NBER WORKING PAPER SERIES

A THEORY OF EQUALITY BEFORE THE LAW

Daron Acemoglu
Alexander Wolitzky

Working Paper 24681
http://www.nber.org/papers/w24681

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018

We thank Bob Gibbons, Suresh Naidu, Jean Tirole, John Wallis, and seminar participants at Northwestern for useful discussions and comments. We gratefully acknowledge financial support from the Carnegie Foundation and the NSF. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Theory of Equality Before the Law
Daron Acemoglu and Alexander Wolitzky
NBER Working Paper No. 24681
June 2018
JEL No. C73,K10,P16,P51

ABSTRACT

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

Daron Acemoglu
Department of Economics
Massachusetts Institute of Technology
50 Memorial Drive
Cambridge, MA 02142-1347
and CIFAR
and also NBER
daron@mit.edu

Alexander Wolitzky
Department of Economics
Massachusetts Institute of Technology
50 Memorial Drive
Cambridge, MA 02142-1347
wolitzky@mit.edu

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, barons, lords, military leaders, etc.—as well as laws that privilege these elites. Some scholars, such as Berman (1983), Hayek (1960), Jones (1981), and Kern (1956), emphasize the historical roots of equality before the law in Europe, dating back to Greek or Roman legal traditions, the customary laws of Germanic tribes, the English common law tradition, or various turning points in the Middle Ages. More recently, North, Wallis and Weingast (2009) have suggested that the broader notion of 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. In none of these cases is it clear, however, why elites with disproportionate political and coercive power find it acceptable—much less in their own interests—to be bound by the same laws as common citizens.[1]

This paper provides a simple framework for addressing this question. Our starting point is that society can be organized without a state, or it can be organized under the auspices of a state with the power to "enforce laws" coercively: that is, to punish individuals who deviate from prescribed behavior.[2] In the former case, "pro-social" behavior is supported only by community enforcement, in particular by the threat of withdrawal of future cooperation; in the latter, incentives are provided by both community enforcement and the threat of coercive punishment.[3] In turn, coercive state

[1] 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 toward regular citizens—but does not focus on whether this is accompanied by equality before the law. We return to our relationship with this literature below.

[2] We thus associate the threat of coercion with "legal enforcement". This may suggest a distinction between "social norms" enforced by the threat of withdrawal of cooperation by the community and "laws" enforced by the threat of coercion. Our favored interpretation is different, however: we view prescribed, on-path behavior as a combination of norms and laws, and put the emphasis on whether there is legal enforcement (threat of coercion) by the state or agents specialized in violence. This is for two reasons. First, as we will see, deviators suffer from both withdrawal of cooperation by the community and coercive punishment, so these two types of incentives are intertwined in our model. Second, as emphasized by Hart (1961) and Tyler (2006), among others, laws that are obeyed are typically embedded in a society's norms, which militates against a sharp distinction between laws and norms in practice.

[3] As emphasized in the context of organizations by Macaulay (1963), Williamson (1985), and Baker, Gibbons and Murphy (1994, 2002), and in the broader context of governance by Granovetter (1985), Ostrom (1990), Milgrom, North, and Weingast (1990), Greif (1993), and Dixit (2003), reputation and the threat of withdrawal of cooperation continue to matter greatly even when legal enforcement becomes commonplace. This feature is a crucial ingredient of our model as well, as we explain below.

enforcement can exist under "elite domination"—where a subset of agents control the means of violence, enforce laws from which they disproportionately benefit, and are themselves above the law—or under "equality before the law", where laws apply equally to all citizens. Our main focus is understanding the transition from elite domination to equality before the law.

In our model, a large number of agents repeatedly take costly actions that generate social benefits. These actions stand for productive effort and other pro-social behavior, such as contributions to public goods or collective defense, and are perfectly observed. In a society without a state, productive effort can be enforced only by the "carrot" of continued societal cooperation. This can be achieved by norms supported by standard community enforcement mechanisms: for example, a deviator can be ostracized and excluded from cooperation until she exerts effort to "repent" for her misdeed. Though community enforcement can support some positive level of equilibrium effort, this level is typically low, owing to the weak nature of this type of enforcement.

The alternative is an organization of society where there is state enforcement of laws, so the carrot of future cooperation is supplemented with the "stick" of coercive punishment. Specifically, we suppose that "the state" acquires the means of violence, which can be used to inflict additional punishments on agents who deviate ("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.[4] Thus, in this elite-dominated equilibrium (which we refer to as "elite enforcement"), all agents face the carrot of future cooperation, while normal agents are additionally confronted with the stick of coercive punishment. Whereas under community enforcement there is relative equality across all agents (in the model, perfect equality), under elite enforcement the threat of punishment makes normal agents work harder than elites, creating inequality. This implies that elites are always better off under elite enforcement, while normal agents may (or may not) also be better off due of the greater level of productive effort induced in this equilibrium.

The most important part of our analysis turns to the question of why the elite would want to 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, but we now also let them choose to what extent they themselves should be subject to "the law", and thus to coercive punishments, when they deviate. We establish several key results.

Most importantly, 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: 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

[4] To be clear, the elite are above the law but are not "above social norms": when they deviate from equilibrium behavior, they still suffer withdrawal of cooperation.

elite and normal agent effort is the central mechanism that may make the elite favor equality before the law.

Table 1 provides a schematic representation of the different enforcement regimes. Community enforcement corresponds to low coercion and low inequality (but also low effort). Elite enforcement involves high coercion and higher effort, but also high inequality favoring the elites. Finally, equality before the law relies on a high level of coercion too, 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	community enforcement	equality before the law
high inequality	?	elite enforcement

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, under full equality before the law, even from the perspective of the elite themselves, it becomes optimal to have complete equality—the elite give up all of their privileges and exert the same level of effort as normal citizens.

What triggers the transition from elite domination to equality before the law? While our model highlights a number of factors affecting this trade-off, we believe the most important one is the role of violence in society. We show that as the extent of punishments that can be imposed on deviators decreases—for technological, political or social reasons—it becomes more attractive for the elites to give up their privileges and transition to equality before the law. This is because of the two levers affecting normal agents' incentives, the stick (coercive punishment) becomes less important and the carrot (the promise of future benefits) becomes more important.[7] This changes the trade-off facing the elite and encourages them to increase their own effort. The elite then find it beneficial subject themselves to coercive punishments in order to achieve this increase in own effort. This comparative static thus links our explanation for the emergence of equality before the law to political and social changes, such as the rise of democratic politics (cfr. footnote 1), which increase the standing and power of non-elites and put natural limits on how harshly they

[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.

[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 enforcement.

[7] In other words, the carrot and the stick are substitutes. This is because of diminishing marginal returns to effort—the more effort is obtained by the threat of withdrawal of cooperation, the less valuable is the marginal effort obtained by additional coercive punishments.

can be treated by the state or the elite, as well as to social forces limiting the acceptability of such punishments (e.g., Elias, 1994, Pinker, 2011). In this light, our theory gives a novel explanation for why moves towards greater mass participation and limits on elite power in politics have often been accompanied by the rise of equality before the law. A complementary comparative static is that if the extent of coercive punishments remains unchanged but the political power of the elite declines (which may again result from the rise of democratic politics), society again moves towards equality before the law.

Several instances of the gradual evolution of equality before the law around the world can be interpreted through the lens of this comparative static. The British case is often emphasized in discussions of the rule of law, with many scholars tracing the roots of these notions to the Middle Ages or even earlier. These important legal and political traditions notwithstanding, Britain remained far from equality before the law as late as the mid-19th century. An emblematic example is provided by a set of laws creating onerous obligations for manual workers and privileges for employers, who could ban workers from quitting their jobs, or even from turning down unattractive offers (Steinfeld, 2001, Naidu and Yuchtman, 2013). The Statute of Laborers, enacted in the 14th century, empowered landowners to compel workers to work at set wages. In Steinfeld's words, "The English laboring poor of this period... were subject to an oppressive regime of legal regulation" (2001, p. 8). This statute was reconfirmed by later, 16th-century statutes, and was extended to a handful of artisanal occupations in the 18th century (it was also imported by the American colonies and formed the backbone of their labor law). 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. Social and political changes during the 19th century, in part spearheaded by democratization, made this coercive institution less and less tenable, however. A first step was the 1867 Master and Servant Act, which prohibited whippings and imprisonment, even as it simultaneously increased the ability of magistrate courts to compel workers to work at the terms offered by their employers. In a second, critical, step towards equality before the law, this act itself was finally repealed in 1875.

