The Paradox of Civilization
Pre-Institutional Sources of Security and Prosperity

Ernesto Dal Bó   Pablo Hernández-Lagos   Sebastián Mazzuca*

February 2019

Abstract

The rise of civilization involved surplus production (“prosperity”) and states that could protect surplus (“security”). But the security-prosperity combination posed a paradox: prosperity attracts predation, which discourages the investments that create prosperity. Drawing from the anthropological and historical literatures, we model the trade-offs facing a proto-state on its path to civilization. We emphasize pre-institutional forces, such as the geographical environment, that shape growth and defense capabilities. We characterize conditions on these capabilities that help escape the civilizational paradox, and provide narrative and quantitative illustration of the model by analyzing the rise and fall of Old World Bronze Age civilizations, with special focus on Egypt and Sumer.

*Dal Bó: UC Berkeley and NBER. Hernández-Lagos: NYU AD. Mazzuca: Johns Hopkins. We thank Robert Allen, Alberto Bisin, Pedro Dal Bó, David de la Croix, Oded Galor, Andrew Hall, Santiago Oliveros, Ben Olken, Demian Pouzo, Robert Powell, Alvaro Sandroni, David Schönholzer, Andrew Schotter, Enrico Spolaore, David Stasavage, Mathias Thoenig, Thierry Verdier, Etienne Wasmer, David Weil and seminar and conference participants at Berkeley, Bocconi, Brown, NBER Political Economy Program, Northwestern, Pompeu Fabra, Princeton, Stanford, UCLA, and Yale for helpful comments.
1 Introduction

Civilization first materialized around 5,000 years ago through fundamental transformations: economic surplus, urbanization, public infrastructure, writing, and states. Although the rise of civilization is arguably more of a qualitative change than the Industrial Revolution, modern political economy has paid much less attention to it.

According to an influential view in archaeology, civilization was driven by favorable endowments and technologies for food production. For V. Gordon Childe (1936), the key features of Lower Mesopotamia, the “cradle of civilization,” were an extremely fertile alluvial soil, an abundance of edible animals, and irrigation technology. Identical factors were emphasized for the rise of Egypt, the first pristine civilization after Sumer. Both in Lower Mesopotamia and Egypt “irrigation agriculture could generate a surplus far greater than that known to populations on rain-watered soil” and “as productivity grew, so too did civilization” (Mann 1986: 80, 108).

Without a surplus above subsistence needs, it was not possible to fund the visible aspects of civilization. However, surplus production was only a necessary condition for civilization, not a sufficient one. In fact, prosperity could attract trouble. Primitive food producers were surrounded by nomadic tribes for whom agricultural surpluses were a most tempting target. The clash between sedentary, agricultural food producers and nomadic pastoralist raiders is a primordial conflict shaping the civilizational process. According to McNeill (1992: 85), “Soon after cities first arose ... the relatively enormous wealth that resulted from irrigation and plowing made such cities worthwhile objects of attack by armed outsiders.” For anthropologists, intergroup violence had been prevalent since before civilization (Keeley 1996), but “the greater the surplus generated, the more desirable it was to preying outsiders” (Mann 1986: 48).

Since civilization entailed the joint achievement of prosperity and security, its emergence is paradoxical in so far as the two outcomes were in tension. In order to reach civilization a primitive society with the capacity for surplus production had to overcome the increased threat of predation associated with prosperity. The conditions leading to prosperity and security are rare, as evidenced by the fact that, out of many primitive societies, only a handful could develop independent civilizations, starting with Sumer and Egypt. These first successes involved a largely pre-institutional process, rooted in tangible assets and technolo-
gies of economic or military nature rather than formal political rules. Pristine civilizations emerged in areas with favorable conditions for food production, and the salient man-made contributions to civilization were productive and defense equipment. Two engineering accomplishments are the mark of pristine Old World civilizations: water management, and perimetral walls in cities throughout Mesopotamia, the Levant, China and the Indus Valley. Each public good had a single, well-defined mission: surplus production and surplus protection. The prominence of the two types of public works reflects the centrality of the production-protection tension in the process of civilization building.

In this paper we develop a model to identify conditions for the rise of civilization under the perspective of the civilizational paradox. Our analysis of the historical record relies both on narratives and quantitative information on the locations of ancient Egyptian and Sumerian cities. These historical cases illustrate the logic of the model, and the model allows for a richer interpretation of the cases. In addition to their historical preeminence as the earliest civilizations, Sumer and Egypt provide evidence that the potential for surplus emphasized by archaeologists was only half of the civilization story. The other half was surplus protection, which occurs in two contrasting ways—defense can be natural as in Egypt, or man-made as in Sumer. The model implies an empirical profile that we illustrate through historical narratives and quantification exercises. The empirical patterns extend to the other two pristine Bronze Age civilizations in the Old World, China and the Indus Valley.

Our pre-institutional theory on the joint achievement of security and prosperity can shed light on the problem of state formation more generally. A broader goal is to understand episodes in which a potentially prosperous region, being surrounded by predatory threats, may flourish, or alternatively fall in the traps of security-preserving stagnation or self-defeating prosperity. These dangers, as well as the role of natural and man-made defenses that we emphasize throughout the paper, likely operated in many historical contexts. For example, Thucydides, who is recognized as the first critical observer of history, characterized the Greeks preceding the ascent of Athens as living by “...cultivating no more of their territory than the exigencies of life required, destitute of capital, never planting their land (for they could not tell when an invader might not come and take it all away, and when he did come they had no walls to stop him)… consequently neither built large cities nor attained to any other form of greatness. The richest soils were always most subject to this change of masters...” (Thucydides 431BC Book 1 [Warner, 1971: 35-36]).
The class of interactions fraught by the tension between prosperity and security includes those between a large number of proto-cities and barbarian invaders from the steppes across the Eurasian continent throughout the Middle Ages; the struggles in 19th century Latin America between elites from port-cities engaged in nation-building and rural warlords, “caudillos”; as well as contemporary state-building efforts in failed states of Sub-Saharan Africa and the Middle East, in which international economic aid, if not coupled with military buildup, may have counter-productive effects by inducing voracity from predatory actors. Echoing concerns in history and anthropology about the reversibility of gains in social complexity, our theory also provides an account for civilization collapse, seen as economic and military reversals in societies that had achieved prosperity and security. For illustration, we use the model to account for the End of the Bronze Age, a much-debated process in which dozens of civilization centers collapsed quickly throughout the Eastern Mediterranean, ushering in the first “dark ages” in the historical record.

Overview of the model and historical illustrations

In our model, an agent (the “incumbent”) may invest and grow future income, which would lead to “prosperity,” but faces potential attacks by a predator (the “challenger”). The possibility of attacks may induce the incumbent to spend resources in consumption and defense instead. Three key parameters govern tradeoffs: initial income, and the respective rates at which income can be turned into defense (defense capability) and future income (growth capability). Productive investment increases future income, but it also attracts stronger predation, and creates a tradeoff between investment-led growth and security. If sufficient defense can be financed, the challenger is deterred (“security” is attained) and investments are safe. Security, however, costs current consumption. These indirect defense costs may make investment too expensive to be desirable. The key formal question is whether some combination of parameter values allows for both security and investment to occur, yielding a civilizational breakthrough. Previous work, which we discuss in Section 5, has analyzed related problems; but the choice between consumption, productive investment, and defense over multiple periods is a natural problem that, to the best of our knowledge, has not been addressed.

In the first part of our analysis the incumbent’s defense capability is exogenous, and in the second part defense capability can be improved. Both parts help rationalize different modalities in the rise of ancient civilizational states. They also offer ways to think about short-run (when defense capabilities are fixed) and long-run scenarios (when defense capabilities are
The analysis in the first part characterizes the unique equilibrium of the game. When both defense and growth capabilities are low, neither prosperity nor security are possible, and societies remain economically stagnant and mired in conflict, as characterized by Keeley (1996), a situation mirroring the Hobbesian “state of nature.”\(^1\) If growth capability is high relative to defense capability, prosperity becomes possible even in the face of attacks. That prosperity may be short-lived since attacks may defeat the incumbent. While anti-Hobbesian, the possibility of growth despite predation is consistent with a widespread occurrence in history, like the Chinese with the Mongolians and the Saxons with the Vikings in the 10th century. Lastly, when both defense and growth capabilities are high and “balanced,” the incumbent can grow and also deter predators. The latter two cases explain the emergence of civilizations. Civilizations occur where high enough returns to productive investment allow the economy to grow, and where the incumbent manages to deter the challenger or defeat it with high probability.

The case of Egypt can be explained in terms of high natural endowments for both growth and defense. Growth capabilities were given by rich alluvial soils that could be improved through investments in water management, and defense was provided by the surrounding deserts, which protected dwellers along the Nile from most types of attack (Bradford 2001).

The rise of civilization in Southern Mesopotamia poses a challenge to our baseline model, however, because the Sumerian settlements did not have natural protection and faced challenges from pastoralist groups. How could the Sumerian city-states ever emerge? The anthropological literature suggests that the groups who formed pristine states exploited an agrarian “staple finance”, which, being highly rewarding, would fund their defense (Johnson and Earle 2000: 305-306). These groups could turn a material advantage into a military one, by relying on walls, weaponry, and numbers, all of which could be used to deter or defeat their enemies. The improvement of defense capabilities can be accounted for in our extended model, where the incumbent can make investments to change its defense technology.

We show that when initial income is high enough, the incumbent can fund its way out

\(^1\)The relationship between security and prosperity has been a perennial concern in the social sciences. A dominant view, inspired by Hobbesian philosophy, is that state-provided security is a precondition for prosperity (Lane 1958; Olson 2000; Bates 2001; see Boix 2015 for a contrasting approach). But the state itself has to be explained and the Hobbesian view provides no clear message on whether state formation requires a modicum of prosperity in the first place.
of the parametric region without security or prosperity, and move into a region with high levels of both, a transition that is easier when growth capabilities are also high. The result is not obvious: investments both productive and defensive confer future advantage but incite immediate predation. Investment in defense occurs in equilibrium due to an intertemporal complementarity with productive investment. If enough defense capability is built today, enhanced security will increase effective returns to, and induce the making of, productive investments tomorrow. The intertemporal pie will grow, justifying defense investments today.

The broad theoretical conclusion is that favorable conditions such as high initial income and growth capabilities cannot on their own produce civilization. In a world where security is a concern, growth capabilities need to interact with high natural defenses, or high initial income to fund artificial defenses. We relate the predictions of our theory to the historical illustration through a quantitative exercise. We divide the world in 1/5th of a degree grid cells (approximately 22kms at the Equator); then, following the historical narrative, we proxy each location’s defense capability as the share of surrounding cells that are occupied by desert, and the growth capability as the differential in an income index when rain-fed agriculture is replaced with irrigation agriculture. Using these measures, we can render empirically the parameter space of the theory. We then locate, in the empirically-proxied parameter space, ancient Egyptian and Sumerian cities as captured in the well-known database by Modelski (1999, 2003).

The quantitative illustration confirms the message from the theory and the historical narratives. Egyptian cities fall in a statistically rare area boasting relative high levels of both growth and defense capabilities (which makes Egypt’s case for civilization viable) while Sumerian cities display lower defense capabilities than Egyptian ones. However, Sumerian cities were disproportionally located in areas that satisfy the joint requirement of high initial income and high growth capabilities. The same pattern holds when we extend attention to the additional two pristine Old World Bronze Age civilizations in the Modelski data. Both the Indus Valley and the Chinese civilizations emerged in sites where substantial investments were made, both military and productive. The quantitative illustration confirms that these locations, like those in Sumer, lacked substantial natural defense, but satisfied the simultaneous requirement of high initial income and high growth capabilities.

**Plan for the paper** In the next section we present the theory, and provide historical illustrations in Section 3. In Section 4 we use the model to rationalize the end of the Bronze
Age. We discuss related literature in Section 5, and conclude in Section 6. Proofs are relegated to the Appendix.

2 Theory

2.1 Basic Model

2.1.1 Setup

Players An “incumbent” controls a productive asset that yields a nonstorable income flow $v_t > 0$ in each of two periods, $t = 1, 2$. The asset can be any bundle of productive resources, including land, a port, and people. The initial level $v_1$ tracks properties of the environment (e.g., climate, quality of the soil) that affect productivity. The incumbent faces a “challenger” in each period, who is interested in gaining control of the asset.

Actions, resources, technology In each period $t$ the incumbent can spend $v_t$ on consumption, productive investment, or mobilizing resources for defense. One unit of productive investment $i_t$ costs one unit of income and it adds $\rho > 1$ units to the future yield of the asset.\footnote{Given that (with linear preferences) investment is never worthwhile if $\rho < 1$, failure to obtain it in equilibrium is inevitable and uninteresting. Hence our assumption $\rho > 1$ which makes investment possible (though not inevitable, due to insecurity).} That is, income evolves according to the relation $v_{t+1} = v_t + \rho i_t$; we abstract from depreciation and discounting for simplicity. $\rho$ captures anything that affects the returns to investments in the asset; like $v_1$, $\rho$ could reflect conditions of the physical environment, and economic factors such as the price of goods sold.\footnote{If the value of what the incumbent produces follows a standard price $\times$ quantity formulation we can write $v_1 = pq$, and $v_2 = pq + \rho' pi \equiv v_1 + pi$, where $q$ and $i$ are physical units. Then, changes in $p$ cause changes in both $v_1$ and $\rho$. Changes in the baseline physical capacity of production $q$ are captured through changes in $v_1$, and changes in the physical returns to investment as changes in $\rho'$ and therefore $\rho$.}

The effectiveness of the incumbent’s defense (or “army”) is denoted $a_t$ and such an army costs the incumbent an amount $\frac{a_t}{\kappa_t}$ where $\kappa_t \geq 0$ is the value of the incumbent’s defense capability. The higher the defense capability of the incumbent, the higher the “firepower” $a_t$ attained by a given conflict expense $\frac{a_t}{\kappa_t}$. The parameter $\kappa_t$ captures anything that affects the costs for the incumbent of producing defense, such as a surrounding desert, better military technology, or expertise. We fix $\kappa_t = \kappa_1$ in our baseline analysis and later endogenize $\kappa_t$ in subsection 2.2.
Timing In each period the incumbent acts as a Stackelberg leader, moving first to choose \( a_t \) and \( i_t \). After observing \((a_t, i_t)\) the challenger selects its own conflict effort \( b_t \). If \( b_t > 0 \), then there is conflict at the end of period \( t \). The winner appropriates the asset, and hence the income it generates in the next period.\(^4\) Whenever the challenger attacks \((b_t > 0)\), it prevails with probability \( \frac{b_t}{a_t + b_t} \) and it gains nothing with the complementary probability (i.e., we adopt the typical Tullock contest success function). If the incumbent does not attack or is defeated, it obtains a payoff of zero. If the challenger selects \( b_t = 0 \) we say the incumbent has successfully deterred the challenger, and this lack of challenge results in full security.