Limits on punishments also played a role in what was arguably the first society approaching equality before the law: Athens between the 6th and 4th centuries BC. Starting with Solon, and continuing with Cleisthenes's reforms, the ability of elite Athenians to command a privileged position towards regular Athenian citizens (and even slaves) was curtailed (Snodgrass, 1980, Ober, 2015). Though many factors likely played a role in the rise of Athenian institutions, an important element was a change in military technology that empowered regular citizens of Athens, now armed with iron weapons as hoplites (citizen infantry). This contrasts with the "palace economies" of

Bronze Age Greece, circa 16th–11th centuries BC, where weapons were more expensive and were thus monopolized by the elite. In the famous words of Gordon Childe (1942), "Iron democratized agriculture and industry and warfare too." This democratization of warfare implies, in the context of our model, a more limited ability of elites to punish the (now more heavily armed) citizens, and hence shifts society towards equality before the law.

Equality before the law can also emerge due to factors other than the diminished power of the elite and limits on their ability to impose punishments. A notable possibility is that a change in the nature of production can alter the trade-off facing the elite, for example, because effort becomes more important for production or for the provision of vital public goods, such as national defense. However, as we will see, 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 productive activities or national defense—when these correspond to greater marginal product of effort—that contribute to the development of equality before the law.

This comparative static can be illustrated by several well-known cases of "defensive modernization" in the 19th century. The abolition of (some of) the privileges of the Japanese military elite following the Meiji Restoration is instructive. Tokugawa Japan had an explicit social class system, where the armed samurai approximate the above-the-law elites in our model. 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 up to foreign (especially American) trade. These events convinced some key Japanese elites that a major reform process was vital to strengthen their economy and national defense. It is in this context of a perceived existential threat that the Meiji Restoration of 1866 took place, disbanding the Tokugawa shogunate and restoring the monarchy (Jansen, 2002, Buruma, 2003, Ravina, 2017). The Meiji government removed the de jure unequal treatment of different social classes and disarmed the samurai (some of whom remained specialists in coercion, but now as police officers under the control of the central state). The Meiji Constitution, drafted in the 1880s and finally promulgated in 1890, introduced such notions as due process before the law, freedom of movement, freedom of speech, and private property for all Japanese. While 19th-century Japan remained an oligarchic society, these changes created a much greater degree of equality before the law. An important question in this context is why the Meiji reforms did not just attempt to modernize the military and the fiscal system, but also took steps towards greater equality before the law. Our model suggests a potential answer—greater equality before the law may have improved elite behavior and

consequently induced greater societal effort towards modernization at a time when the need for (i.e. the marginal returns from) such effort was very high. Similarly, the process of legal reforms in 19th-century Prussia—which abolished various vestiges of serfdom and can thus be viewed as an important step towards equality before the law (Fisher, 1903, Blanning, 1989, Acemoglu et al., 2011)—and the *Tanzimat* reforms in the Ottoman Empire promulgated in the Rose Garden Edict in 1839—which introduced some degree of equality before the law, including for various non-Muslim minorities (Zürcher, 2004, Owen, 2004)—were responses to foreign threats as well.

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

In addition to the literatures on the historical origins of rule of law and democratization mentioned above, three others need to be highlighted. The first is the literature pioneered by North and Weingast (1989), which interprets constitutions and other institutional features as commitment devices for respecting other groups' property rights, and thus encouraging greater investment and economic participation.[8] 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 property rights and residual control rights within an organization strengthens some agents' investment incentives by reducing holdup. The result that equality before the law encourages normal citizens to exert effort by removing elite privileges and increasing elite effort bears some similarity to these insights, but with several important differences. First, equality before the law is not a commitment to a constitutional provision but an alternative organization of society

[8]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).

leading to a different repeated game equilibrium. Second, equality before the law impacts 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 withdrawal of cooperation and does not consider costly punishments (Kandori, 1992, Ellison, 1994, Wolitzky, 2013, Ali and Miller, 2014). 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, 2018). These papers investigate neither enforcers' willingness to subject themselves to punishment nor equality before the law.

Finally, our paper is also 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 the elites control the means of violence and are privileged. The key mechanism that equality before the law enhances community enforcement does not arise in this literature.

The rest of the paper is organized as follows. Section 2 introduces our baseline environment. Section 3 characterizes the best equilibrium under community enforcement (without the state). Section 4 analyzes the same environment under elite domination, while Section 5 studies the optimal degree of equality before the law from the viewpoint of the elite. Section 6 presents our main comparative static results, which delineate factors that encourage the emergence of equality before the law. Section 7 generalizes the baseline environment to a matching model in which, in addition to benefitting society at large, effort generates private benefits for one's partner. While in the baseline model elites are privileged only because they exert lower effort than others, in this extended environment the best equilibrium from the viewpoint of elites also involves normal agents working harder when they match with elites. Section 8 extends the model to study within-elite heterogeneity in terms of productivity and the implications of a hierarchical structure within the elite. Section 9 concludes. All proofs are presented in the Appendix.

2 Environment

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

2.1 The Baseline Environment

There is a continuum of infinitely-lived agents that discount the future with discount factor $\delta \in (0,1)$. Fraction 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") $x_i \in \mathbb{R}_+$.[9] 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 the "benefit production functions" that map units of disutility of effort to units of benefits for society. They are strictly increasing, strictly concave, and bounded, and satisfy $f_N(0) = f_E(0) = 0$ and $f_N'(0), f_E'(0) > 1/\delta$. The assumption that $f_N'(0), f_E'(0) > 1/\delta$ (and hence $f_N'(0), f_E'(0) > 1$) implies that the stage game is a continuous-action version of the prisoners' dilemma. We allow the functions f_N and f_E to differ for normal and elite agents as these agents may have different roles in production—for example, "effort" by elites could simply correspond to "not expropriating others" (see footnote 14 below), or it could represent business investment 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 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.[10]

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

[9] 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.

[10] Combining a continuum population and perfect monitoring/observability raises measurability issues that make formally defining strategies complicated. Rather than addressing these issues formally, we simply assert that our model is obviously the limit of a large finite population. Indeed, the only reason we assume a continuum rather than a finite population is to ensure that, for both a normal agent and an elite agent, the fraction of *other* agents with elite status is 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.

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.

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

Throughout, we focus on stationary, symmetric, subgame perfect equilibrium (*equilibrium* henceforth) as the solution concept. 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.[11]

2.2 A Random Matching Interpretation

The economy described so far is "centralized" in two ways: each individual's effort directly benefits everyone in society, and a centralized state directly allocates punishments. We remark that it is straightforward to give a mathematically equivalent decentralized interpretation (or a hybrid interpretation where only one of these dimensions is decentralized).

Suppose first that effort is still a pure public good, but the means of coercion are controlled by the elite. Players randomly match in pairs, and an elite agent can punish her partner in the match. Suppose also that punishing one's partner is costless (so a player is indifferent as to whether or not to punish her partner), and that punishment inflicts disutility g/α on a normal agent and $\rho g/\alpha$ on an elite (this scaling by $1/\alpha$ keeps the expected disutility of punishment fixed at g, as there are α

[11] 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 "long-run welfare".

elites in the population). This variant of the model where punishments are carried out by elites is completely equivalent to the baseline model.

Next, suppose that benefits are also generated within matches, and a player only benefits from the effort of her partner. Then, provided that effort levels are chosen before players observe their partners' status as normal or elite (while status is subsequently observed at the punishment stage), each agent must choose the same level of effort regardless of her partner's status, and therefore her effort generates the same expected benefit for everyone. This "anonymous" matching model thus endogenously generates the pure public good feature that was assumed in the centralized model. This version of the model—where all economic interactions take place within matches—remains mathematically equivalent to the baseline model. In Section 7, we study a variant of this model where matching is non-anonymous, so players know their partners' status when choosing effort. In this case, effort is no longer a pure public good, but we will see that our most important results continue to apply.

3 Community Enforcement

We first consider the model with $\alpha = g = 0$, where all agents are identical and no coercive punishments are available. This gives a model of community enforcement of cooperation. The following result is standard: for this result, and throughout the paper, we denote the *first-best* (surplus-maximizing) normal agent effort level by

$$x^{FB} = \left(f_N'\right)^{-1}(1).$$

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

$$x = \delta f_N(x). \tag{1}$$

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

Grim trigger strategies are one way of supporting the unique optimal equilibrium effort level characterized in Proposition 1, but not the only one. In a different optimal equilibrium, a player's punishment for deviating in period t is that in period $t+1$ she must play $x_i = x$ while her opponents all play $x_j = 0$, and all players restart the original equilibrium in period $t+2$ if this

punishment is successfully carried out. Relative to grim trigger, this "repentance" equilibrium has the advantage that it is renegotiation proof (Farrell and Maskin, 1989, Van Damme, 1989). Whether the withdrawal of cooperation that supports effort level x^{CE} is carried out via grim trigger strategies, repentance, or some combination of the two is irrelevant for our results—in particular, our results do not require "extreme" community-wide punishments for individual deviations. The same comment will apply in later sections where cooperation is supported by the threat of both the withdrawal of cooperation and coercive punishment.