Payoffs Both challenger and incumbent are risk neutral and care linearly about consumption, which equals income net of costs (of defense for both players, and of investment for the incumbent). The incumbent chooses \( a_t \) and \( i_t \) to maximize the value of his expected intertemporal consumption \( V_t = v_t - \frac{a_t}{\kappa_t} - i_t + \frac{a_t}{a_t + b_t}V_{t+1} \), while observing the budget constraint (or non-negative consumption condition), \( v_t - \frac{a_t}{\kappa_t} - i_t \geq 0 \). The challenger chooses \( b_t \) to maximize the value of her own expected intertemporal consumption \( \frac{b_t}{a_t + b_t}V_{t+1} - b_t \).

Solution concept We solve for a Subgame Perfect Nash Equilibrium by backward induction.

2.1.2 Solution

Second period The rewards from success in conflict accrue one period later, so the challenger does not fight in the second and last period, leaving \( b_2^* = 0 \) (asterisks denote equilibrium choices). Anticipating this, the incumbent chooses \( a_2^* = 0 \). Since the proceeds from productive investment only materialize in the next period, the incumbent selects \( i_2^* = 0 \) yielding \( V_2 = v_2 = v_1 + \rho i_1 \).

First period The challenger observes the pair \((a_1, i_1)\) and chooses \( b_1 \) to maximize \( \frac{b_1}{a_1 + b_1}v_2 - b_1 \). Since the first order condition is \( \frac{a_1}{(a_1 + b_1)^2}v_2 = 1 \), the best response function of the challenger is, \( b_1(a_1, v_2) = \sqrt{a_1v_2} - a_1 \) if \( a_1 < v_2 \) and zero otherwise.

Since \( v_2 = v_1 + \rho i_1 \), the best response \( b_1(a_1, v_2) \) exhibits a key trade-off of the model. Productive investments \( i_1 \) raise the value of the productive asset since \( \rho > 1 \). Thus, conditional on maintaining control of the asset, investment is a good idea for the incumbent.

\(^4\)In the two period model it is immaterial whether expropriation involves the income flow or the asset itself. Both cases were observed historically: intermittent raids, and invasion with “replacement,” such as when Sargon of Akkad took over the Sumerian city-states, or the Mongols invaded China.
However, investment raises the incentives of the challenger to arm itself since it makes it more attractive to become the incumbent (formally, \( \frac{\partial b_1}{\partial i_1} > 0 \) if \( a_1 < v_2 \)). In sum, while productive investments increase the value of the asset, they lower the chance that the current incumbent gets to reap that value. This is the civilizational paradox: future prosperity raises insecurity, which in turn depresses incentives to invest and undermines the creation of that future prosperity.\(^5\) The resulting question is: are there any parameter values \( v_1, \kappa_1, \) and \( \rho \) that map into security and prosperity?

The incumbent maximizes his initial expected intertemporal consumption \( V_1 \) subject to the budget constraint \((BC)\) \( v_1 - i_1 - \frac{a_1}{\kappa_1} \geq 0 \), and a deterrence constraint \((DC)\) \( a_1 \leq v_2 \) stemming from the facts that the challenger will not fight if \( a_1 \geq v_2 \), and the incumbent does not arm beyond the point that attains deterrence. The incumbent’s problem in period 1 can then be written as,

\[
\max_{a_1, i_1} \left\{ v_1 - \frac{a_1}{\kappa_1} - i_1 + \frac{a_1}{a_1 + b_1(a_1, v_1 + \rho i_1)}(v_1 + \rho i_1) \right\}
\]

subject to,

\[
\begin{align*}
v_1 - \frac{a_1}{\kappa_1} - i_1 & \geq 0 \quad (BC) \\
v_1 + \rho i_1 - a_1 & \geq 0 \quad (DC) \\
a_1 & \geq 0 \\
i_1 & \geq 0.
\end{align*}
\]

The Lagrangian, which expresses the expected utility \( V_1 \) of the incumbent, is:

\[
\mathcal{L} = v_1 - \frac{a_1}{\kappa_1} - i_1 + \frac{a_1}{a_1 + b_1(a_1, v_1 + \rho i_1)}(v_1 + \rho i_1) \\
+ \lambda_{BC}(v_1 - \frac{a_1}{\kappa_1} - i_1) + \lambda_{DC}(v_1 + \rho i_1 - a_1) + \lambda_a a_1 + \lambda_i i_1,
\]

\(^5\)The civilizational paradox, involving as it does the incentives of a challenger, is related to, but differs from, Hirshleifer’s (1991) paradox of power. Hirshleifer’s paradox consists of the fact that the poorer contender can end up better off. We instead use the term “paradox” to denote a tension: investments leading to prosperity reduce security and therefore the motivation to bring about that prosperity.
where $\lambda_{BC}$, $\lambda_{DC}$, $\lambda_a$ and $\lambda_i$ are the Lagrange multipliers for each constraint (2)-(5). We characterize the solution $(a_1^*, i_1^*, \lambda_{BC}^*, \lambda_{DC}^*, \lambda_a^*, \lambda_i^*)$ to this problem for each parameter combination $(\rho, \kappa_1, v_1) \in (1, \infty) \times \mathbb{R}_+ \times \mathbb{R}_+$. The first order and complementary slackness conditions that characterize the optimum are given by,

$$\frac{\partial \mathcal{L}}{\partial a_1} = \frac{1}{2} \sqrt{\frac{v_1 + \rho i_1}{a_1}} - \frac{1}{\kappa_1} - \frac{\lambda_{BC}}{\kappa_1} - \lambda_{DC} + \lambda_a = 0; a_1 \geq 0, \lambda_a \geq 0, \lambda_a a_1 = 0 \text{ c.s.} \quad (7)$$

$$\frac{\partial \mathcal{L}}{\partial i_1} = \frac{\rho}{2} \sqrt{\frac{a_1}{v_1 + \rho i_1}} - 1 - \lambda_{BC} + \lambda_{DC} \rho + \lambda_i = 0; i_1 \geq 0, \lambda_i \geq 0, \lambda_i i_1 = 0 \text{ c.s.} \quad (8)$$

$$\lambda_{BC}(v_1 - \frac{a_1}{\kappa_1} - i_1) = 0 \text{ c.s.}., \lambda_{DC}(v_1 + \rho i_1 - a_1) = 0 \text{ c.s.} \quad (9)$$

The conditions in (7)-(9) are necessary and sufficient for a maximum because the constraints are linear and the objective is concave in the control variables $(a_1, i_1)$. Note from (7) that the marginal benefit of $a_1$ goes to infinity as $a_1$ goes to zero (a typical feature of contests), so the optimum must feature $a_1 > 0$ and $\lambda_a = 0$. Beyond this, the method for solving the problem requires checking which combinations of values for the endogenous variables yield the highest value of the program while being consistent with the constraints in each part of the parametric space. The following proposition summarizes the solution.

**Proposition 1** There is a unique equilibrium, which yields a partition of the parameter space $(\kappa_1, \rho, v_1)$ into four regions:

**Region 1** (R1): $\{(\kappa_1, \rho, v_1)|\kappa_1 > \rho, \rho > \kappa_1/(\kappa_1 - 1), \kappa_1 > 1\}$ Security and prosperity

Solution: \[
\left\{ a_1^* = v_1^{\kappa_1/(\kappa_1 + \rho)}, i_1^* = v_1^{(\kappa_1 - 1)/(\kappa_1 + \rho)}, V_1 = v_1^{\kappa_1/(\kappa_1 + \rho)} \right\}
\]

**Region 2** (R2): $\{(\kappa_1, \rho, v_1)|\rho > \kappa_1, \rho > 4/\kappa_1\}$ Prosperity without security

Solution: \[
\left\{ a_1^* = \frac{\kappa_1 v_1}{2} \left( 1 + \frac{1}{\rho} \right), i_1^* = \frac{v_1}{2} \left( 1 - \frac{1}{\rho} \right), V_1 = \frac{v_1}{2} \left( 1 + \frac{1}{\rho} \right) \sqrt{\rho \kappa_1} \right\}
\]

**Region 3** (R3): $\{(\kappa_1, \rho, v_1)|2 > \kappa_1, \rho < 4/\kappa_1\}$ Neither prosperity nor security

Solution: \[
\left\{ a_1^* = v_1 \left( \frac{\kappa_1}{2} \right)^2, i_1^* = v_1 \left( 1 + \frac{4}{\kappa_1} \right) \right\}
\]

**Region 4** (R4): $\{(\kappa_1, \rho, v_1)|\kappa_1 > 2, \rho < \kappa_1/(\kappa_1 - 1)\}$ Security without prosperity

Solution: \[
\left\{ a_1^* = v_1, i_1^* = 0, V_1 = v_1 \left( 2 - \frac{1}{\kappa_1} \right) \right\}
\]

Panel (a) in Figure 1 contains a graphical representation of the solution. We restrict attention to the bidimensional space $(\kappa_1, \rho)$ because the shape of the four regions is invariant
in $v_1$. One implication of this invariance is that income advantage does not ensure full security. Only a technological advantage at defense does.

A salient feature of the solution is that all four combinations of security and prosperity can be observed depending on the values of the parameters $(\kappa_1, \rho)$. For low values of both defense and growth capabilities, the incumbent will be stuck in a situation of stagnation and conflict (R3). In R3 the prospect of conflict lowers the net return to investment, preventing growth. If defense capability $\kappa_1$ is higher but growth capability $\rho$ is still low, the incumbent will be in region R4, where the challenger is deterred but there is no investment. In this region growth is foreclosed not by existing but by potential conflict: investment would foster challenge and raise the costs of maintaining deterrence. If growth capability is relatively high and defense capability relatively low—i.e., in R2—growth occurs despite the fact that full security is not attained. The high growth capability makes the incumbent willing to invest even if the prospect of a larger income makes the challenger more aggressive.\(^7\) If, starting from R2 or R4, defense capability $\kappa_1$ were to become sufficiently higher, the incumbent would enter R1, featuring both investment and deterrence. Note that strict military superiority ($\kappa_1 > 2$) by the incumbent is needed for complete security (the challenger has a military capacity of 1).

Figure 1(a) highlights “extensive margin” variation, as there is either no growth or some growth, and no insecurity or some insecurity. But the equilibrium magnitudes of growth and security vary continuously. The other panels in Figure 1 display contour plots of relevant equilibrium magnitudes. The continuous lines within each region represent level curves, and the lighter shades of color represent higher values of the respective magnitude. Figure 1(b) shows that since the army $a_1$ of the incumbent increases as defense capability is higher this increases security, proxied by the probability that the incumbent will prevail. This probability is 1 in R4 and R1, and it decreases in R2 as defense capability goes down

---

\(^6\)Initial income does not affect the four regions because (in equilibrium) it does not affect the incumbent’s incentives at the margin. These incentives depend on a ratio involving the prize under dispute (future income) and arming expense (as seen in conditions 7 and 8). Both the prize $v_2$ and arming $a_1$ are proportional to initial income in equilibrium, and income changes become irrelevant.

\(^7\)It may surprise that growth increases with $\rho$ in R2. When $\rho$ goes to infinity, the challenger grows infinitely aggressive, against an incumbent who is resource-constrained and cannot keep up, which suggests investments are almost sure to be lost and should go to zero. However, the relative speed at which gross returns grow relative to the challenger’s threat justify positive investments even at the limit. In R2 growth is unresponsive to defense capability because any increase in $\kappa_1$ is met with a similar increase in $a_1$, which keeps the resources devoted to defense $\frac{a_1}{\kappa_1}$ and investment $i_1$ constant.
Figure 1: Equilibrium with exogenous defense capability. Assumption: $v_1 = 1$. 
or growth capability goes up (as this fires up the challenger). The areas in R2 that are sufficiently close to R1 display very high levels of security (approaching full security where R2 meets R1). But civilization requires more than security; it also requires the creation of surplus, which in our model amounts to growth \( (v_2 - v_1 = i_1\rho) \). Figure 1(c) shows how there is no growth in R3 and R4 (since there is no investment) and that there is growth in R2 and R1. Growth increases with returns \( \rho \) and in R1 it also increases with defense capability, as a higher defense capability lowers the costs of arming and releases resources for investment.

The panels (b) and (c) in Figure 1 show that, all else equal, increases in defense capability increase both security and prosperity, while increases in growth capability help growth but undermine security. Panel (d) of Figure 1 shows the combination of growth and security, as given by the expected continuation value perceived by the incumbent in period 1. This is the value \( EV_2 = \frac{a_1}{a_1 + b_1} v^*_2 \), which involves investment-led future income \( v^*_2 \) and the probability \( \frac{a_1}{a_1 + b_1} \) that the incumbent prevails in equilibrium.

2.1.3 Modeling choices, interpretation, and robustness

Internal vs external conflict and social structure of incumbent polity The civilization process precedes nations and hence distinctions between domestic and external conflict. The classification of challengers as internal vs. external (akin to classifying conflict as about internal order vs. sovereignty) partly depends on whether the eventual civilization incorporated or excluded formerly hostile populations. Both cases occurred in history. Also, we do not distinguish between ruler and subjects within each actor. The incumbent and challenger in our model can be taken to be representative agents of their respective groups, perfectly benevolent rulers acting on their behalf, or perfectly extractive rulers who are residual claimants.

Asymmetries We have kept as many aspects as possible symmetric between the incumbent and the challenger, and only introduced asymmetries that we deemed helpful to

---

8The latter interpretation is more suitable if growth in our model is interpreted in per capita terms, because extraction and concentration of surplus in a few hands would be a way in which an ancient society could escape Malthusian population adjustment. Alternatively we can interpret growth in our model in terms of per working capita, or at the level of groups, where wealth becomes population. For models of conflict over land in prehistoric settings that put Malthusian dynamics front and center, see Dow and Reed (2003) and Dow, Mitchell and Reed (2017).
analyze the type of interaction of interest. The incumbent acting as a Stackelberg leader helps make deterrence possible. Also, while defense effort costs the incumbent resources, it does not deplete a budget for the challenger. This is for tractability. It would be possible to include a budget constraint for the challenger and the advantage of a wealthier incumbent at being able to finance higher defense effort, and then attaining deterrence, would operate in similar fashion. However, the reaction function of the challenger would hit its constraint eventually and have a kink, making the analysis less elegant. A third asymmetry is that the incumbent can accumulate and the challenger does not. This maintains tractability, and matches historical situations where one settled, food-producing, group has more room to accumulate than rival nomadic groups.

Trade and other interactions We do not consider trade between the incumbent and the challenger. Exchange between populations has been documented around the time of the birth of the first cities (see Mann 1986: 79). In this paper we seek to understand the interaction between prosperity—as driven only by investment, also the focus of traditional growth models—with security. The way this interaction is further affected by the possibility of exchange is an important next step for future research (Martin et al. (2008), for instance, address trade and war in modern states). Similarly, we do not consider potential interactions between multiple incumbents (food producers who can accumulate), which are also part of the complex civilizational process. These are certainly limitations in the current analysis. Our model has empirical traction to the extent that (i) conflict among incumbents involves asymmetries similar to those in our model or (ii) however such conflict is resolved, civilization still requires prevailing against non-accumulating challengers.

Dynamic considerations Our basic model has two periods. Would things change qualitatively if more periods were considered? With two periods, it does not matter whether the challenger is different across periods or not, nor whether the challenger aims to steal the flow income or permanently replace the incumbent. An extension to three periods is direct if a different challenger arrives each period who only aims to steal the income flow. If the challenger aims to replace the incumbent, an extension to three periods is not trivial for at least two reasons. First, the challenger now cares directly about $\rho$ as he may want to invest upon becoming the incumbent, and he values the asset for its future, as well as current, yield. In addition, with a longer horizon the incumbent may want to invest in order to be able to finance a larger army in the future and attain more security. We show in the Online
Appendix A that an extension to more periods preserves our qualitative results.

The use of contests Like many authors before us, we use a Tullock contest and abstract from transfers.\(^9\) As is well known, even when transfers are possible inefficient conflict may occur, for example due to inconsistent priors across players, agency issues, commitment problems, or asymmetric information (see Jackson and Morelli (2011) for a survey of reasons for war). Every paper on conflict must take a stand on whether or not to microfound conflict onset by reference to one of those phenomena. The cost of the added structure is justified when a specific distortion responsible for conflict is particularly likely or germane given the problem under investigation. When the researcher remains agnostic about such connections, the more parsimonious approach that we take seems justifiable.

Destruction from conflict Our results do not change qualitatively if conflict generates some destruction.\(^{10}\)

2.2 Endogenous defense capability and the transition to prosperity and security

In our basic model, an incumbent with low enough defense capabilities cannot attain prosperity and security. As we show later, some societies succeeded despite an initial situation of relative insecurity. In fact, the mark of many civilizations was the erection of defensive structures to enhance the effectiveness of defense efforts. We now extend the model to allow for the endogenous expansion of defense capabilities.

2.2.1 Setup

We consider the arrival of a different challenger each period who aims to replace the incumbent and inherit its productive and defense capabilities. Like in our baseline model, in each period the incumbent selects productive investment \(i_t\) and defense \(a_t\); but the incumbent can now spend resources \(m_t\) in one period to increase its defense capability in the

---

\(^9\)Generalized versions of the ratio-based contest success function exist but are less tractable. Hirshleifer (2001) explores some of the difficulties. The typical generalization is to consider functions of the type \(a^{\alpha n + b}\). The most important feature of our model, which is the possibility of generating deterrence, obtains for any function satisfying \(\alpha \in [0, n/(n-1)]\) in a symmetric contest with \(n\) players. For \(\alpha > n/(n-1)\) pure strategy equilibria cease to exist.

\(^{10}\)The model presented here represents the limit case of a more general model where a fraction \(\sigma \in [0, 1]\) of the asset survives the war. The solution of that model is continuous in \(\sigma\), so the solution remains qualitatively similar when \(\sigma\) dips below 1 (proof available upon request).
next, and move horizontally in the \((\kappa_{t+1}, \rho)\) space. Defense capabilities evolve according to the relation \(\kappa_{t+1} = \kappa_t + m_t\), and the incumbent’s budget constraint in period \(t\) becomes \(v_t - m_t - \frac{a_t}{\kappa_t} - i_t \geq 0\).

Now we need to consider three periods, 0, 1 and 2. Since the challenger will never fight in period 2, the incumbent will never spend in expanding defense capability in period 1. Thus, the decision to augment defense capability is only relevant in period 0. All other aspects of the interaction between challenger and incumbent remain as before. After observing \((m_0, a_t, i_t)\) the challenger selects \(b_t\). If \(b_t = 0\), the incumbent retains his position in the next period. If \(b_t > 0\), then there is war at the end of period \(t\). The winner becomes the incumbent in the next period, and faces a new challenger then.

To make things as stark as possible, consider an incumbent that, barring investments in defense capability, will find itself in region \(\textbf{R3}\) in period 1, by imposing the following,

**Assumption 1**  \(\kappa_0 \rho (1 + \rho) < 4\) and \(\kappa_0 < 2\).

We ask whether a society with high enough initial productivity \(v_0\) can transition into prosperity and security despite a low initial defense capability \(\kappa_0\). More formally: can investments in defense capabilities land the incumbent in \(\textbf{R1}\) or \(\textbf{R2}\) in \(t = 1\)? The answer is not obvious because investments in defense capability become useful only in the future, and since they can be appropriated by the challenger, they incite predation now.

### 2.2.2 Solution

As before, we solve the model through backward induction. Action in periods 1 and 2 follows the logic in our baseline model. Given the initial parameters \((v_0, \kappa_0, \rho)\), the choices \((i_0, m_0)\) of the incumbent in period 0 generate a continuation value of incumbency \(V_1(i_0, m_0) = (v_0 + \rho i_0) \times S(m_0)\), where \(S(m_0)\) is a function (detailed in the Appendix A.2) that captures changes in payoffs in period 1 depending on what region \(\textbf{R1}, \textbf{R2}, \textbf{R3}\) or \(\textbf{R4}\) of the parametric space \((\kappa_1, \rho)\) the incumbent lands in. Given the continuation value \(V_1(i_0, m_0)\), we can solve for decisions in period 0. After the incumbent has selected \(m_0, a_0\) and \(i_0\), the challenger decides whether to fight. Using the same logic as in the baseline model, the challenger’s best response function is \(b_0(a_0, i_0, m_0) = \sqrt{a_0 V_1(i_0, m_0)} - a_0\) if \(a_0 < V_1(i_0, m_0)\) and zero otherwise. Since the value of incumbency \(V_1(i_0, m_0)\) in period 1 is increasing in investments.
both productive \((i_0)\) and defensive \((m_0)\), investments of both kinds incentivize challenges and represent non-trivial decisions.

Given the challenger’s best response function, the incumbent chooses \(a_0, i_0\) and \(m_0\) to maximize his expected utility,

\[
\max_{a_0, i_0, m_0 \geq 0} v_0 - m_0 - \frac{a_0}{\kappa_0} - i_0 + \frac{a_0}{a_0 + b_0(a_0, V_1(i_0, m_0))} V_1(i_0, m_0)
\]

subject to the nonnegativity constraints \(a_0 \geq 0, i_0 \geq 0, m_0 \geq 0\), the budget constraint \(v_0 - m_0 - \frac{a_0}{\kappa_0} - i_0 \geq 0\) and the deterrence constraint \((v_0 + \rho i_0)S(m_0) - a_0 \geq 0\).

We now establish,

**Proposition 2** Suppose that \(\rho > 1\) and Assumption 1 holds. Then,

1. If \(v_0\) is low enough, the incumbent is trapped in stagnation and insecurity. More formally, there exists \(\underline{v}(\kappa_0, \rho)\) such that if \(v_0 < \underline{v}(\kappa_0, \rho)\) investments in defense capability are zero \((m_0^* = 0)\) and in \(t = 1\) the incumbent remains in \(R_3\);

2. (i) If \(v_0\) is high enough, then in \(t = 1\) the incumbent will land somewhere in \(R_1 \cup R_2\) and enjoy increased security and growth. More formally, there exists \(\bar{v}(\kappa_0, \rho)\) such that if \(v_0 > \bar{v}(\kappa_0, \rho)\) the incumbent makes positive investments in defense capability \((m_0^* > 0)\) in \(t = 0\) to land somewhere in \(R_1 \cup R_2\) in \(t = 1\); (ii) If \(v_0 > \bar{v}(\kappa_0, \rho)\), higher \(v_0\) yields (weakly) higher levels of growth and security, and a strictly higher continuation value \(EV_2\).

3. Consider any point in \(R_1 \cup R_2\) that can be reached in \(t = 1\) by making an investment \(m_0^*\) in defense capability. The initial income \(v_0\) that makes such investment optimal is decreasing in \(\rho\), implying that, given a distribution of initial incomes \(v_0\), more income levels allow the incumbent a transition into \(R_1 \cup R_2\) when \(\rho\) is high than when it is low.

The intuition for the result is as follows. Investments both productive and defensive augment challenges, and become discouraged for \(v_0\) low enough. It is preferable to consume in the present rather than risk any investment. This holds even if \(v_0\) is sufficient to finance defenses that would allow the incumbent to exit \(R_3\) – the key to the result is not just financial feasibility. When \(v_0\) is high enough, a complementarity arises between defense and productive investments. A large investment in defense at \(t = 0\) increases security in \(t = 1\) so much, that the effective rate of return increases to make productive investment in \(t = 1\) incentive-compatible. In other words, a large enough investment in defense grows the intertemporal pie, and makes the risk of additional immediate predation worth taking.
2.3 Predictions

We use the theoretical propositions to derive predictions for the likelihood of observing civilizational success as a function of parameters. The first step in connecting the model to the historical record is to relate the parametric space to the event of a civilization rising. Consider first the world with exogenous defense capabilities. If civilization is the joint attainment of growth and a substantial degree of security, \( R_3 \) and \( R_4 \), which feature no growth, are incompatible with civilization. Civilization requires that defense and growth capabilities \( (\kappa_1, \rho) \) be high enough so that the polity can be somewhere in \( R_1 \) or \( R_2 \) and attain growth with some degree of security. Within these areas, as seen in panels (b) and (c) of Figure 1, there are locations with high growth and low security and vice versa. A finer proxy for the likelihood of civilization is one that combines security and growth outcomes in a measure of “expected growth,” such as the expected continuation value \( EV_2 \) in panel (d) of Figure 1. In areas where there is growth, this value increases with both defense and growth capabilities (in \( R_1 \), \( EV_2 = v_1 \frac{\kappa_1(1+\rho)}{\kappa_1+\rho} \), and in \( R_2 \), \( EV_2 = \frac{v_1(1+\rho)}{2} \sqrt{\frac{\kappa_1}{\rho}} \)). In sum, higher values of both capabilities \( (\kappa_1, \rho) \) make civilization more likely, both because it is more likely that the polity will be in regions \( R_1 \) or \( R_2 \) and because conditional on being in those regions, higher values of those parameters increase prosperity and security. Initial income \( v_1 \) also raises \( EV_2 \), but notably, this is conditional on growth and defense capabilities being high enough that the polity finds itself in areas where growth and security are possible in the first place. Summarizing,

**Remark 1** In a world with exogenous defense capabilities, higher values of both capabilities \( (\kappa_1, \rho) \) make civilization more likely. Conditional on relatively high values of \( (\kappa_1, \rho) \), civilization is more likely for higher values of initial income \( v_1 \); initial income is otherwise irrelevant.

High exogenous defense capabilities are no longer necessary for civilization when the incumbent can strengthen those capabilities. But such investments are costly, so if initial income is low, stagnation and insecurity are unavoidable (point 1 in proposition 2). In contrast, if initial income is high enough, stronger defense capabilities will help improve security and attain prosperity (point 2). This effect, however, should arise mainly in areas where growth capabilities are relatively high (point 3). These observations immediately yield the following,
Remark 2 If (exogenous) defense capabilities are low, locations associated with civilizational success should have relatively high initial income and growth capabilities, while at least one of these parameters should be low among civilizational failures.

These remarks lay out the patterns that we seek to illustrate in the next section. The predictions differ from the blunter observation that it is good if every parameter is high. They highlight that there is a positive effect of initial income $v_1$ on the likelihood of civilization that is conditional on capabilities being relatively high; in an insecure world, high growth capability $\rho$ cannot produce civilization on its own either. Growth capabilities must be paired with high defense capabilities or with high income.