In practice, the most common way in which cooperation is withdrawn from deviators is *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 either permanently ostracized or ostracized until they repent by exerting effort without receiving any benefits as described in the previous paragraph. It is straightforward to verify that introducing ostracism in this way does not affect our equilibrium conditions, and hence does not affect any of our results, except that the community can now discourage deviations with the threat of ostracism.[12]

Proposition 1, especially with the repentance or ostracism interpretation, provides a stylized representation of social order in stateless (small-scale) societies. First, the equilibrium involves low levels of inequality across agents (in our simple model, no inequality at all). This is consistent with the evidence from the anthropological and archaeological literatures on the strong emphasis on and practice of egalitarianism in most stateless societies (Bohannan and Bohannan, 1953, Boehm, 1999, 2012, Flannery and Marcus, 2014). Second, little coercion is used to support pro-social behavior (in our model, no coercion). Although there is continuous infighting, blood feuds, and endemic violence in many stateless societies (Chagnon, 1968, Boehm, 1986, LeBlanc and Register, 2004), there is limited use of coercion to support cooperation. Indeed, much violence in stateless societies appears to result from inter-group conflict (LeBlanc and Register, 2004), 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, 2018). In contrast, 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 such societies (see, e.g., Briggs, 1970, on the

[12] 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).

Inuit, Woodburn, 1982, on the Hadza, or Wiessner, 2005, on the !Kung Bushmen; see Baumard, 2010, for a general discussion). Rather, in all of these cases, cooperation appears to be supported by a combination of low social regard directed at non-cooperators and the threat of withdrawal of future cooperation, for example via social isolation. The same appears to be true in societies with nascent but still weak state institutions, such as Germanic tribes and subsequently Frankish states shortly after the fall of the Western Roman Empire, as well as early Anglo-Saxon England: in these cases, most infractions were punished by payments from perpetrators to victims or their families, for example via the "wergeld" as specified by the Salic Law of the Franks or King Alfred's Law Code (Drew, 1991, Acemoglu and Robinson, 2018). This arrangement closely resembles community enforcement supported by repentance and/or ostracism, as described above.[13]

4 Elite Enforcement

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

Proposition 2 *Under elite enforcement,*

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

$$\max_{x \geq 0, y \geq 0} (1-\alpha) f_N(x) + \alpha f_E(y) - y \tag{2}$$

subject to

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

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

2. *Constraint (3) binds at the optimum.*

3. *Let us denote the unique pair $(x, y) > (0, 0)$ such that both (3) and (4) bind by $\left(\bar{x}^{EE}, \bar{y}^{EE}\right)$,*

[13] The example of wergeld raises the question of whether introducing monetary transfers would matter for the model. The answer is essentially no: as long as $f_N'(x) > 1$ and $f_E'(y) > 1$ for effort levels that arise in equilibrium, it is more efficient to demand effort rather than on-path transfers, and replacing off-path "repentance effort" with "repentance transfers" would not affect any of our results.

and denote the solution to (2) subject to (3) and (4) by $\left(x^{EE}, y^{EE}\right)$*. Then we have*

$$\begin{aligned}
\alpha f'_E\left(y^{EE}\right) + \delta\left(1-\alpha\right) f'_N\left(x^{EE}\right) &\leq 1 \text{ if } y^{EE} = 0, &(5)\\
\alpha f'_E\left(y^{EE}\right) + \delta\left(1-\alpha\right) f'_N\left(x^{EE}\right) &= 1 \text{ if } y^{EE} \in \left(0, \bar{y}^{EE}\right), &(6)\\
\alpha f'_E\left(y^{EE}\right) + \delta\left(1-\alpha\right) f'_N\left(x^{EE}\right) &\geq 1 \text{ if } y^{EE} = \bar{y}^{EE}. &(7)
\end{aligned}$$

Note that (2) is elite welfare, since elites receive per-period benefits of cooperation $\left(1-\alpha\right) f_N\left(x\right) + \alpha f_E\left(y\right)$ and exert effort y. In this maximization problem, (3) is the incentive constraint for a normal agent, and (4) is the incentive constraint for an elite agent.[14] These constraints are intuitive: any player who deviates loses an expected benefit of $\left(1-\alpha\right) f_N\left(x\right) + \alpha f_E\left(y\right)$ in the next period. Moreover, normal agents that deviate face an additional coercive punishment of g.[15] There is no such punishment for elite agents (as elites are "above the law"), so this second term is not present in (4).[16] Furthermore, in the best equilibrium for elites, normal agents are always required to work as hard as possible, so (3) binds.

For the last part of the result, (6) is the first-order condition with respect to y, once x has been substituted out of the objective function using (3). This expression captures the fact that elites benefit in two ways from working harder. First, there is a direct marginal benefit of elites' effort on other elites' utility (the $\alpha f'_E\left(y\right)$ term). Second there is an indirect marginal benefit (the $\delta\left(1-\alpha\right) f'_N\left(x\right)$ 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—is the crux of our theory and is responsible for our comparative static results below. It is also this indirect effect that captures the repeated game aspect of the equilibrium, as can be seen by noting that this effect disappears when $\delta = 0$.

To better understand the indirect effect and 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 $\delta\left(1-\alpha\right) f'_N\left(x\right)$ units of benefit to both normal agents and elites. These units of benefit in turn increase normal agents' effort by another $\delta^2\left(1-\alpha\right) f'_N\left(x\right)$ units, which provide $\delta^2\left(1-\alpha\right)^2 f'_N\left(x\right)^2$ units of benefit, and so on. The total marginal benefit to

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

[15] This role of coercive punishment g in deterring deviations is somewhat similar to that in Acemoglu and Wolitzky (2011), where we assumed that employers/principals could use coercion in order to reduce the outside option of their employees/agents, thus forcing them to accept contracts that they would otherwise reject.

[16] However, there are still "norms" that trigger withdrawal of cooperation if elite agents deviate from equilibrium behavior. It is these norms that incentivize $y > 0$.

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 (6).[17]

There are once again multiple ways of supporting the unique optimal equilibrium path: for example, we can specify either that cooperation breaks down forever once an agent deviates ("grim trigger"), or that cooperation breaks down for only a single period while the deviator continues to cooperate ("repentance"). Again, repentance has the advantage of being renegotiation-proof.[18]

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

$$\begin{aligned}
u^{CE} &= f_N\left(x^{CE}\right)-x^{CE},\\
u_N^{EE} &= (1-\alpha)f_N\left(x^{EE}\right)+\alpha f_E\left(y^{EE}\right)-x^{EE}, \text{ and}\\
u_E^{EE} &= (1-\alpha)f_N\left(x^{EE}\right)+\alpha f_E\left(y^{EE}\right)-y^{EE}
\end{aligned}$$

denote (elite-)optimal payoffs under community enforcement and elite enforcement, for (N)ormal and (E)lite agents. It is clear that elites prefer elite enforcement to community enforcement: $u_E^{EE}\geq u^{CE}$. However, normal agents may or may not prefer elite enforcement to community enforcement. The tradeoff is that under elite enforcement normal agents work harder than elites (i.e., $x^{EE}>y^{EE}$) and therefore receive a smaller share of the total social surplus $(1-\alpha)(f_N(x)-x)+\alpha(f_E(y)-y)$ than under community enforcement, but total social surplus can be higher under elite enforcement than under community enforcement because the threat of coercive punishment increases the maximum sustainable effort level (i.e., $\min\left\{\bar{x}^{EE},\bar{y}^{EE}\right\}>\bar{x}^{CE}$).[19]

[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 Section 7, we show similar results obtain when effort creates a mix of public benefits and private returns for one's partner.

[18] In addition, in Acemoglu and Wolitzky (2018), we show that if punishments are costly to carry out, then another advantage of repentance over grim trigger is that it improves incentives for punishment. The implications of our analysis here are very different from that paper, not only because punishments are costless, but also because the elite take productive actions and there is a choice of how much punishment the punishers/elite should be subject to.

[19] To see that $\bar{y}^{EE}>\bar{x}^{CE}$, note that $\bar{x}^{CE}$ is the positive root of the concave function $\delta f(x)-x$, while $\bar{y}^{EE}$ is the positive root of the concave function $\delta[(1-\alpha)f(y+\alpha g)+\alpha f(y)]-y$. The latter function is everywhere greater than the former, so it has the greater root.