3 Historical illustrations

We illustrate the model through two quantitative case studies on the earliest two civilizations, namely Sumer and Egypt. These cases embody the historical narratives that motivate the parameters of our model, namely defense capabilities $\kappa_1$, growth capabilities $\rho$, and initial income $v_1$. The historical cases are further characterized through a quantitative exercise linking the theoretical parameters to measurable environmental conditions, yielding an empirical rendering of the theoretical parametric space. We then check if civilization, represented by cities in Egypt and Sumer, tended to arise in parts of the parametric space that the model highlights as conducive to civilization. We conclude the exercise by extending attention to pristine civilizations in China and the Indus Valley, the other two Old World’s Bronze Age civilizations.

Rather than select cities on an ad hoc basis, we rely on Modelski’s (1999) data on ancient cities, which was digitized by Reba et al. (2016). This dataset contains cities that are believed to have surpassed 10,000 inhabitants. We consider Sumerian and Egyptian cities meeting that criterion at any time until 1,200BC, the estimated date for the end of the Bronze Age in the Eastern Mediterranean.¹¹ The dataset includes 5 Egyptian cities and 18 cities in

¹¹These criteria have pros and cons. The high population cutoff restricts attention to cities that were undoubtedly marks of civilization, but causes us to miss others that did not grow as much. The timeline is a compromise between restricting attention to the time of earliest urban settlements and allowing for more observations. If we consider even earlier cities, for example, until 2500BC, most of the Sumerian cities (12 out of 16) would still be part of our illustrations, but only one (out of five) Egyptian cities would do so. Modelski (2003) raises the population cutoff to 100,000 after 1000BC, and 1 million after 1000AD which precludes consistent attention to the New World.
Sumer. We eliminated 2 Sumerian cities because upon further investigation their locations turn out to be uncertain or unknown (Akkad and Akshak), leaving us with 16 Sumerian cities. Table A.1 in the Appendix lists the cities in the sample. Cities are then located in a spatial grid with cells of 1/5th of a degree side length, roughly 22km at the Equator. The construction of the empirical proxies relies on various datasets. We describe below how we compute each empirical measure. We offer further details, including the sources utilized, in Online Appendix B.

3.1 Egypt and Sumer

Southern Mesopotamia gave rise to the first major civilization, based on a cluster of city-states. The Egyptian civilization emerged slightly later, but its development after the adoption of agriculture was faster: in only about 1,000 years after the adoption of farming, a state emerged that managed a relatively wealthy economy and was also able to protect it for long stretches of time (Bard 1994, Allen 1997). It is commonly argued that geography played a role in the development of both civilizations. Sumer was located in a riverine valley, along the Tigris and Euphrates, exceptionally endowed for alluvial agriculture. And the Nile river and surrounding deserts are credited with shaping outcomes in Ancient Egypt. Such consensus highlights connections between properties of the natural environment and the conditions for prosperity and security. In what follows we detail these connections.

1) Growth capabilities - The potential of infrastructure for water management and irrigation. Egyptians could vastly increase their economic output by investing in water management, which in the Nile valley took the form of basin irrigation. Egyptians used a grid of basins to trap the floodwater and increase soil fertility before planting. Scholars agree that in Egypt irrigation agriculture “could generate crop-to-seed yields of between 12:1 and 24:1 . . . but only at the cost of high capital investments” (Morris and Manning 2005: 141). For Mann, artificial irrigation was one of the earliest forms of substantial economic investment. Both in Egypt and Mesopotamia, irrigation agriculture could “generate a surplus far greater than that known to populations on rain-watered soil” (1986: 80).

12According to a long scholarly tradition (Weber [1909] 2013, Wittfogel 1957), water management and state formation were closely linked in ancient societies. The thesis of “hydraulic empires” claims that irrigation was a public good with large fixed costs, and that pristine states formed in order to provide it. However, there is evidence that irrigation was not always preceded by the emergence of state administrations.
Like Egyptians, Sumerians made massive investments in irrigation infrastructure, securing extraordinary returns. According to Mann (1986: 78), “If [the alluvium] can be diverted onto a broad area of existing land, then much higher crop yields can be expected. This is the significance of irrigation in the ancient world: the spreading of water and silt over the land. Rain-watered soils gave lower yields.” Liverani (2008: 5) gives an idea of the increase in yields: “The agricultural production of barley underwent a notable, possibly tenfold, increase thanks to the construction of water reservoirs and irrigation canals, of long fields adjacent to the canals watered by them, and thanks to the use of the plow, of animal power, of carts, of threshing sledges, of clay sickles, and of improved storage facilities.”

In our model, a high value of the parameter $\rho$ reflects an environment in which investments yield large increases in productivity, mirroring the way in which the construction of irrigation and flood control systems resulted in major expansions of production in Egypt and Sumer. Following this narrative, when we proxy $\rho$ empirically for a given cell in the map, we use the difference in an income index (the baseline income index is explained below). That difference captures how much extra income, in terms of the caloric potential of agricultural yields, can be generated when moving from rain-fed agriculture to irrigation agriculture. We do this in a way that keeps track of the availability of riverine water, to ensure that irrigation was feasible. The availability of riverine water is proxied through a standard measure of river flow accumulation. See Online Appendix B.1 for details.\footnote{Our quantitative measure of growth capabilities is admittedly narrow. Ancient populations could make investments in things other than water management, such as storage facilities, kilns, or domesticated animals that consume fodder but raise human productivity. We have chosen a measure that seems to capture something central about the mode of production of these societies, and that can be quantified with available data.}

(2) Defense capabilities - Territorial isolation as natural protection. The Nile basin is surrounded by deserts, which made invasions much less likely than in other food-producing centers. According to Bradford (2001: 9), “The sea to the north and the deserts west and east isolated the Egyptians from the rest of mankind, except for merchants, some infiltrators, and the occasional raid.” The sea (as we discuss below) eventually became a threat rather than a protection, but the desert is considered to have provided two durable kinds of protection. It discouraged the emergence and settlement of hostile neighbors nearby, and acted as a barrier against distant rivals. In terms of our model, Egypt’s territorial isolation due to deserts maps into a naturally high $\kappa_1$. Note that the desert is not an economically beneficial feature of the
geography, but the historical narrative suggests it played a role through its security effects. Following this narrative, when we proxy $\kappa_1$ empirically for a given cell in the map, we use the percentage of the territory in surrounding cells that is covered by desert.

In contrast to Egypt, Sumer was exposed to numerous threats. As Bradford (2001: 4) puts it, “Their neighbors to the west, the Amorites, nomads of the desert, infiltrated Mesopotamia... The neighbors to the east, who dwelled in the mountains, were the Gutians and the Elamites. The Gutians and, to a lesser extent, the Elamites considered Sumer and Akkad a treasurehouse to be raided.” Finer (1997: Book I, 105) located Sumer in a plain “ringed to north and east by mountains, the millenial home of barbarous highlanders, always ready and eager to descend on the wealthy cities below.” In terms of our model, the vulnerability of Sumerian settlements to invaders suggests that defense capability was lower than in Egypt.

The narratives above suggest that Egypt had high growth capabilities $\rho$ and high exogenous defense capabilities $\kappa_1$. Therefore, as predicted by Remark 1, Egypt’s candidacy for civilizational success is already viable. Egypt would appear to be located in a favorable section of $R_2$ or in $R_1$, in the North-East of Figure 1(d), which we argued would be favorable to civilization. This pattern is confirmed quantitatively by the location of Egyptian cities in our empirically-proxied parametric space – see Figure 2 panel a). Out of five Egyptian cities, four display high values of both parameters and all of them show high natural defense capability.

In contrast, the case for civilization in Sumer is not as well established. Sumerian locations were described as having high growth capabilities like Egypt, but low exogenous defense capabilities. Given their vulnerability, the trajectories of most Sumerian cities must have begun in relatively insecure parts of $R_2$, or directly in the conflict-stagnation region, $R_3$. This is borne out in the quantitative approximation in Figure 2 panel a), where it is shown that the natural defense capabilities of Sumerian cities was lower.\footnote{One might think that having nearly 50\% of surrounding territory constituting a desert, as is often the case with Sumerian cities, may offer non-trivial defense. But a few pathways could suffice for enemies to attack, explaining why the historical narrative describes Sumer as an insecure land. Thus, it is safer to restrict attention to comparative, rather than absolute, statements between Egypt and Sumer.} But if output was insecure, how could the first human civilization emerge?

The answer is that Sumerians invested in their defense capabilities. The archaeological record offers evidence of large investments to improve defense, such as perimeter walls that
Figure 2: Egypt and Sumeria: natural defense capabilities, growth capabilities, and initial income. Darker dots represent higher density of world cells.

made Sumerian cities large-scale fortifications.\textsuperscript{15} Figure C.1 in the Online Appendix includes illustrations of four Sumerian cities. All of them had walls. In fact, virtually every city in ancient history had walls. Walls were the endogenous, artificial substitute for the missing natural protection that was present in Egypt (where cities did not typically have walls).

Remark 2 indicates that vulnerable locations could attain combined prosperity and security if they had relatively high initial income with which to finance stronger defenses. This leads to consider the third parameter of the model, initial income \(v_1\).

(3) \textit{Initial income - Rivers and ecology as directly productive resources.} Egyptian economic life has been characterized as strongly dependent on the Nile, which had at least two important properties: a yearly flood that fertilized the soil, and a two-way navigability that facilitated exchange along the entire valley. According to Bradford (2001: 9), "\(T\)he Nile was perfectly ordered—its current carried boats downstream, the wind blew them back upstream—and the Nile’s regular flooding renewed the fields and made farming so easy...” Although

\textsuperscript{15}According to van de Mieroop’s (1997) study of Mesopotamian cities, “\textit{Perhaps the presence of walls was the main characteristic of a city in the eyes of an ancient Mesopotamian.}” The archaeological record substantiates not only the generalized presence of defense investments in rising city-states, but also their costliness, which would have been prohibitive to societies with low initial productivity. Both walls and the often complementary moats have been estimated to involve large investments (e.g., the cost estimate for the moat in the Babylonian city of Dur-Jakin is ten thousand men working for three and a half months (Van de Mieroop (1997): 76)).
the Tigris and Euphrates had less attractive properties, it is also agreed that Sumer shared
strong advantages, namely the alluvium combined with an unparalleled initial endowment
of plant and animal domesticates. According to Trigger (2003: 281), domesticated animals
afforded large gains in agricultural labor productivity, and may help explain why Sumer and
Egypt were the first areas in the world to develop civilization.\footnote{Diamond (1997: Ch. 8) highlights that all eight founder crops in the Neolithic were present in the area as well as four of the five most important domesticated animals. He further states (1997: 135) \textit{any attempt to understand the origins of the modern world must come to grips with the question why the Fertile Crescent’s domesticate plants and animals gave it such a potent head start}. Olsson and Hibbs (2005) present empirical evidence that corroborates this observation.}

In addition, the rivers offered variation in terms of diet (e.g., fish). All of the aforementioned properties, fertility, easy exchange, ecological diversity, map into a relative high $v_1$ in our model. We take these factors into account in our empirical proxy for the baseline ability of humans to generate income: to capture the presence of rivers we use a measure of river flow accumulation, to proxy for agricultural suitability we use the caloric potential of pre-Columbian exchange crops, and to track variation in plant and animal resources we use a measure of ecological diversity (used previously by Fenske 2014). The index is constructed as a sum of z-scores for the three factors we consider.\footnote{The use of agricultural suitability scores raises the question of whether modern measures can reliably track ancient conditions. The quality of soil has changed over time, largely due to human use, and estimated yields are sensitive to assumptions on the sophistication of land management. One partial solution to the first problem is to use measures that do not take into account soil constraints. In the Online Appendix B.3 we show that our results are robust to using these measures. The second problem can be partially addressed by focusing on yields assuming the lowest sophistication in inputs and management allowed by the FAO-GAEZ data. Our Online Appendix B offers further elaboration on the assumptions behind the measures we use and on the robustness to using alternative measures.}

Our quantification exercise again matches the historical narrative and the predictions from the theory. Remark 1 says that conditional on both capabilities being high (Egypt’s case), a higher income makes civilization more likely. Remark 2 indicates that if defense capabilities are low but growth capabilities are high (Sumer’s case), a high income makes civilization more likely. Figure 2 panel b) shows that initial income in the locations that gave rise to Egyptian and Sumerian cities was high relative to other cells in the globe. Four out of five Egyptian city locations and fourteen out of sixteen Sumerian city locations have incomes at or above the median. Seventeen of the twenty-one cities across both civilizations fall in the upper-right ($v, \rho$) quadrant. This is not due to a strong positive correlation between initial income and growth capabilities, as seen in Figure 2 b), where we shade the parametric

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure2}
\caption{Initial incomes and growth capabilities of ancient cities.}
\end{figure}
space to indicate the density of cells. (The same conclusion obtains if we restrict attention to only those cells in Egypt and Sumer.)

One potential concern with this exercise—beyond the small number of successes—is that the locations we deem as failures may be homes to other civilizations that developed relatively soon after Egypt and Sumer. There are two other Old World pristine civilizations fitting that description, to which we now turn.

### 3.2 China and the Indus valley

The Indus valley gave rise to a multi-city civilization in locations both upstream (e.g., Harappa) and downstream (e.g., Mohenjo Daro), along the Indus and nearby rivers, reaching its peak in the 2600-1900BC period. Civilization in China is believed to have emerged in the Yellow River, giving rise to the Erlitou and Erligang cultures in the 1900-1200BC period. Roughly at the same time, the Sichuan basin gave rise to the largely independent Sangxingdui culture. Both the Chinese and Indus civilizations emerged, like Egypt and Sumer, in riverine valleys.