Several takeaways are worth emphasizing. First, in contrast to community enforcement, elite enforcement involves high inequality and high coercion. Both of these are characteristics of early societies that developed state institutions (either in the form of chieftaincies, proto-states or what anthropologists label "states"; Johnson and Earle, 2000, Flannery and Marcus, 2014). These characteristics 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.

Second, an important debate concerns whether the transition from stateless societies to those with more organized institutions and coercion was welfare-improving for the population at large because it encouraged better cooperation or dispute resolution (as maintained by various social contract theories going back to Thomas Hobbes and John Locke; see also Huntington, 1968, Bates, 2001, Fukuyama, 2011), or welfare-reducing for most because it led to exploitation by the elite (as maintained by Scott, 2017, and suggested by evidence of affluence and relatively good health among some stateless societies, e.g., Sahlins, 1974, Suzman, 2017). Our analysis shows either outcome is possible. Under elite enforcement (relative to community enforcement), there is greater inequality favoring the elite, which tends to make normal agents worse off. At the same time, because higher effort benefits everyone, normal agents may become better off as well. Elite agents are always better off under elite enforcement, because they benefit both from the higher effort of normal agents and from their privileged position resulting from their monopoly on coercion and their above-the-law status. This feature is also consistent with the existing archaeological and historical evidence (e.g., Flannery and Marcus, 2014).

We end this section by discussing how the comparison between u^{CE} and u_N^{EE} depend on g. This dependence is ambiguous in general, but it can be characterized when the fraction of elite agents, α, is small.

Proposition 3 *Assume $f'_E(0) < \infty$. There exists $\bar{\alpha} > 0$ such that if $\alpha < \bar{\alpha}$, then u_N^{EE} is single-peaked in g.*

The proof shows that when α is sufficiently small so that elites do not find it in their interest to exert effort under elite enforcement, normal agent welfare under elite enforcement is maximized at an intermediate level of g (or is monotonically decreasing in g): a very low g implies insufficient production, while a very high g implies excessive coercion (from the viewpoint of normal agents). This proposition thus highlights some of the forces that determine whether normal agents will benefit from a transition to elite enforcement. If the extent of coercion is very high (for example, as in ancient empires relying on large-scale labor coercion, such as Egypt or Sparta), the inequality effect dominates and normal agents are worse off. In contrast, if there is very little coercion, then

elite enforcement does not lead to a significant increase in effort, and normal agents cannot benefit much from this transition.

Finally, it is useful to note that the elite enforcement model includes the special case $g = 0$ where there is no coercive technology but elites and normal agents may still be treated asymmetrically. In other words, if $\alpha > 0$ while $g = 0$, the model allows political hierarchy—and potentially some degree of inequality—even in the absence of coercion. Note, however, that when $g = 0$, we have $\bar{x}^{EE} = \bar{y}^{EE}$: that is, if it is optimal for elites to work at the maximum sustainable level when $g = 0$, then egalitarianism is their most preferred option. While it is not always optimal for elites to exert maximal effort, we have that $\bar{x}^{EE}$ and $\bar{y}^{EE}$ are increasing in g, so (7) is easiest to satisfy—and thus equal levels of effort are most likely—when g is small. This result that egalitarianism is most likely to arise when g is small establishes our earlier claim that low coercion and equality go hand-in-hand: when the elite cannot use coercion effectively, it is optimal from their viewpoint for them to exert the same level of effort as normal agents. This is also the reason why we believe the low coercion, high inequality cell in Table 1 is not well-populated.

5 Equality Before the Law

We now allow for the possibility that $\rho \geq 0$, so that elites may also be subject to some degree of coercive punishment. We take $\rho \in [0, 1]$ to be a choice variable for the elite, and continue to focus on the elite-optimal equilibrium.[20] The interpretation is that we view the elite as holding the political power to choose both the institutional environment (ρ) and the equilibrium. The resulting problem for the elites is

$$\max_{x \geq 0, y \geq 0, \rho \in [0,1]} (1 - \alpha) f_N(x) + \alpha f_E(y) - y \tag{8}$$

subject to

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

$$y \leq \delta [(1 - \alpha) f_N(x) + \alpha f_E(y)] + \rho g, \tag{10}$$

where (10) is the incentive compatibility constraint for elites. Here (9) is identical to (3), while (10) differs from these constraints only in that an elite agent's minmax payoff is $-\rho g$ rather than $-g$.

Since increasing ρ relaxes the incentive constraint of the elite and we focus on the elite-optimal equilibrium, the elite are always willing to choose $\rho = 1$ (full equality before the law) and not punish themselves: intuitively, the elite are happy to allow themselves to be subject to coercion, provided the equilibrium specifies they are never actually coerced. To rule out this rather artificial possibility, we assume the elites always choose the smallest level of ρ when indifferent—this corresponds to

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

imposing a small cost of increasing ρ and then considering the limit where this cost vanishes. 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 (9) to bind, as increasing x increases the objective and relaxes (10): hence, $x^{EL} = x^*\left(y^{EL}\right)$, where again $x^*(y)$ is the value of x that makes (9) hold with equality. Let $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ be the unique pair (x, y) such that (10) binds with $\rho = 1$. That is, $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ are the greatest sustainable effort levels under equality before the law. Note that $\bar{x}^{EL} = \bar{y}^{EL}$, which implies that the maximum sustainable effort level for normal and elite agents is the same under full equality before the law.

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

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

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

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

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

$$\alpha f_E'\left(y^{EL}\right) + \delta\left(1-\alpha\right) f_N'\left(x^{EL}\right) = 1. \tag{11}$$

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*

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

The maximization problem (8) differs from (2) only in that ρ is now a choice variable, rather than being fixed exogenously at 0. As in the earlier problem, the incentive compatibility constraint of normal agents, (9), always binds, and that of elite agents, (10), binds only if the best equilibrium for the elite involves maximum elite agent effort. Hence, if (10) with $\rho = 0$—or if equivalently the corresponding constraint (4) under elite enforcement—is slack, then elites have no interest in committing themselves to a higher level of effort, and instead prefer to remain in the elite enforcement regime with $\rho = 0$. In contrast, if (4) binds under elite enforcement (or equivalently, if (7) holds with strict inequality), then the elites opt for at least partial equality before the law, where the optimal level of equality before the law is just sufficient to commit themselves to the effort level y^{EL} satisfying the first-order condition (11).[21] Finally, in the case where $\alpha f_E'\left(\bar{y}^{EL}\right) +$

[21] The intuition for this first-order condition with endogenous ρ is the same as for the one with $\rho = 0$ given in (6):

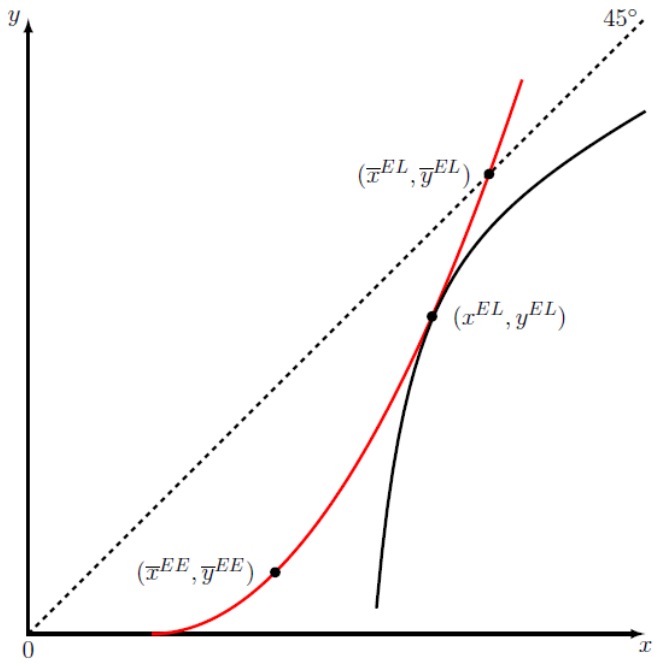

Figure 1: The black curve represents an indifference curve for the elite, while the red curve represents the boundary of the incentive compatibility constraint (9). The point $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ corresponds to full equality before the law ($\rho = 1$) and the point $\left(\bar{x}^{EE}, \bar{y}^{EE}\right)$ corresponds to elite enforcement ($\rho = 0$).

$\delta(1-\alpha) f'_N\left(\bar{x}^{EL}\right) \geq 1$, elites prefer full equality before the law. Interestingly, when this is the 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 completely equal allocations.[22] This then yields another way of viewing the last part of the proposition: the elite prefer to establish full equality before the law only when they are willing to work as hard as normal agents.

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

$$\max_{x \geq 0, y \in [0, \bar{y}]} (1-\alpha) f_N(x) + \alpha f_E(y) - y \tag{12}$$

subject to (9), where $\bar{y} = \bar{y}^{EE}$ under elite enforcement and $\bar{y} = \bar{y}^{EL}$ under endogenous choice of ρ. We illustrate this problem diagrammatically in Figure 1. The thick curve represents combinations of normal agent and elite effort that satisfy the normal agents' incentive compatibility constraint,

the direct marginal benefit to elites of increasing their effort is $\alpha f'_E(y)$, and the indirect marginal benefit—coming through the induced increase in the maximum incentive compatible level of normal agent effort—is $\delta(1-\alpha) f'_N(x)$. The first-order condition sets the total marginal benefit of $\alpha f'_E(y) + \delta(1-\alpha) f'_N(x)$ equal to the total marginal cost of 1.

[22] Normal and elite agents exert the same effort even though f_N and f_E may differ. This is because effort levels of the two types of agents under equality before the law are determined by their binding incentive compatibility constraints, which are identical and thus imply the same level of effort. This is no longer the case in Section 7, where elites may receive greater benefits from cooperation.

(9), as an 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 $\left(x^{EE}, y^{EE}\right)$, corresponding to elite enforcement with $\rho^* = 0$, is plotted as well. The figure also superimposes the indifference curves of (12), which are convex (since (12) is concave). The point of tangency, if any, between these indifference curves and the boundary of (9) gives the combination of (x, y) that is optimal from the viewpoint of the elite; such a point of tangency corresponds to an intermediate value of $\rho^* \in (0, 1)$. When there is no tangency, the highest indifference curve is reached either at the corner where $(x, y) = \left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ with full equality before the law ($\rho^* = 1$), or at the point where $(x, y) = \left(x^{EE}, y^{EE}\right)$ with elite enforcement ($\rho^* = 0$).

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

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

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

Several points are worth noting here. Equality before the law, just like elite enforcement, makes heavy use of the threat of coercive punishment in order to encourage pro-social behavior. However, in contrast to elite enforcement, it features a low degree of inequality: elite agents are not treated in a privileged manner. As already anticipated in the Introduction, this feature of equality before the law in our model has much in common with the ideal of "rule of law" of Hayek (1960), who in particular emphasized the defining role of equal application of laws and equal protection from coercion. This type of equality before the law is also a critical component of the concept of "open access order" proposed 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. It is also a key aspect of inclusive economic institutions in Acemoglu and Robinson (2012), which depend on a level economic playing field among all individuals and thus the removal of various privileges before the law. Indeed, the evolution of many Western societies towards more democratic and inclusive institutions can be viewed precisely as such a process of stripping away the privileges of elites.

Finally, we have so far assumed that if the best equilibrium for the elite involves some degree of equality before the law—$\rho > 0$—then the elite can freely choose and commit to such an arrangement. An important question is how this can be secured in practice. For example, in the matching

environment outlined in Section 2.2 where the elite control the means of coercion, they may be unable to commit to subjecting themselves to punishments. One obvious and historically common solution is to separate coercive functions from elite status. The vital aspect of this solution is to transfer the means of coercion from elites to agents specialized in law enforcement, similar to what the Meiji government did by disarming the samurai and creating a professional police force. A related solution is to create a (sufficiently independent) government bureaucracy and judiciary to resolve conflicts and decide whom should be subject to punishment. In both cases, the practical challenge is to ensure the independence and impartiality of the agents charged with law enforcement or judiciary functions.

6 Comparative Statics: Towards Equality Before the Law

We now turn to our key comparative statics results on how the elite-optimal levels of production and equality before the law vary with parameters.

6.1 Comparative Statics for Coercive Capacity

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

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

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

Figure 2 provides a diagrammatic intuition for Proposition 6. An increase in g has no impact on the indifference curves of the elite, but shifts the boundary of (9) to the right. The figure also shows that the indifference curves of the elite become less steep as we move to the right along a horizontal line.[23] Consequently, the shift out of (9) leads to a new tangency point with not only greater x, but also lower y. Lower elite effort then translates into a lower level of equality before the law.

A complementary intuition is that coercive punishments and incentives provided by norms/threat of withdrawal of future cooperation are substitutes at the margin. The greater is g, the less need there is for additional incentives coming from norms, and this allows the elite reduce y. More precisely, recall that part of the elites' incentive to choose greater effort y is that this indirectly

[23] 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).

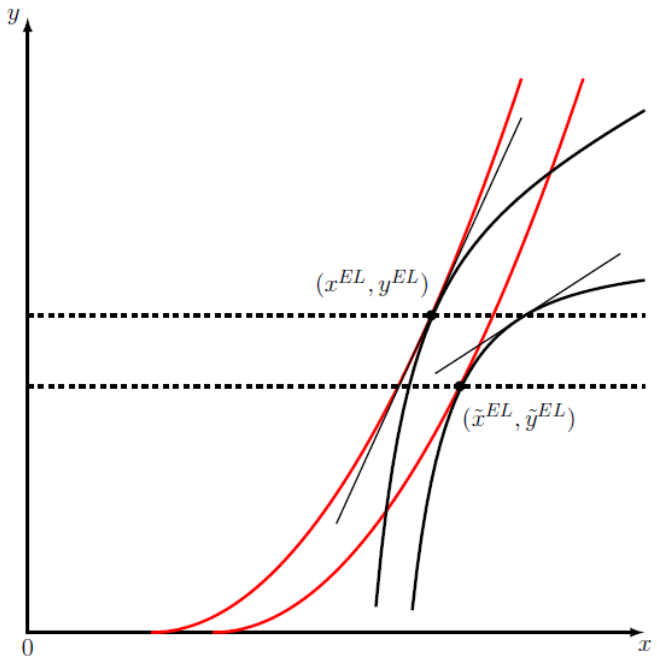

Figure 2: The indifference curves of the elite become flatter as we move to the right along a horizontal line. An increase in g shifts out the red curve representing the boundary of the incentive compatibility constraint (9) to the right, and thus leads to a new point of tangency with greater x and lower y, and thus lower ρ^*.

increase normal agent effort x. An increase in g raises x for a fixed level of y. Because f_N is concave (i.e., there are diminishing returns to effort), the term $\delta(1-\alpha) f'_N(x)$ in (6), which captures this indirect effect, declines when x increases. This encourages the elite to choose a lower y. Since increasing ρ is a way to raise y (by making the elite subject to greater coercive punishments), an increase in g also leads to a reduction in ρ.[24] The fact that this comparative static ceases to be strict when $\delta = 0$ confirms this intuition, since in this case there are no repeated game considerations and hence no indirect effect.[25]

This comparative static sheds light on one set of powerful forces leading to the emergence of equality before the law. As already mentioned above, there are at least three distinct but related reasons for why effective coercive capacity g may decline. First, technological changes may reduce the ability of the state and elites to impose punishments on normal agents or increase the ability of normal agents to resist such punishments. This may be relevant for understanding the social and political changes that Athens underwent between the 6th and 4th centuries BC. Providing greater legal equality to non-elites was one of the major objectives of the political reforms initiated by Solon in 594 BC and continued by Cleisthenes at the end of the 6th century BC. For

[24] 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.

[25] 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 $\delta = 0$.

example, Solon promulgated a hubris law, which made it illegal—in fact, a capital offense—to act "hubristically" toward (violently intimidate or humiliate) any Athenian, even a slave. The hubris law may be viewed as a nascent form of equality before the law. Cleisthenes's subsequent ostracism law, which enabled Athenians to ostracize and exile powerful politicians and elites, further curtailed the political privileges of the elite, and can thus be viewed as another step in this direction. An important factor paving the way for this institutional revolution was the change in military technology from the Bronze Age to the Iron Age. When weapons were made of bronze, they were expensive and were consequently monopolized by the elite. The Iron Age, as emphasized by Childe (1942) "democratized warfare", and led to more active involvement of free Athenian citizens in war, as hoplites armed with iron weapons (Snodgrass, 1980). Consistent with our comparative static result, this greater equality of access to the means of violence within Athenian society appears to have been an important factor in increasing the demand for institutional change and creating the conditions for the development of laws that applied equally to all Athenian citizens and even provided some degree of protection to slaves (see, e.g., Ober, 2015).

Second, the extent of punishments may also be curtailed because of political changes empowering normal agents. For example, as discussed in the Introduction, mass enfranchisement may have been important in the repeal of the Master and Servant Acts in mid-19th-century Britain.