**Defense capabilities** - None of the Chinese or Indus locations were, like Egypt, completely surrounded by deserts that could isolate them, suggesting their natural defense capabilities were comparatively low. Chinese cities were often walled, and there is plenty of evidence of fighting in Neolithic times (Liu and Chen 2012: 230). The Sangxingdui culture was less connected to areas outside of the Sichuan basin due to mountains, but the role of mountains as defense isolators remains unclear.\(^{18}\) The Sanxingdui site was also walled, suggesting defense was a concern (Watabe 2002). A similar concern is reflected in the walls surrounding (most often the citadels) of Indus valley centers. However, evidence of violence has been elusive, prompting debates on whether Indus sites were truly peaceful and if so how they might have achieved peace (Parpola 2015, Ch. 4). In sum, both the Chinese and Indus civilizations resemble Sumer more than Egypt, in that they were seemingly located in less impregnable territories and built defenses, which in the Indus case may have attained deterrence.

**Growth capabilities** - Abundant scholarship has established that the Chinese and Indus

\(^{18}\)See Finer's statement quoted earlier about the mountains surrounding Sumer as the source of, rather than a barrier against, threats. There are other examples of abrupt landscapes (e.g., the Andes) not preventing groups from attacking each other or a group projecting its power.
civilizations made impressive advances in crafts production, construction, metallurgy, and the development of water management systems. All of these required investments. Both civilizations used animals to do work – for example driving the plough, or providing force to lift water. This made humans more productive, and entailed a form of investment (e.g., deferring consumption to produce fodder for animals). The investment in water management in urban spaces is well established for both Yellow river (see Storozum et al 2017 and references therein) and Indus civilizations (Kenoyer 1998). The cities in the Indus Valley had impressive systems to circulate water and dispose of sewage; the sanitary benefits would raise the returns to agglomerated living and production. Their use of irrigation infrastructure in rural settings is less clear. The exact investments that were made to capitalize on floods is the subject of discussion. In the case of the Chinese, farming occurred near rivers and large scale irrigation infrastructure is only attested through excavations in sites pertaining to the late Bronze Age – water management was otherwise more basic, including the digging of moats and small-scale canals (Storozum et al 2017). In the case of the Indus Valley, scholars continue to discuss the specific ways in which irrigation was utilized. Some investment and maintenance-demanding forms of irrigation were used, such as small-scale canals, oxbow irrigation, and wells (Miller 2006). While some of the growth capabilities of these regions may involve aspects we do not measure (e.g., minerals), the presence of flooding rivers makes clear that there were gains to water management, flood control, and to forms of artificial irrigation (albeit likely less developed or large-scale than those in Sumer and Egypt).

Initial income - the land near the Yellow river, rich in loess, and that in the Sichuan basin were highly productive (Trigger 2003, Watabe 2002). The land in the Indus valley would be fertilized by flooding (although flooding patterns would be irregular; see Miller 2006), and all of these areas would have (progressive) access to valuable grain and animal domesticates.

In sum, the description of Chinese and Indus locations is compatible with cases of relatively low exogenous defense capabilities, but relatively high growth capabilities and initial income. In Figure 3 we add cities from the Chinese and Indus valley civilizations and place them in the (empirically proxied) three-dimensional space $\langle \kappa_0, \rho, v_0 \rangle$. The additional Old World pristine civilizations seem to extend the Sumerian pattern of low defense capabilities with high growth potential and high initial income (the only exception is Dholavira in the Indus Valley, located in an island inside a lake - it is associated with high initial income, but low natural defense and growth capabilities - it is thought to have been a commerce town).
Taking all four civilizations together, it emerges that civilization needs growth potential, but as the model indicates more is needed. Cities are associated with either high initial defense capabilities (as in Egypt), or if that fails, a high initial income that can help create those capabilities (as in Sumer, China, and the Indus Valley). The world has a majority of its locations in areas where all capabilities and income are low, or where either income, growth or defense capabilities are high on their own (74% of world cells; 68% if we restrict attention to cells in the Old World). Yet those are almost exclusively failures in civilizational terms, as the theory would predict.

We analyze the robustness of our exercise in the Online Appendix B.3. While the exact location of civilization centers in the empirically proxied parametric space can vary, the broad patterns remain when we alter various features, such as the information used to construct our empirical measures, and the size of cells into which we divide the map.
4 The end of the Bronze Age

After identifying conditions for the rise of civilizations, we now address their fall. For a period of almost 400 years, multiple states emerged in the Eastern Mediterranean that improved their productive capacity and were able—mainly through fortified walls and chariots—to defend their wealth against “barbarian” populations. This set of thriving states included the city-ports of the Levant, the kingdoms of Anatolia, the Egyptian empire, and the city-states of Mesopotamia and Cyprus. But a collapse epidemic swept across the entire region around 1200BC. As Cline puts it (2014: 241), “…the world as they had known it for more than three centuries collapsed and essentially vanished”. According to Drews (1993: 3), “Altogether the end of the Bronze Age was arguably the worst disaster in ancient history, even more calamitous than the collapse of the western Roman Empire.”

A long debate on the causes of the collapse at end of the Bronze Age has considered earthquakes (Schaeffer 1948), droughts and famines (Carpenter 1968), internal rebellions (Zuckerman 2007 and Carpenter 1968), and innovations in military technology (Drews 1993). The hypothesis of earthquakes has been discredited in the face of new archaeological evidence showing that most urban destruction was caused not by natural forces but by human attack. Hittites and Egyptians left unequivocal testimonies of attacks by the “Sea Peoples,” as the Egyptians called them, a diverse array of intruders with different origins (Sandars 1987). The same evidence challenges a pure internal rebellion story. The possibility of invasions remains, but begs the question of what caused them in the first place. Two hypotheses consistent with available evidence are:

1. A severe climate shock (draught), which caused famines, and compelled populations in the periphery to invade in search for food. Cities that were storehouses of grain fell victim to “a final resort to violence by a drought sicken people” (Carpenter 1968: 69).\(^{19}\)

2. A revolution in the means of war, which tipped the military balance in favor of nomadic intruders. According to Drews (1993: 33), “the Catastrophe was the result of a new style of warfare that appeared toward the end of the thirteen century BC, [which] opened up new and frightening possibilities for various uncivilized populations that until that time had been no cause of concern to the cities and kingdoms of the eastern Mediterranean”. What

---

\(^{19}\)A recent paleobotany study confirms a substantial climate shock around the time of the collapse that could have caused a famine (Langgut, Finkelstein, and Litt 2013). In the interpretation of these authors the shock may have caused internal rebellions rather than foreign invasions.
were the changes introduced by the “uncivilized populations”? Chrissanthos (2008: 11) summarizes them: “these tribes developed better and lighter body armor, [...] lighter and smaller round shields, [...] revolutionary longer, stronger swords [...] They also invented a new weapon, the javelin, which could be used as a missile to hurl at an enemy. They [managed to] overcome the civilizations’ chariot advantage [...] Once these tribes mastered sea travel, no shore was too far for an attack. The failure of the chariot in the face of this new warfare marks the beginning of the Bronze Age world’s collapse”.20

We now use our incumbent-challenger model with exogenous defense capabilities as a tool for short-run analysis given the (relative) sudden nature of the shocks at the end of the Bronze Age. Consider the challenger’s valuation to be parameterized as $h \times v$ ($h > 0$ for hunger), and the challenger’s military capability $\kappa_c \geq 1$, so the challenger’s expected benefit reads $\frac{h_t}{a_t + b_t} h V_{t+1} - \frac{h_t}{\kappa_c}$. The historical debate has sometimes considered changes in the challenger’s motivation to attack ($h$) and its military effectiveness ($\kappa_c$) as the rival explanations (1) and (2) above. But the “hunger” and the “barbarian military innovation” hypotheses, while historically distinct, are formally identical in our model. At the margin, $h$ and $\kappa_c$ affect the aggressiveness of the challenger in the same way. Therefore, fixing $h = 1$ and studying the comparative statics of $\kappa_c$ can illuminate the role of changes in both the motivation and aggressiveness of challengers.

The parameter $\kappa_c$ was assumed equal to 1 in the baseline model. We now consider a move to $\kappa_c > 1$. How will the incumbent fare when facing a tougher challenger? In other words, how does a higher $\kappa_c$ affect the partition of the parameter space derived in Proposition 1? The following proposition yields the answer.

**Proposition 3** For any point in the $(\kappa_1, \rho)$ space where either security or prosperity (or both) are attained, a higher $\kappa_c$ implies that security, prosperity or both may be lost. A higher

---

20 A climate shock and military innovation are not mutually incompatible causes and can be combined under the form of a “perfect storm” (Cline 2014: Chapter 5). Another recent explanation builds on the idea of “system collapse” since late Bronze Age societies were tightly connected through commerce, the fall of a few of them could set off domino effects. This allows for the theoretical possibility that climate- and technology-induced invasions had devastated only a critical number of nodes in the interconnected Eastern Mediterranean, but eventually provoked a general collapse.

21 The parameter $h$ could also track the differential ability of the challenger at “operating” the asset under the interpretation that a successful challenge leads to replacement. One issue we do not take up here is the case where a challenger has a high valuation for the stream of production (as when looting) but a low valuation for the asset due to an inability to operate it. These are interesting variations left for future research.
\( \kappa_c \) reduces the area of \( R1 \) (where both security and prosperity obtain). In addition, \( R3 \), which yields insecurity and stagnation, grows at the expense of all others, making it less likely that the incumbent capabilities \((\kappa, \rho)\) are compatible with civilization.

The comparative statics of \( \kappa_c \) are intuitive: a more aggressive challenger reduces the set of parameter values for which the incumbent attains both growth and security, and enlarges the set in which he attains neither. Figure E.1 in the Online Appendix includes a graphical representation of the comparative static effects of \( \kappa_c \).

The historical victory of the “Sea Peoples” over the kingdoms of the Ancient Near East can be seen as a shift from the security-prosperity region (\( R1 \), or good parts of \( R2 \)) to the conflict-stagnation region (\( R3 \)), as a result of changes in challenger motivation (an increase in \( h \), in turn an effect of a climate shock) or military technology (an increase in \( \kappa_c \)). More broadly, this exercise illustrates arguments made by social theorists that the evolution of political complexity is not unilinear, but plagued by reversals that could stem from environmental or military shocks.

5 Related Literature

5.1 Conflict and state capacity

Our work belongs in the literature on conflict, which is too extensive to review here, in particular the papers investigating the tradeoff between “guns” and “butter” (see Garfinkel and Skaperdas 2007 for a review), some of which have become canonical in the area of conflict. The closest example is perhaps Grossman and Kim (1995). In their model, agents choose between contestation and producing to consume; their agents, like our incumbent, are concerned with deterrence. Like them, we consider a tradeoff between consumption and security. However, their model does not include a separate choice to invest, and therefore it cannot speak to the other tradeoffs we analyze, pitting investment against both consumption and security. In addition, they assume interior solutions for some of their variables, while we investigate all interior and corner solutions.

A recent literature studies investments in state capacity by a ruler who may lose control of the polity to a competing faction (Besley and Persson 2011), or to a foreign power (Gennaioli and Voth 2015). One difference with Besley and Persson’s model is that our incumbent
controls defense and the economy directly, while their incumbent faction controls different forms of state capacity investment. As importantly, investments in our model augment the virulence of challenges. Gennaioli and Voth (2015) formalize and investigate empirically Tilly’s (1992) argument that modern European states formed as a result of the competitive pressures of military conflict, which created a need to centralize fiscal control. Gennaioli and Voth (2015) model the problem of a ruler choosing the level of fiscal centralization. In their setup, centralization contains elements of both defense and investment: higher centralization today generates resources that, if there is conflict tomorrow, become the funding for the war effort tomorrow, and the disputable loot the day after tomorrow. Crucially, however, their ruler makes a single choice, while our incumbent selects investment and defense separately. The lack of a separate decision to invest is not sensitive for the purposes of Gennaioli and Voth, nor Grossman and Kim. But it is important for us to separate investment from defense in order to characterize the tension between growth and security. If we tied together the defense and investment decisions, a lot would ride on how exactly we do so; we might observe investment due to an implicit defense component to the investment choice that would be absent in a world where the incumbent can make separate decisions.

Our paper is also related to models of state consolidation (Powell 2012, 2013); the key difference is that in our model consolidation is studied in relation to investment and growth.

5.2 Early states

Archaeologists like V. Gordon Childe (1936), who first conceptualized the advent of the Neolithic era as an “agricultural revolution,” focused on the innovations in the means and relations of production while abstracting from the necessary accompanying innovations in military protection. On the other hand, several archaeologists have noted the paramount role of investments in protection, such as fortifications, walls, and moats, in the erection of the first cities (Service 1975, 299). According to Near Eastern archaeologist Volkmar Fritz, “in the Jordan Valley, settlements were surrounded by a wall even before it is possible to speak of the city proper” (1997 II: 19). Other authors, like Mann (1986), explicitly connected food production with protection needs, as mentioned earlier. However, we are not aware of any account that has explicitly focused on the interplay of surplus production and surplus protection to point out a solution to the civilizational paradox.
Our approach builds on, but departs from, historical accounts that emphasize the geographic sources of economic prosperity. The approaches emphasizing the availability of domesticable plants and animals to explain why some regions generated surpluses while others did not (e.g., Diamond 1997) contribute a necessary building block for understanding the prosperity of the first settled societies. However, a purely geographic approach is incomplete, for it misses the role of incentives and strategic action. Our approach incorporates both strategic actors and geographic factors such as food production potential or protective terrain.

Our investigation comes at the cost of abstracting from some aspects that have been considered in anthropological theories of the state, such as social stratification and domination (Fried 1960, 728; Carneiro 1970). We abstract from social hierarchy not because we think stratification is unimportant, but because it helps to focus attention on the incumbent-challenger interaction. Some similarities arise, however. For Carneiro (1970), states originated as growing populations contested fertile areas surrounded by less productive land. The fertile Nile valley, surrounded by deserts, is a good candidate location. Our model generates a similar empirical implication; however, it is not driven by exploitation but by the fact that low quality surrounding land can protect against challengers. Unlike Carneiro’s theory, our model does not appeal to population pressure, an assumption that has been challenged by some writers (Allen 1997), and it highlights the role of investment returns.

It is customary in the social sciences to view the state as the monopoly on violence. Adapting from Weber, we define the state not in binary terms but as a matter of degrees (Weber 1978: ch. I, s. 16), so that state formation involves higher degrees of protection from attacks. We focus on the state as “sovereignty,” defense from threats, and abstract from “rulership,” the creation of a political hierarchy and institutions within a society. The exclusion of rulership from our model helps identify a minimalist view of early civilizations as the intersection of surplus production and statehood seen as sufficient surplus protection.

Our work is related to both theories of state formation (Tilly 1975, 1992, Spruyt 1996) and theories of the political sources of prosperity (North and Weingast 1989; Olson 1993, 2000, Bates 2001; Acemoglu et al. 2005, Boix 2015). In contrast to our model, theories of state formation do not place the state in the context of the “security-prosperity” tradeoff,

---

22For a review of anthropological theories of early states see for example Claessen and Skalnik (1978).
23See Schönholzer (2017) for a recent investigation of Carneiro’s hypothesis.
and theories of the political sources of prosperity focus on rules of the political game once
the state is already in place rather than on pre-state forces.\textsuperscript{24}

Our work has important complementarities with that by Mayshar, Moav and Neeman (2017), and Mayshar, Moav, Neeman and Pascali (2015). They also combine a focus on early states, an emphasis on geographic drivers, and the use of formal theory. For us, geography matters because it defines both productive and defense capabilities, while for them it determines the observability of production (the former paper) or its appropriability (the latter). Mayshar, Moav and Neeman (2017) use a principal-agent model to show how monitoring capabilities shape the extent of political centralization, and account for contrasting trajectories in Sumer and Egypt, where observability of the Nile allowed for a more unified and lasting state. Our focus is not on the form of states, but on the conditions for their emergence. This is also the focus of Mayshar, Moav, Neeman and Pascali (2015), who focus on crop appropriability. They equate the state with the political hierarchy that results from appropriability and assume it results in the full prevention of conflict. We abstract from appropriability and internal hierarchy, and investigate whether it is true that conflict can be eliminated.

6 Conclusion

We build a model to investigate the combined operation of growth and defense capabilities in a society that seeks to consolidate security and grow its economy. The components of the model are chosen by reference to the anthropological and historical literature, in order to capture relevant environmental parameters and the minimalistic strategic dilemma facing proto-civilizations.

The theory makes contributions on two levels. One is conceptual, by formalizing the emerging tradeoffs when a polity must choose between consumption, defense, and growth. This helps evaluate claims about the relative role of security and prosperity that are central to classic theories of state formation. We show that if defense capabilities are fixed (or there

\textsuperscript{24}We share with Boix (2015) an interest in mechanisms of state formation extending back into prehistoric times, as well as in “hard” causes related to the physical environment. Although Boix finds sources of pre-institutional cooperation under conditions of anarchy (absence of state), he conceives of state formation as the selection of either republican or monarchic institutional settings. By contrast, we focus on the conditions for state formation that allow for investment and security before political institutions become central.
is not enough income to change them substantially), all four combinations involving the presence or absence of security and prosperity are possible. This precludes statements of security and prosperity being necessary or sufficient for one another.

A tension arises between prosperity and security, because investments attract predation, and a combination of relatively high growth and defense capabilities can relax the tension. If defense capabilities are low but can be augmented, the security-prosperity tension can be relaxed as long as initial income is high enough. A large enough investment in defense creates conditions for productive investments to eventually be made safely. Isolating formally the role of defense capability in the civilizational process contributes to understanding how economic shocks can affect state formation and political stability more generally.

The second contribution of the theory is to offer predictions that can be contrasted with historical narratives and data. The predictor of civilizational success is a combination of high defense and growth capabilities, and if defense capabilities start low, a simultaneous presence of high growth capabilities and initial income. Thus, high initial income or growth capabilities alone are not sufficient for civilization.

We use these predictions to rationalize the emergence of civilization, and the first historical episode of civilization collapse. We construct a quantitative representation of the theoretical parametric space by reference to environmental characteristics. In a context of sparse data, quantifying the model’s parametric space offers a way to enrich the historical analysis beyond a purely narrative approach. Our exercise does not amount to a formal statistical test, and we do not use it to rule out alternative theories. But we show that the quantitative patterns match the historical narrative and that our model has some empirical traction. The pristine Bronze Age civilizations in the Old World tend to appear in locations of the parametric space predicted by the model.
A Appendix

A.1 Proof of Proposition 1:

This is a particular case of the model with general values of $\kappa_c$, studied in Proposition 3.

A.2 Proof of Proposition 2:

We use the convention that in period 0 the incumbent starts by selecting $m_0$, and then picks $a_0$ and $i_0$.\textsuperscript{25} Backward inducting within period 0, we first characterize solutions to $(a_0, i_0)$ given $m_0$ and then prove the statements in the proposition concerning $m_0^*$.

From the expressions for $V_1$ in Proposition 1, as the incumbent lands in each of the regions $R1 - R4$, the value $V_1$ is,

$$V_1(i_0, m_0) = (v_0 + \rho i_0) \times \begin{cases} \frac{(\kappa_0 + m_0)(1+\rho)}{\kappa_0 + m_0 + \rho} & (\kappa_0 + m_0, \rho) \in R1 \\ \sqrt{\frac{(\kappa_0 + m_0)(1+\rho)}{\rho}} & (\kappa_0 + m_0, \rho) \in R2 \\ \frac{1}{\kappa_0 + m_0} & (\kappa_0 + m_0, \rho) \in R3 \\ 2 - \frac{1}{\kappa_0 + m_0} & (\kappa_0 + m_0, \rho) \in R4 \end{cases}$$

Once $m_0$ is set, the problem of choosing $(a_0, i_0)$ so solve the program in (10) is similar to the choice of $(a_1, i_1)$ in the baseline model, except now the continuation value depends explicitly on $m_0$ through $S(m_0)$. The objective function is differentiable and concave in $a_0$ and $i_0$, and the constraints are linear, so the first order and complementary slackness conditions below are necessary and sufficient for a unique maximum:

$$a_0 : \frac{1}{2} \sqrt{(v_0 + \rho i_0) S(m_0)} - \frac{1}{\kappa_0} - \frac{\lambda_{BC}}{\kappa_0} - \frac{\lambda_{DC}}{\kappa_0} + \lambda_a = 0 \quad (11)$$

$$i_0 : \frac{\rho S(m_0)}{2} \sqrt{\frac{a_0}{(v_0 + \rho i_0) S(m_0)}} - 1 - \frac{\lambda_{BC}}{\kappa_0} + \frac{\lambda_{DC}}{\kappa_0} + \lambda_i = 0 \quad (12)$$

\textsuperscript{25}The assumption that $m_0$ is decided before $a_0$ and $i_0$ is just to simplify exposition. The results do not vary if $(m_0, a_0, i_0)$ are selected simultaneously since the challenger moves after the incumbent.
\[ \lambda_{BC}(v_0 - m_0 - \frac{a_0}{\kappa_0} - i_0) = 0, \quad \lambda_{DC}\left( (v_0 + \rho i_0)S(m_0) - a_0 \right) = 0, \quad \lambda_a a_0 = 0, \quad \lambda_i i_0 = 0 \quad (13) \]

As before, \( \lambda_{BC}, \lambda_{DC} \) are the Lagrange multipliers for the budget constraint and deterrence constraints, and \( \lambda_a, \lambda_i \) are the multipliers for the non-negativity constraints for the control variables. The infinite marginal utility of \( a_0 \) at zero implies \( a_0 > 0 \) and \( \lambda_a = 0 \), so there are eight possible cases depending on whether the remaining three Lagrange multipliers are positive or zero. The following Lemma, the proof of which is in the Online Appendix D.1, shows that given our Assumption 1 there are only two feasible cases in period 0.

**Lemma 1** If Assumption 1 holds, then in period 0 the incumbent chooses:

i) \( i_0 = 0 \) and \( a_0 = \frac{\kappa_0^2}{4} v_0 S(m_0) \) when \( \frac{v_0 S(m_0)}{v_0 - m_0} < \frac{4}{\kappa_0} \); or

ii) \( i_0 = 0 \) and \( a_0 = \kappa_0 (v_0 - m_0) \) when \( \frac{v_0 S(m_0)}{v_0 - m_0} > \frac{4}{\kappa_0} \).

We now proceed with the statements in the proposition. We analyze separately the cases where \( \rho \leq 2 \) and \( \rho > 2 \). We focus on the first case and relegate to the Online Appendix D the second, analogous, case. There are three steps. Step 1 is to compute the expected utility in period \( t = 0 \) as a function of \( m_0 \) fixing all the other parameters for the two cases highlighted in Lemma 1. Step 2 is to note that there are levels of \( m_0 \) that shift the regime the polity is in, both in period 0 (depending on which case of Lemma 1 obtains) and also in period 1 (depending on which region \( R_1, R_2, R_3, R_4 \), the polity lands in period 1). Each of these regime changes is reflected in the expected utility function. Step 3, having correctly characterized expected utility, is to show that a low enough \( v_0 \) yields \( m_0^* = 0 \) for part 1, and for part 2(i) that high enough \( v_0 \) yields a positive \( m_0^* \) that is as high as desired.

**Proposition 2, part 1.** We prove this by showing that for \( v_0 \) small enough, \( m_0^* = 0 \).

Step 1 - We write the incumbent’s expected utility in each of the two cases in Lemma 1:

- **Case** \( \lambda_{BC} = 0 \) (BC not binding), \( \lambda_{DC} = 0 \) (DC not binding, conflict), and \( \lambda_i > 0 \) (\( i_0 = 0 \))

  The proof to Lemma 1 shows that the Lagrange multiplier conditions defining the case imply \( \frac{v_0 S(m_0)}{v_0 - m_0} < \frac{4}{\kappa_0}, \kappa_0 < 2, \rho S(m_0) < \frac{4}{\kappa_0}, \) and expected utility is, \( V_0 = v_0 - m_0 + \frac{\kappa_0}{4} v_0 S(m_0) \).

- **Case** \( \lambda_{BC} > 0 \) (BC binds), \( \lambda_{DC} = 0 \) (DC not binding, conflict) and \( \lambda_i > 0 \) (\( i_0 = 0 \))
The Lagrange multiplier conditions imply, \( \frac{v_0 S(m_0)}{v_0 - m_0} > \frac{4}{\kappa_0} \), \( \frac{v_0 S(m_0)}{v_0 - m_0} \geq \kappa_0 \), \( \frac{v_0 S(m_0)}{v_0 - m_0} > \rho S(m_0) \), and expected utility is,

\[
V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 S(m_0).
\]

Step 2 - We denote with \( \bar{m} \) the value of \( m_0 \) that satisfies \( \frac{v_0 S(\bar{m})}{v_0 - \bar{m}} = \frac{4}{\kappa_0} \) and which makes the polity switch from case 1 to case 2 in Lemma 1 in period 0. We denote with \( m_{RX,RY} \) the value of \( m_0 \) such that regimes change in period 1 from RX to RY. In the case of R3 and R4 that value is \( m_{R3,R4} = 2 - \kappa_0 \). Because \( \bar{m} \) is an implicit function of \( S(\cdot) \), we compute conditions on the parameters when \( \bar{m} \) lies respectively below and above \( m_{R3,R4} \). The reason that it is important to know where \( \bar{m} \) lies relative to \( m_{R3,R4} \) is that it will indicate which expected utility expression to use to evaluate choices of \( m_0 \). It turns out that if \( v_0 \) is low enough, more precisely, if \( v_0 < \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \), then \( \bar{m} < m_{R3,R4} \). To see this, notice that \( \frac{v_0 S(m_0)}{v_0 - m_0} \) is increasing in \( m_0 \). Using the definition for \( \bar{m} \), note \( \bar{m} < m_{R3,R4} \iff \frac{v_0 S(m_{R3,R4})}{v_0 - m_{R3,R4}} > \frac{4}{\kappa_0} \iff v_0 < \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \).

Step 3 - Now we know how to write expected utility depending on the value of \( v_0 \), given all other parameters. We only need to find a cutoff in the space of possible values for \( v_0 \) such that when \( v_0 \) is under the cutoff the incumbent prefers to stay in R3. We propose \( \nu = \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \). In this case, \( v_0 < \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \) is equivalent to a regime described by \( \bar{m} < m_{R3,R4} \), which means we have to use two different expected utility expressions in the interval \([0, m_{R3,R4}]\) depending on whether \( m_0 < \bar{m} \), or \( m_0 > \bar{m} \). A useful fact is that \( \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \) is strictly decreasing in \( \kappa_0 \) so its maximum value is at \( \kappa_0 = 0 \) (since \( \kappa_0 \geq 0 \)). In this case \( \frac{8(2 - 1)}{8 - 3 \times 1} = 2 \).

**Segment [0, \bar{m}]** Expected utility in period \( t = 0 \) is given by case (i) in Lemma 1 and by considering \( S(m_0) \) to be given by the expectation of remaining in R3:

\[
V_0 = v_0 - m_0 + \frac{\kappa_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{\kappa_0}{4} v_0 (1 + \frac{\kappa_0 + m_0}{4}) = v_0 (1 + \frac{\kappa_0}{4} + \left( \frac{\kappa_0}{4} \right)^2) + m_0 \left( \frac{\kappa_0 + m_0}{16} - 1 \right)
\]

Since \( \bar{m} < m_{R3,R4} \iff v_0 < \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \) then \( \frac{\kappa_0 + m_0}{16} - 1 < 0 \). To see why, replace \( v_0 = \frac{8(2 - \kappa_0)}{8 - 3 \kappa_0} \) in \( \frac{\kappa_0 + m_0}{16} \) so \( \frac{\kappa_0 + m_0}{16} = \frac{\kappa_0}{16} + \frac{2(2 - \kappa_0)}{16} \leq \frac{2 \times 2}{16} < 1 \). This implies \( V_0 \) is decreasing in \( m_0 \) and the optimal choice is \( m_0 = 0 \).