Third, changing values and social conventions may also limit the extent to which harsh punishments are viewed as socially acceptable (see, e.g., Elias, 1994, Pinker, 2011). This too will correspond to a decline in g in our model and potentially trigger greater equality before the law. This may have been another factor contributing to the repeal of the harshest punishments on workers for contract breach as well as to the removal of other privileges of 19th-century British elites.

Finally worth noting are historical examples where, rather than advancing, equality before the law retrenches. One well-known case from European history is the establishment of the medieval feudal hierarchy, where the well-armed nobility monopolized the means of coercion and significantly increased its privileges. A famous theory about the origins of this feudal order, advanced by Lynn White (1962), links it to the invention and spread of the iron stirrup in Europe starting in the 8th century. According to White, the iron stirrup increased the effectiveness of heavily-armored cavalry in combat and thus intensified the coercive capability of those who could afford horses, armor, and weapons. Several aspects of this thesis, especially concerning the timing of the introduction of the stirrup and the rise of the feudal order, are controversial, however. Our key comparative static here provides an alternative channel through which the spread of the stirrup may have over time contributed to social changes favoring the elite—by improving the technology of coercion and thus creating a force against equality before the law.

6.2 Comparative Statics for Political Power

Our previous comparative static focused on restrictions on the extent of coercion that the elite can exert while still maintaining their political power. Many of the social changes emphasized in that context, most notably the emergence of mass democratic politics, not only put restrictions on the use of coercion but reallocated political power away from the elite towards normal agents (see the discussion and references in Acemoglu and Robinson, 2006). In this subsection, we show that a decline in the relative political power of the elite will also contribute to the emergence of equality before the law. We now establish this result in the simplest possible fashion (without introducing a micro-founded model of the political power of the elite) by simply focusing on the set of equilibria that maximize a weighted average of the utilities of the elite and normal agents, and then reducing the weight of the elite in this social welfare function. In the process, we also confirm that none of our results so far depend on focusing on the best equilibrium from the viewpoint of the elite.

Our first result establishes that under elite enforcement (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 is true whenever normal agents' incentive constraints bind (for example, whenever effort is below the first-best level). The intuition is that elite agents work more at the optimum under more equal Pareto weights, and this in turn induces higher effort from normal agents. Thus, inequality of political power reduces production.

Proposition 7 *Under elite enforcement, let* $\left(x^{EE}(\gamma), y^{EE}(\gamma)\right)$ *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 (3) and (4). For all Pareto weights $\gamma > \gamma' \geq \alpha$*, if* $x^{EE}(\gamma) < x^{FB}$ *then* $x^{EE}(\gamma) \leq x^{EE}(\gamma')$ *and* $y^{EE}(\gamma) \leq y^{EE}(\gamma')$*.*

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

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

Proposition 8 *Let* $\left(x^{EL}(\gamma), y^{EL}(\gamma), \rho^*(\gamma)\right)$ *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 (9) and (10). For all Pareto weights $\gamma > \gamma' \geq \alpha$, if $x^{EL}(\gamma) < x^{FB}$, then $x^{EL}(\gamma) \leq x^{EL}(\gamma')$, $y^{EL}(\gamma) \leq y^{EL}(\gamma')$, and $\rho^*(\gamma) \leq \rho^*(\gamma')$.

6.3 Comparative Statics for the Returns to Effort

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

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

Suppose the production functions f_N and f_E are parameterized by $\theta \in [0,1]$. Let (x_0, y_0, 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 9 *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, 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^*.

The result that $\frac{d\rho^*}{d\theta}$ is non-negative is somewhat subtle. Suppose increasing θ raises marginal returns while leaving average returns unchanged (a case allowed by the proposition). It is quite intuitive that this leads to an increase in x^* and y^*. But why does this encourage greater equality before the law? In other words, why is the increased carrot of future cooperation not enough to justify the resulting higher level of elite effort? Intuitively, increasing 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 will always fall short of the former, because it is incentivized only by the increased benefits that elite agents enjoy in equilibrium, and since the initial allocation was chosen to maximize net benefits to the elite, the implied increase in elite effort from these greater benefits will be small. Hence to achieve the desired increase in y^*, the elite collectively need to make themselves subject to greater coercive punishments.[26]

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 result sheds light on the question we posed in the Introduction—why defensive modernization efforts, such as those in 19th-century Prussia, Japan, and the Ottoman Empire, not only modernize the military and the fiscal system but also attempt to expand equality before the law. The answer that follows from our analysis is that equality before the law is a way of improving the behavior of the elite, and thus indirectly increasing effort on the part of all agents in society.

6.4 Comparative Statics for Inequality

In this subsection, we slightly modify our baseline setup to discuss the effects of economic inequality between elite and normal agents on the emergence of equality before the law. Recent increases in wealth and income inequality around the world (e.g., Atkinson, Piketty, and Saez, 2011) have raised concerns about whether a system based on equal opportunity—and in our setting, equality before the law—can survive in a highly unequal society. Scheidel (2017) argues this has not been possible historically, and only war and revolution have tended to limit inequality and bring some

[26] To see this in a little more detail, denote total benefits from cooperation (gross of costs) by $B = (1-\alpha) f_N(x^*(\theta), \theta) + \alpha f_E(y^*(\theta), \theta)$. Since (10) binds at 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 this argument out in greater detail.

type of equal opportunity. There are indeed several historical cases where early steps towards equality before the law have been reversed following increases in economic and political inequality, for example, in the Roman Republic and medieval Venice (see, e.g., Acemoglu and Robinson, 2012, and Puga and Trefler, 2014). We now show that one type of increase in inequality—where the elite get richer while normal agents do not—makes equality before the law less likely to emerge (and perhaps harder to maintain) in our model.

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

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

$$u_i(e_i + f(x_j)) - x_i.$$

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

Proposition 10 *An increase in elites' endowments e_E leads to lower normal and elite agent effort. Formally, x^{EE}, x^{EL}, y^{EE}, and y^{EL} are nonincreasing in e_E.*

The intuition for this result can be seen from Figure 1. The modified problem here again generates a set of convex indifference curves, and an increase in elites' endowments has no effect on the constraint set but rotates the indifference curves clockwise, decreasing both elite and normal agent effort, and consequently reducing equality before the law

Proposition 10 focuses on a rise in "inequality" driven by an increase in elite endowment with the endowment of normal agents remaining constant. What happens if simultaneously the endowment of normal agents, e_N, declines? It turns out that the implications of this change are ambiguous: on the one hand, with a lower endowment, normal agents work harder and the greater returns that this creates for the elite discourages them from exerting effort, reinforcing the result in Proposition 10. On the other hand, with a lower endowment, the sensitivity of normal agents' effort to elite effort increases and this might encourage the elites to increase their effort. Nevertheless, it can be

shown that if u_N is not very concave, this second effect is dominated and thus the same result as in Proposition 10 applies when we consider a simultaneous increase in elite endowment and decrease in normal agent endowment.

What about the effect of increasing elites' endowments on equality before the law? Recall that ρ is defined so that the elite incentive compatibility constraint, now given by

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

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

6.5 Comparative Statics for the Size of the Elite

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

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

To see the intuition for this result, 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.[27] 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 (6) in addition to the indirect effect. Finally, note that the overall effect of an increase

[27] 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.

in α on x is ambiguous, because the direct, negative effect on x may be offset by the indirect, positive effect coming through the increase in y.

7 Private Benefits of Cooperation

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

Specifically, we analyze the random matching model described at the end of Section 2, where effort decisions are taken non-anonymously. In every period, agents first randomly match in pairs and observe their partner's status (normal or elite) and then exert effort (which disproportionately benefits one's partner), and then each elite agent has the option of punishing her partner. Note that any equilibrium of this non-anonymous random matching model in which players do not condition their effort choices on their partners' status reduces to an equilibrium of the anonymous random matching model—and thus an equilibrium of our baseline, centralized model—so the non-anonymous random matching model is effectively a generalization of the baseline model. In this section, we show how this generalization affects the structure of incentives, and we establish that our most important comparative static result generalizes to this environment: a reduction in coercive capacity g encourages greater equality before the law.[28]

To model the fact that cooperation imposes positive externalities on society without being a pure public good, we assume that a fraction $1-\lambda \in [0,1]$ of the benefits of cooperation accrue only to one's partner rather than to society at large. Thus, $\lambda = 0$ corresponds to pure private goods (i.e., cooperation generates no positive externalities), and $\lambda = 1$ corresponds to pure public goods (and is thus identical to our baseline environment). Formally, when player i chooses effort x_i, her

[28] 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.

partner chooses effort x_j, and the distributions of effort levels among normal agents and the elite are, respectively, F_N and F_E, player i's stage payoff is

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

if her partner is normal, and

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

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.