**Segment [\bar{m}, m_{R3,R4}]** Expected utility in period \( t = 0 \) is given by case (ii) in Lemma 1 and by considering \( S(m_0) \) to be given by the expectation of remaining in R3:

\[
V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 S(m_0) = \sqrt{\kappa_0 (v_0 - m_0)} v_0 (1 + \frac{\kappa_0 + m_0}{4}), \text{ and we now show this to}
\]
decrease in $m_0$. Note,

$$\frac{dV_0}{dm} = \frac{1}{2} \left( \kappa_0 (v_0 - m_0) v_0 \left( 1 + \frac{\kappa_0 + m_0}{4} \right) \right)^{-\frac{1}{2}} \left( -\kappa_0 v_0 \left( 1 + \frac{\kappa_0 + m_0}{4} \right) + \frac{1}{4} \kappa_0 (v_0 - m_0) v_0 \right)$$

and,

$$\frac{dV_0}{dm} < 0 \iff \frac{1}{2} \kappa_0 (v_0 - m_0) v_0 < \kappa_0 v_0 \left( 1 + \frac{\kappa_0 + m_0}{4} \right)$$

or, iff $v_0 < 4 + \kappa_0 + 2m_0$.

If $4 + \kappa_0 + 2m_0$ is higher than $\frac{8(2-\kappa_0)}{8-3\kappa_0}$, then the condition $\bar{m} < m_{R3|R4}$ is also a sufficient condition for $V_0$ in the segment $[\bar{m}, m_{R3|R4}]$ to be decreasing. So, it is sufficient to show that $4 + \kappa_0 + 2m_0 > \frac{8(2-\kappa_0)}{8-3\kappa_0}$. Because the right hand side is decreasing in $\kappa_0$, it attains a maximum at $\kappa_0 = 0$ and it is equal to 2 which is smaller than any feasible value of the expression in the left hand side, which is at least 4. Therefore, in this segment utility is maximized at $m_0 = \bar{m}$, and equals $V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 S(m_0) = \sqrt{4(v_0 - \bar{m})^2} = 2(v_0 - \bar{m})$.

**Segment** $[m_{R3|R4}, m_{R4|R1}]$ Since now $m_0$ can only be larger than $\bar{m}$, we know expected utility will be given by case (ii) in Lemma 1 and by the expectation of landing in $R4$, implying $V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 S(m_0) = \sqrt{\kappa_0 (v_0 - m_0)} v_0 \left( 2 - \frac{1}{\kappa_0 + m_0} \right)$. Computing the first derivative with respect to $m_0$, we get,

$$\frac{dV_0}{dm_0} = \left( \kappa_0 (v_0 - m_0) v_0 \left( 2 - \frac{1}{\kappa_0 + m_0} \right) \right)^{-\frac{1}{2}} \left[ -\kappa_0 v_0 \left( 2 - \frac{1}{\kappa_0 + m_0} \right) + \frac{\kappa_0 (v_0 - m_0) v_0}{(\kappa_0 + m_0)^2} \right]$$

which is negative whenever $2(\kappa_0 + m_0) - 1 > \frac{(v_0 - m_0)}{\kappa_0 + m_0}$, or, $2(\kappa_0 + m_0)^2 - \kappa_0 > v_0$. Note $2(\kappa_0 + m_{R3|R4})^2 - \kappa_0 = 8 - \kappa_0$. Now note $8 - \kappa_0 > \frac{8(2-\kappa_0)}{8-3\kappa_0} \equiv 2$, since the LHS is at least 6 and the RHS is at most 2. Thus, $\frac{dV_0}{dm_0} < 0$ and utility would be maximized at $m_{R3|R4}$ in this segment.

**Segment** $[m_{R4|R1}, \infty]$ Expected utility is given by case (ii) in Lemma 1 and by the expectation of landing in $R1$, implying $V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 \left( \frac{\kappa_0 + m_0(1+\rho)}{\kappa_0 + m_0 + p} \right)$ and

$$\frac{dV_0}{dm_0} = \left( \kappa_0 (v_0 - m_0) v_0 \left( \frac{\kappa_0 + m_0(1+\rho)}{\kappa_0 + m_0 + p} \right) \right)^{-\frac{1}{2}} \left[ -\kappa_0 v_0 \left( \frac{\kappa_0 + m_0(1+\rho)}{\kappa_0 + m_0 + p} \right) + \kappa_0 (v_0 - m_0) v_0 \frac{(1+\rho)(\kappa_0 + m_0 + p) - (\kappa_0 + m_0)(1+\rho)}{(\kappa_0 + m_0 + p)^2} \right].$$

Note $\frac{dV_0}{dm_0} < 0$ whenever $v_0 < \frac{(\kappa_0 + m_0 + p)(\kappa_0 + m_0)}{p} + m_0$. The right hand side of this expression is increasing in $m_0$, so the minimum is attained at $m_0 = m_{R4|R1}$ and it equals $\frac{\rho}{(p-1)^2} + \frac{2\rho}{p-1} - \kappa_0$. The highest possible value of $v_0$, $v = \frac{8(2-\kappa_0)}{8-3\kappa_0}$ is smaller than 2 which, in turn, is always smaller than $\frac{\rho}{(p-1)^2} + \frac{2\rho}{p-1} - \kappa_0$ given that $\rho < 2$ (its minimum value is 4, at $(\kappa_0, \rho) = (2, 2)$). Therefore the maximum of $V_0$ in this segment is attained at $m_0 = m_{R4|R1}$. 

37
Considering all the segments together, we now show that the global maximum in this case is \( m^*_0 = 0 \). This follows from the just demonstrated fact that the maximum within each segment of the support is at the minimum value, and from the fact that \( V_0 \) is continuous. This in turn follows from the fact that \( S(.) \) is continuous for all \( m_0 \) and \( V_0 \) in period \( t = 0 \) is also continuous at \( \bar{m} \): it is easy to show that in \( t = 0 \), \( V_0 \) in segment \([0, \bar{m}]\) evaluated at \( \bar{m} \) is \( 2(v_0 - \bar{m}) \), which is equal to \( V_0 \) in segment \([\bar{m}, m^*_{R3,R4}]\) evaluated at \( \bar{m} \). Thus, if \( v_0 < v^* \) the incumbent does not invest in defense capabilities, and the polity will stay at \( R3 \) in period 1.

**Proposition 2, part 2 (i).** We show that if \( v_0 \) is high enough, then the optimal choice \( m^*_0 \) is a (unique) strictly positive, strictly increasing, unbounded function of \( v_0 \).

Define \( \bar{v} = \left( \frac{\rho}{\rho - 1} \right)^2 \frac{4}{\kappa_0} \).

Steps 1&2 - In the frontier \( R4 | R1 \), \( \rho = \kappa_1 / (\kappa_1 - 1) = (\kappa_0 + m_0) / (\kappa_0 + m_0 - 1) \), so \( m^*_{R4,R1} \) satisfies \( \rho = (\kappa_0 + m^*_{R4,R1}) / (\kappa_0 + m^*_{R4,R1} - 1) \)

\[
\rho \kappa_0 + \rho m^*_{R4,R1} - \rho = \kappa_0 + m^*_{R4,R1} \\
\rho (\kappa_0 - 1) = \kappa_0 (1 - \rho) m^*_{R4,R1} \\
\frac{\rho (\kappa_0 - 1) - \kappa_0}{1 - \rho} = m^*_{R4,R1}
\]

It turns out that if \( v_0 > \frac{4}{\kappa_0} \left( \frac{\rho}{\rho - 1} - \kappa_0 \right) \), then \( \bar{m} > m^*_{R4,R1} \). To see this, note that from Lemma 1 and the fact that \( \frac{v_0 S(m_0)}{v_0 - m_0} \) is increasing in \( m_0 \), \( \bar{m} > m^*_{R4,R1} \) requires \( \frac{v_0 S(m^*_{R4,R1})}{v_0 - m^*_{R4,R1}} < \frac{4}{\kappa_0} \), and this follows iff \( v_0 > \frac{4}{\kappa_0} \left( \frac{\rho}{\rho - 1} - \kappa_0 \right) \).

Therefore everywhere in \( R3 \) and \( R4 \) the expected utility of the incumbent corresponds to case (i) in Lemma 1, but in \( R1 \) the expected utility switches to that corresponding to case (ii) in Lemma 1 for \( m_0 \geq \bar{m} \).

Step 3 - We now show that if \( v_0 > \bar{v} = \left( \frac{\rho}{\rho - 1} \right)^2 \frac{4}{\kappa_0} \), expected utility is increasing in \( m_0 \) for as long as \( m_0 \) is in \( R3 \) and \( R4 \), and that there must be an interior solution \( m^*_0 \) in \( R1 \). This solution is generically unique because the maximum stems from a concave expected utility, regardless of whether it lies in the segment \([m^*_{R4,R1}, \bar{m}] \) or \([\bar{m}, \infty) \).

**Segment \([0, m^*_{R3,R4}] \)** Expected utility is given by case (i) in Lemma 1 and by the expectation of landing in \( R3 \), implying,

\[
V_0 = v_0 - m_0 + \frac{\kappa_0}{4} v_0 S(m_0) = v_0 - m_0 + \frac{\kappa_0}{4} v_0 \left( 1 + \frac{\kappa_0 + m_0}{4} \right) = v_0 (1 + \frac{\kappa_0}{4} + \left( \frac{\kappa_0}{4} \right)^2) + m_0 \left( \frac{\kappa_0^2}{16} - 1 \right).
\]
For \( V_0 \) to be increasing over \([0, m_{R3|R4}]\), we need \( v_0 > \frac{16}{\kappa_0} \). This is guaranteed if \( \bar{v} = \left( \frac{\rho}{\rho-1} \right)^2 \frac{4}{\kappa_0} > \frac{16}{\kappa_0} \Leftrightarrow \left( \frac{\rho}{\rho-1} \right)^2 > 4 \Leftrightarrow \frac{\rho}{\rho-1} > 2 \) which holds strictly for \( \rho < 2 \) and with equality at \( \rho = 2 \).

**Segment \([m_{R3|R4}, m_{R4|R1}]\)** Expected utility is given by case (i) in Lemma 1 and by the expectation of landing in \( R4 \), implying, \( V_0 = v_0 - m_0 + \frac{\kappa_0}{2} v_0 S(m_0) = v_0 - m_0 + \frac{\kappa_0}{2} v_0 \left( 2 - \frac{1}{\kappa_0 + m_0} \right) \).

Note this expected utility is maximized at \( m_0^* = \sqrt{\frac{2v_0}{\kappa_0} - \kappa_0} \). But this maximum lies above \( m_{R4|R1} \) whenever \( \sqrt{\frac{2v_0}{\kappa_0}} - \kappa_0 > \frac{\rho(\kappa_0 - 1) - \kappa_0}{\kappa_0} \). Or, equivalently, whenever \( v_0 > \frac{4}{\kappa_0} \left( \frac{\rho}{\rho - 1} \right)^2 \equiv \bar{v} \).

Thus, the polity must reach \( R1 \).

**Segment \([m_{R4|R1}, \bar{m}]\)** Expected utility is given by case (i) in Lemma 1 and by the expectation of landing in \( R1 \), implying, \( V_0 = v_0 - m_0 + \frac{\kappa_0}{2} v_0 S(m_0) = v_0 - m_0 + \frac{\kappa_0}{2} v_0 \left( \frac{\kappa_0 + m_0}{\kappa_0 + m_0 + \rho} \right) \). This is a concave function with a maximum at \( m_0^* = \left( \frac{\sqrt{\frac{2v_0}{\kappa_0}}}{\kappa_0} \right) \rho(1 + \rho), -\kappa_0 - \rho \), which shows that if the maximum lies in \([m_{R4|R1}, \bar{m}]\), it is unique. Note if \( v_0 > \left( \frac{\rho}{\rho - 1} \right)^2 \frac{4}{\kappa_0} \equiv \bar{v} \) then \( m_0^* > m_{R4|R1} \).

To see this, rewrite the inequality \( m_0^* > m_{R4|R1} \) as \( \sqrt{\frac{2v_0}{\kappa_0}} - \kappa_0 - \rho > \frac{\rho}{\rho - 1} - \kappa_0 \) then plug \( \bar{v} \) into the LHS to obtain \( 1 + \rho > \rho \). Therefore for \( v_0 > \bar{v} \) the polity will be in the interior of \( R1 \) in period 1. Moreover, for \( v_0 > \bar{v} \) we have \( m_0^* \) is a strictly increasing, unbounded, function of \( v_0 \).

**Segment \([\bar{m}, \infty)\)** Expected utility is given by case (ii) in Lemma 1 and by the expectation of landing in \( R1 \), implying,

\[
V_0 = \sqrt{\kappa_0 (v_0 - m_0)} v_0 S(m_0) = \sqrt{\kappa_0 (v_0 - m_0)} v_0 \frac{\kappa_0 + m_0}{\kappa_0 + m_0 + \rho}.
\]

The first order condition for a maximum is:

\[
\frac{1}{2} \frac{\kappa_0 v_0 (\kappa_0 + m_0)(1 + \rho)}{\kappa_0 + m_0 + \rho} - \frac{\kappa_0 (v_0 - m_0)}{2} v_0 (1 + \rho)(\kappa_0 + m_0 + \rho)\frac{(\kappa_0 + m_0)(1 + \rho)}{\kappa_0 + m_0 + \rho} = 0,
\]

which, after some algebra boils down the quadratic,

\[
m_0^2 + 2 (\kappa_0 + \rho) m_0 + \kappa_0^2 + \rho (\kappa_0 - v_0) = 0,
\]

which has only one (strictly) positive root \( m_0^* = - (\kappa_0 + \rho) + \sqrt{\rho (\kappa_0 + \rho + v_0)} \), an increasing,
increasing in $\kappa$, function of $v$ and unbounded function of $v$ segments $[m, \infty)$.Lemma 2, point 2. in the Online Appendix D.3) and there is only one solution to the first order condition.