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

Proposition 12 *Under endogenous equality before the law, suppose the elite-optimal level of equality before the law 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 $d\rho^*/dg \leq 0$.*

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

8 Heterogeneous Elites

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

8.1 Heterogeneous Productivity within the Elite

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

Formally, we assume there are two elite types that differ according to a productivity parameter b: an elite agent with productivity b who exerts effort y generates output $f_E(by)$. Fraction α_H of the population are (high-productivity, commercial) elites with productivity b_H, and fraction α_L of the population are (low-productivity, landed) elites with productivity $b_L \leq b_H$. We assume that an individual's elite status and output are observable, but her productivity is unobservable. Thus, members of the two elite subgroups cannot be asked to produce different output levels, since otherwise each would pretend to be a member of the group that produces less output.[29] If the equilibrium effort level of high-productivity elites is y_H, then the equilibrium effort level of low-productivity elites is $(b_H/b_L)\, y_H$. Noting that all elites produce output $f_E(b_H y_H)$, the resulting incentive constraints are

$$\begin{aligned} x &\leq \delta\left[(1-\alpha_H-\alpha_L) f_N(x) + (\alpha_H+\alpha_L) f_E(b_H y_H)\right] + g \\ y_H &\leq \delta\left[(1-\alpha_H-\alpha_L) f_N(x) + (\alpha_H+\alpha_L) f_E(b_H y_H)\right] + \rho g \\ \frac{b_H}{b_L} y_H &\leq \delta\left[(1-\alpha_H-\alpha_L) f_N(x) + (\alpha_H+\alpha_L) f_E(b_H y_H)\right] + \rho g. \end{aligned} \tag{13}$$

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

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

[29] This is one part of our analysis that does depend on the continuum population assumption: the claim in the text is clearly true with a continuum, but would require more careful justification with a finite population.

maximization problem:

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

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

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

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

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

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

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

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

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

8.2 Enforcement Hierarchy

Suppose again that there are two elite groups, now corresponding to "minor elites" (say barons) and more powerful, "major elites" (say dukes). These two groups are now equally productive but differ in their vulnerability to coercion. Specifically, suppose that—in the absence of equality before the law—normal agents are vulnerable to coercion from both types of elites, while minor elites (type 1

elites) can be coerced by more powerful elites (type 2 elites), and the latter are initially completely immune to coercion. The level of equality before the law $\rho \in [0,1]$ 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. The resulting incentive constraints are

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

Intuitively, as ρ increases, this closes both the gap in privilege between normal agents and elites as a whole and the gap between the minor and major elites

With two different elite incentive constraints, the issue of what level of equality before the law is optimal for the elites is delicate. For example, if either minor elites or major elites could choose both ρ and the resulting equilibrium, they would choose $\rho = 1$ while requiring more effort from the other elite group. These unintuitive possibilities disappear when all three incentive constraints bind, and this is the case on which we focus in this subsection.[30] This focus thus rules out equilibria where one elite group sets a high level of equality before the law to coerce the other elite group while exerting low effort itself. Consequently, for any value of ρ, the two elite groups differ in their vulnerability to coercion, but the full force of the level of ρ that is chosen applies to both groups.

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

$$\begin{aligned} &\max_{\rho\in[0,1]} \begin{array}{c} (1-\delta)\left[\left(1-\alpha_1-\alpha_2\right) f_N(x(\rho))+\alpha_1 f_E(y_1(\rho))+\alpha_2 f_E(y_2(\rho))\right] \\ -\beta\left(\rho\alpha_1+\alpha_2\right) g-(1-\beta)\rho\left(\alpha_1+\alpha_2\right) g \end{array} \\ = &\max_{\rho\in[0,1]} \begin{array}{c} (1-\delta)\left[\left(1-\alpha_1-\alpha_2\right) f_N(x(\rho))+\alpha_1 f_E(y_1(\rho))+\alpha_2 f_E(y_2(\rho))\right] \\ -\rho\left(\alpha_1+\alpha_2\right) g-\beta\alpha_2 g+\beta\rho\alpha_2 g \end{array}. \end{aligned}$$

It is straightforward to see that this objective function is supermodular in (β, ρ). Hence, the set of optimal values of ρ is increasing in β in the strong set order. Thus, when minor elites have more political power, the resulting level of equality before the law is higher. The intuition is that since minor elites are already exposed to coercion by major elites, greater equality before the law

[30] The assumptions that $\delta f_N'(0) > 1$ and $\delta f_E'(0) > 1$ and that these functions are concave and bounded imply that there exists a positive vector (x_1, y_1, y_2) where all three constraints bind. The concavity of these functions also implies that there is only one such vector.

increases the effort of major elites by relatively more than it increases the effort of minor elites. This makes minor elites more inclined to favor equality before the law. Thus, an increase in minor elites' political power leads to greater equality before the law.

This comparative static result, like the one with respect to α discussed above, is related to North, Wallis and Weingast's (2009) argument that rule of law among the elite is a precursor to the emergence of equality before the law for all agents. Consistent with this comparative static (and with North, Wallis and Weingast), several historical episodes support the notion that political changes that strengthen minor elites encourage greater equality before the law. For example, the Magna Carta was an agreement imposed by barons on King John in 1215, limiting his powers and ability to act without the approval of the barons. But the final charter was formulated as a concession from the king "to all the free men of our kingdom", and went so far as to restrict the ability of landowners to impose forced labor on their own serfs (see Holt, 2015, and the discussion in Acemoglu and Robinson, 2018). Our extension in this subsection is a simple formalization of these ideas: as the political power of "minor elites" increases relative to that of more powerful elites, this encourages an extension of equality before the law for all agents in society.

9 Conclusion

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

productive effort; on the other, it privileges elites at the expense of normal agents.

The most important part of our analysis concerns situations where the elite can choose between elite enforcement and various degrees of "equality before the law", which in our model is interpreted as the elite also being subject to coercive punishments for breaking the law. We show that it may be optimal—even from the viewpoint of the elite—to introduce full equality before the law, which combines high coercion with low inequality. The key mechanism is that by stripping the elite of their privileges, equality before the law enhances the carrot of future cooperation for normal agents. This encourages normal agents to exert greater effort, which can benefit everyone in society, including the elite. Interestingly, we show that equality before the law also leads to low inequality—in the case of our baseline model with pure public goods, complete equality—in that elites exert the same level of effort and receive the same utility as normal agents.

What factors encourage the emergence of equality before the law? We first show that a decline in the extent of coercive punishments that elites can impose on citizens favors equality before the law. 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 the citizens resulting from democratization, to social changes that make certain harsh punishments simply unacceptable (as emphasized by Elias, 1994, and Pinker, 2011). 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 levels of effort from the elite, and thus towards greater equality before the law. We also show that a direct increase in the political power of normal agents has a similar effect. We then establish that an increase in marginal returns to effort (but not average returns) also leads to greater 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—leading to equality before the law. We also explore the implications of economic inequality for equality before the law, and show that when the elite become richer, this may discourage them from exerting additional effort and thus hinder the emergence of equality before the law. When the elite are heterogeneous in terms of their economic investments and productivity (e.g., divided between landowners and commercial interests), a strengthening of more productive segments of the elite also favors greater equality before the law. Finally, consistent with the emphasis of North, Wallis and Weingast (2009), we show that various changes encouraging "rule of law among the elite"—resulting either from an increase in the size of the elite or a change in the balance of power within a heterogeneous elite towards its weaker members—encourage greater equality before the law as well.

Many interesting areas remain to be explored. First, several important extensions of our framework would be interesting to study. These 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. Second, it could be fruitful to apply similar ideas to the internal organization of firms. A key aspect of organizations that has received much less attention than others in the economics literature is the balance of power between "management" and "workers". Tilting this balance in a way that induces managers to exert more or better effort can then incentivize workers, either via repeated game incentives or gift-exchange type considerations. The analogue of our comparative static with respect to coercive capacity here might be studied by considering changes in societal values, social norms, and institutions that make it less acceptable for managers to ask for certain actions from their employees. There are interesting issues to consider in this context. These include the effect of exit options and markets on the choice internal organization, as well as what aspects of firm architecture affect the balance between management and workers. Yet another direction in this context might be to merge a model of labor coercion as in Acemoglu and Wolitzky (2011) with repeated game considerations, so that the carrot of future cooperation interacts with coercive behavior by employers.

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

Appendix: Proofs

Proof of Proposition 1

If x is an equilibrium effort level, then

$$f_N(x) - x \geq (1-\delta) f_N(x).$$

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

Proof of Proposition 2

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

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

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 course of punishments. Rearranging this expression yields (3). The argument for (4) is the same, except that an elite agent's minmax payoff is 0 rather than $-g$. Moreover, (3) and (4) are sufficient as well as necessary for (x, y) to be a pair of equilibrium effort levels, because under these conditions grim trigger strategies combined with coercive punishment of any deviator support constant effort at x and y for normal and elite agents, respectively. Finally, it is clear that (3) binds at the optimum, as increasing x increases the objective and also relaxes constraint (4).

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

$$\frac{dx^*(y)}{dy} = \frac{\delta\alpha f'_E(y)}{1-\delta(1-\alpha) f'_N(x^*(y))}.^{31} \tag{14}$$

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

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

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

Proof of Proposition 3

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

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

$x = \delta(1-\alpha) f_N(x) + g)$, and

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

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

Proof of Proposition 4

If the solution to the elites' problem involves $\rho^* = 0$, the problem reduces to that under elite enforcement. If instead $\rho^* > 0$, then (10) binds by the assumption that ρ^* is minimal. As (9) 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 (8), subject to $y < \bar{y}^{EL}$. Hence, y^{EL} must satisfy the first-order condition (11) derived in the proof of Proposition 2. ■

Proof of Proposition 5

First, note that $\left(x^{EL}, y^{EL}\right) \geq \left(x^{EE}, y^{EE}\right)$, with strict equality if $\rho^* > 0$. To see this, note that x^{EE} 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 $y^{EL} \geq y^{EE}$ is similar.

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

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

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

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

Proof of Proposition 6

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

Next, let μ_N be the Lagrange multiplier on (3). 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 (10), when $y \in \left(\bar{y}^{EE}, \bar{y}^{EL}\right)$. In this case, implicitly differentiating (10) 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 $y \leq \bar{y}^{EE}$ and $\rho^* = 1$ when $y = \bar{y}^{EL}$, this implies that ρ^* is everywhere nonincreasing in g. ■

Proof of Proposition 7

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

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

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

Proof of Proposition 8

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

Proof of Proposition 9

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

As (3) binds at the optimum,

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

To simplify notation, let $f^N = f_N(x^*(\theta), \theta)$ and let $f^E = f_E(y^*(\theta), \theta)$. Totally differentiating with respect to θ yields

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

Recall that $1 > \delta(1-\alpha) f_x^N$ (because (3) 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 (11) using this notation, we have

$$\alpha f_y^E + \delta(1-\alpha) f_x^N = 1. \tag{16}$$

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 (10). 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

$$\frac{dy^*}{d\theta} - \delta\left[(1-\alpha)\left(f_x^N \frac{dx^*}{d\theta} + f_\theta^N\right) + \alpha\left(f_y^E \frac{dy^*}{d\theta} + f_\theta^E\right)\right]. \tag{17}$$

Note that by (15),

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

Moreover, by (16),

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

Hence, (17) 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 (16). Hence, $\frac{dy^*}{d\theta} \geq 0$ and $\delta < 1$ imply $\frac{d\rho^*}{d\theta} \geq 0$. ■

Proof of Proposition 10

The elites' problem in this case becomes

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

subject to

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

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

$$\frac{dx^*}{dy} = \frac{\delta\alpha u_N'(e_N + f_E(y))f_E'(y)}{1-\delta(1-\alpha)u_N'(e_N + f_N(x))f_N'(x)}.$$

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

$$(1-\alpha)u_E'(e_E + f_N(x^*(y)))f_N'(x^*(y))\frac{dx^*}{dy} + \alpha u_E'(e_E + f_E(y))f_E'(y) = 1$$

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

Proof of Proposition 11

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

$$\max_{y\in[0,\bar{y}]} (1-\alpha)f(x^*(y,\alpha)) + \alpha f(y) - y, \tag{18}$$

where $x^*(y,\alpha)$ is the value of x that makes (3) hold as equality when the fraction of elite agents is α. We now show that the solution to (18) 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 $y \in \left(\bar{y}^{EE}, \bar{y}^{EL}\right)$, and hence (10) binds, we have

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

We may thus rewrite (10) as

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

Implicitly differentiating yields

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

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

Proof of Proposition 12

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

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

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*

$$\begin{aligned}
(1-\delta\alpha)\, w + \delta\alpha x &\leq \delta B_N(w,x,y,z) + \delta\alpha g && (19)\\
(1-\delta(1-\alpha))\, x + \delta(1-\alpha)\, w &\leq \delta B_N(w,x,y,z) + (1-\delta(1-\alpha))\, g && (20)\\
(1-\delta\alpha)\, y + \delta\alpha z &\leq \delta B_E(w,x,y,z) + \delta\alpha\rho g && (21)\\
(1-\delta(1-\alpha))\, z + \delta(1-\alpha)\, y &\leq \delta B_E(w,x,y,z) + (1-\delta(1-\alpha))\,\rho g. && (22)
\end{aligned}$$

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

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

Hence, a normal agent's equilibrium payoff is

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

and elite agent's equilibrium payoff is

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

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

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

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

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

as in this case a normal agent can be punished in the current period. This rearranges to (20). The argument for elite agents is similar, noting that an elite agent's minmax payoff is $-\rho\alpha g$ rather than $-\alpha g$. ■

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

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

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

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

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

$$\begin{aligned} 1 - \lambda\alpha f_E'(y^*) &\geq 0, \quad (23) \\ 1 - (1 - \lambda(1-\alpha)) f_E'(z^*) &\geq 0. \quad (24) \end{aligned}$$

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

To derive the desired inequalities, let $u_E^{EL}(g)$ be the value of the elites' problem for parameter g. Note that $u_E^{EL}(g)$ is a concave function of g. To see this, suppose (w, x) is a solution given coercive capacity g and (w', x') is a solution given coercive capacity $g' > g$. Then, for all $\beta \in (0, 1)$, $(w^*, x^*) = (\beta w + (1-\beta)w', \beta x + (1-\beta)x')$ is feasible given coercive capacity $\beta g + (1-\beta)g'$ (as f_N and f_E are concave), and elite utility at (w^*, x^*) is greater than the β-weighted average of elite utility at (w, x) and (w', x').

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

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

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

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

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

Finally, suppose that both (19) and (20) bind. In this case, $g = x-w$, so substituting $\delta\alpha(x-w)$ for $\delta\alpha g$ in (19) and (20) lets us rewrite the elite's problem as

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

subject to

$$x = \delta\left[(1-\lambda\alpha)\left[(1-\alpha)f_N(x-g)+\alpha f_E(y)\right]+\lambda\alpha\left[(1-\alpha)f_N(x)+\alpha f_E(z)\right]\right]+g. \tag{25}$$

Let $\mu_{NE} \geq 0$ be the multiplier on (25). 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 $dw^*/dg \geq 0$), $dy^*/dg \leq 0$, $dz^*/dg \leq 0$, (23), and (24). ■

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 to Liberty: The Red Queen and the Struggle of State against Society*. Book manuscript, 2018.

[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". Working paper, 2018.

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

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

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

[13] 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.

[14] Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. "Top Incomes in the Long Run of History". *Journal of Economic Literature* 49.1 (2011): 3-71.

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

[16] 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.

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

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

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

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

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

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

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

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

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

[26] Boehm, Christopher. *Moral Origins: The Evolution of Virtue, Altruism, and Shame.* Basic Books, 2012.

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

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

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

[30] Buruma, Ian. *Inventing Japan: 1853-1964.* Modern Library, 2003.

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

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

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

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

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

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

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

[38] Farrell, Joseph, and Eric Maskin. "Renegotiation in Repeated Games". *Games and Economic Behavior* 1.4 (1989): 327-360.

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

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

[41] 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.

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

[43] 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.

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

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

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

[47] 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.

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

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

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

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

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

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

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

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

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

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

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

[59] Kern, Fritz. *Kingship and Law in the Middle Ages.* Translated by S. B. Chrimes, The Law Book Exchange Ltd., New Jersey, 1956.

[60] LeBlanc, Steven A., and Katherine E. Register. *Constant Battles: Why We Fight.* Macmillan, New York, 2004.

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

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

[63] 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.

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

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

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

[67] 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.

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

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

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

[71] 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.

[72] 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.

[73] 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.

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

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

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

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

[78] 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.

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

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

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

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

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

[84] Scheidel, Walter. *The Great Leveler: Violence and the History of Inequality from the Stone Age to the Twenty First Century.* Princeton University Press, Princeton New Jersey, 2017.

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

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

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

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

[89] Van Damme, Eric. "Renegotiation-Proof Equilibria in Repeated Prisoners' Dilemma". *Journal of Economic Theory* 47.1 (1989): 206-217.

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

[91] White, Lynn. *Medieval Technology and Social Change.* Clarendon Press, Oxford, 1962.

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

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

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

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

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

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