To show $V_0$ is concave, we need to show $\frac{d^2V_0}{dm_0^2} < 0$. Note that,

$$\frac{d^2V_0}{dm_0^2} = \frac{1}{2V_0^2} \left\{ \left[ -\kappa_0 v_0 \frac{dS}{dm_0} + \kappa_0 v_0 \frac{d}{dm_0} \left( (v_0 - m_0) \frac{dS}{dm_0} \right) \right] V_0 - \left( -\kappa_0 v_0 S(m_0) + \kappa_0 (v_0 - m_0) \frac{dS}{dm_0} \right) \frac{dV_0}{dm_0} \right\}.$$

The second term within curly brackets in $\frac{d^2V_0}{dm_0^2}$ is zero when evaluated at $m_0^*$, yielding,

$$\frac{d^2V_0}{dm_0^2} = \frac{1}{2V_0} \left[ (\kappa_0 v_0) \frac{dS}{dm_0} \right] < 0,$$

where the sign follows from the fact that $\frac{dS}{dm_0} = (1 + \rho) \rho > 0$ and that

$$\frac{d}{dm_0} \left( (v_0 - m_0) \frac{dS}{dm_0} \right) = -\frac{dS}{dm_0} + (v_0 - m_0) \frac{d^2S}{dm_0^2} < 0$$

because $v_0 - m_0 \geq 0$ and $\frac{d^2S}{dm_0^2} = -\frac{2\rho(1+\rho)}{(\kappa_0+m_0+\rho)^2} < 0$.

We have now proved statement 2(i) in the proposition, namely that $m_0^*$ must lie in segments $[m_{R4}, m]$ or $[m, \infty)$ and in each case it is a unique solution and a positive, increasing, unbounded function of $v_0$.

**Proposition 2, part 2 (ii).** We show that growth, security, and $EV_2$ are weakly increasing in $\kappa_1$, which is an increasing function of $m_0$, and by virtue of 2(i) an increasing function of $v_0$. For $v_0 > \bar{v}$, we have $EV_2 = \frac{a_1^*}{a_1 + b_1} V_2 = \frac{a_1^*}{\sqrt{a_1^* V_2}} V_2 = \sqrt{\frac{a_1^*}{V_2} V_2}$. Since landing in $R1$ implies $a_1^* = V_2 = v_1 \frac{\kappa_1(1+\rho)}{\kappa_1 + \rho}$ from Proposition 1, and $i_0 = 0$ from Lemma 1, we get $EV_2 = v_0 \frac{\kappa_1(1+\rho)}{\kappa_1 + \rho}$, which is increasing in $\kappa_1$ and hence in $v_0$. Note once in $R1$ security is complete and invariant in $v_0$ but prosperity increases. The same is true when landing in $R1$ in the case with $\rho > 2$ (available in the Online Appendix D.3), while growth is constant in $v_0$ and security increasing when landing in $R2$.

**Proposition 2, part 3.** In segment $[m_{R4}, m]$ the solution is $m_0^* = \sqrt{\frac{\kappa_0 + \rho}{4}} (1 + \rho) - \kappa_0 - \rho$, and in segment $[m, \infty)$ we have the solution $m_0^* = -\kappa_0 - \rho + \sqrt{\rho (\kappa_0 + \rho + v_0)}$. In
each respective case, the \( v_0 \) that is required to produce such solution is,

\[
v_0 = \frac{4}{\kappa_0 \rho (1 + \rho)} (m_0^* + \kappa_0 + \rho)^2
\]
\[
v_0 = \frac{1}{\rho} (m_0^* + \kappa_0 + \rho)^2 - (\kappa_0 + \rho).
\]

Differentiating \( v_0 \) with respect to \( \rho \) shows that such needed income is decreasing in \( \rho \) in each case.

\[\blacksquare\]

A.3 Proof of Proposition 3:

The problem is to maximize,

\[
\mathcal{L} = v_1 - \frac{a_1}{\kappa_1} - i_1 + \frac{a_1}{a_1 + b_1} (v_1 + \rho i_1)
\]
\[
+ \lambda_{BC} (v_1 - \frac{a_1}{\kappa_1} - i_1) + \lambda_{DC} (\kappa_c v_1 - a_1 + \kappa_c \rho i_1) + \lambda_a a_1 + \lambda_i i_1.
\]

(14)

We characterize the solution \((a_1, i_1, \lambda_{BC}, \lambda_{DC}, \lambda_a, \lambda_i)\) to this problem for each parameter combination \((\rho, \kappa_1, \kappa_c, v_1)\). To save on notation, let us define \(PS = \{(\kappa_1, \kappa_c, \rho, v_1) | \kappa_1 \geq \kappa_c > 0, \rho > 1, v_1 > 0\}\) the parameter space we consider throughout the proof. The first order and complementary slackness conditions that characterize the optimum are given by,

\[
\frac{\partial \mathcal{L}}{\partial a_1} = \frac{1}{2 \sqrt{\kappa_c}} \sqrt{\frac{v_1 + \rho i_1}{a_1}} - \frac{1}{\kappa_1} - \frac{\lambda_{BC}}{\kappa_1} - \lambda_{DC} + \lambda_a = 0; a_1 \geq 0, \lambda_a \geq 0, \lambda_a a_1 = 0 \text{ c.s.} \quad (15)
\]

\[
\frac{\partial \mathcal{L}}{\partial i_1} = \frac{\rho}{2 \sqrt{\kappa_c}} \sqrt{\frac{a_1}{v_1 + \rho i_1}} - 1 - \lambda_{BC} + \lambda_{DC} \kappa_c \rho + \lambda_i = 0; i_1 \geq 0, \lambda_i \geq 0, \lambda_i i_1 = 0 \text{ c.s.} \quad (16)
\]

\[
\lambda_{BC} (v_1 - \frac{a_1}{\kappa_1} - i_1) = 0 \text{ c.s., } \lambda_{DC} (\kappa_c v_1 - a_1 + \kappa_c \rho i_1) = 0 \text{ c.s.} \quad (17)
\]

Since \(\lambda_a = 0\), there are eight possible cases to be analyzed, given by whether the remaining Lagrange multipliers \(\lambda_{BC}, \lambda_{DC}, \text{ and } \lambda_i\) are zero or positive. We assume in each case that the conditions defining it hold, and then determine which part if any of the parameter space
This solution is consistent with \( \lambda^*_BC > 0 \) (BC binds), \( \lambda^*_DC > 0 \) (DC binds, consolidation), and \( \lambda^*_i = 0 \) \( (i^*_i > 0) \) Since \( a_1 \) is always positive, and in this case \( i_1 \) is also positive, the FOCs in (15), (16) must hold with equality. Because this case involves binding BC and DC, constraints also hold with equality. This means,

\[
a_1 : \frac{1}{2\sqrt{\kappa_c}} \sqrt{\frac{v_1 + \rho i_1}{a_1}} - \frac{1}{\kappa_1} - \frac{\lambda_{BC}}{\kappa_1} - \lambda_{DC} = 0;
\]

\[
i_1 : \frac{\rho}{2\sqrt{\kappa_c}} \sqrt{\frac{a_1}{v_1 + \rho i_1}} - 1 - \lambda_{BC} + \lambda_{DC} \rho \kappa_c = 0
\]

\[
DC : \kappa_c v_1 - a_1 + \kappa_c \rho i_1 = 0; \quad BC : v_1 - \frac{a_1}{\kappa_1} - i_1 = 0.
\]

BC and DC yield \( i^*_i = v_1 \frac{(\kappa_1 - \kappa_c)}{(\kappa_1 + \kappa_c \rho)} \) and \( a^*_i = v_1 \frac{\kappa_1 \kappa_c (1 + \rho)}{\kappa_1 + \kappa_c \rho} \). Then \( \lambda^*_i = 0 \) (or \( i^*_i > 0 \)) is supported by \( \kappa_1 > \kappa_c \). Using \( a^*_i \) and \( i^*_i \), the FOCs are seen to involve only \( \lambda_{DC}, \lambda_{BC} \), and parameters \( \kappa_1, \kappa_c \), showing \( \lambda_{DC} > 0 \Leftrightarrow \kappa_1 > \rho \kappa_c \) and \( \lambda_{BC} > 0 \Leftrightarrow \rho > \kappa_1/(\kappa_1 - \kappa_c) \). The parameter set supporting this solution is therefore \( R1 = \{(\kappa_1, \rho, v_1) \in PS | \rho < \kappa_1/\kappa_c, \rho > \kappa_1/(\kappa_1 - \kappa_c), \kappa_1 > \kappa_c \} \) and in this region there is investment and deterrence. The expected utility is \( V_1 = v_1 \frac{\kappa_1 (1 + \rho)}{(\kappa_1 + \kappa_c \rho)} \).

When \( \kappa_c > \kappa_1 \) this case would be infeasible, as it contradicts \( i_1 > 0 \).

2. Case \( \lambda^*_BC > 0 \) (BC binds), \( \lambda^*_DC = 0 \) (DC does not bind, conflict), and \( \lambda^*_i = 0 \) \( (i^*_i > 0) \) The first order and complementary slackness conditions yield

\[
a_1 : \frac{1}{2\sqrt{\kappa_c}} \sqrt{\frac{v_1 + \rho i_1}{a_1}} - \frac{1}{\kappa_1} - \frac{\lambda_{BC}}{\kappa_1} = 0; \quad i_1 : \frac{\rho}{2\sqrt{\kappa_c}} \sqrt{\frac{a_1}{v_1 + \rho i_1}} - 1 - \lambda_{BC} = 0 \]

\[
BC : v_1 - \frac{a_1}{\kappa_1} - i_1 = 0
\]

Investment and army solutions are respectively \( i^*_1 = \frac{v_1}{2} \left(1 - \frac{1}{\rho} \right) \) and \( a^*_1 = \frac{\kappa_1 \rho v_1}{2} \left(1 + \frac{1}{\rho} \right) \). This solution is consistent with \( \lambda^*_DC = 0 \) (DC holds with strict inequality) and \( \lambda^*_i = 0 \Leftrightarrow \rho > \frac{\kappa_1}{\kappa_c} \) and \( \rho > 1 \) respectively. The solution is consistent with \( \lambda^*_BC > 0 \Leftrightarrow \rho > \frac{4\kappa_c}{\kappa_1} \).
(from checking $\lambda_{BC} > 0$ in the FOCs). As a result, the parameter set supporting this case is $R2 = \{(\kappa_1, \rho, v_1) \in PS|\rho > \kappa_1/\kappa_c, \rho > 4\kappa_c/\kappa_1\}$. Expected utility for the incumbent is $V_1^* = \frac{v_1}{2} \left(1 + \frac{1}{\rho}\right) \sqrt{\rho \kappa_1}$.

3. Case $\lambda_{BC}^* = 0$ (BC does not bind), $\lambda_{DC}^* = 0$ (DC does not bind, conflict), and $\lambda_i^* > 0$ ($i_1^* = 0$) In this case $a_i^* = v_1 \kappa_1^2/(4\kappa_c)$ and $i_1^* = 0$. This solution is consistent with $\lambda_{BC}^* = 0$ and $\lambda_{DC}^* = 0 \iff \kappa_1 < 2\kappa_c$. Also for $\lambda_i^* > 0$ we need $1 - \rho \kappa_1/(4\kappa_c) > 0$ (from the FOC of $i_1$). Thus, this holds for any triple $(\rho, \kappa_1, v_1) \in PS$ such that $\kappa_1 < 2\kappa_c$ and $\rho < 4\kappa_c/\kappa_1$. In other words, the parameter set for which this region contains the solution to the incumbent’s problem is $R3 = \{(\kappa_1, \rho, v_1) \in PS|2\kappa_c > \kappa_1, \rho < 4\kappa_c/\kappa_1\}$. Expected utility in this case is given by $V_1^* = v_1 \left(1 + \frac{\kappa_1}{4\kappa_c}\right)$.

4. Case $\lambda_{BC}^* = 0$ (BC does not bind), $\lambda_{DC}^* > 0$ (DC binds, consolidation), and $\lambda_i^* > 0$ ($i_1^* = 0$) In this case the system of conditions is given by $a_1^* = \frac{1}{2\kappa_c} - \frac{1}{\kappa_1} - \lambda_{DC} = 0; i_1 : \frac{\kappa}{2} - 1 + \lambda_{DC} \rho \kappa_c + \lambda_i = 0$; and $BC : \kappa_c v_1 - a_1 = 0$. Since $i_1^* = 0$, the DC yields $a_1^* = \kappa_c v_1$. For this to be consistent with $\lambda_{DC}^* > 0$, we must have from the first equation that $\kappa_1 > 2\kappa_c$, and to be consistent with $\lambda_i^* > 0$ we need $\rho < \kappa_1/(\kappa_1 - \kappa_c)$, yielding $R4 = \{(\kappa_1, \rho, v_1) \in PS|\kappa_1 > 2\kappa_c, \rho < \kappa_1/(\kappa_1 - \kappa_c)\}$. The expected utility is $V_1^* = v_1 \left(2 - \frac{\kappa_c}{\kappa_1}\right)$.

5. Case $\lambda_{BC}^* > 0$ (BC binds), $\lambda_{DC}^* > 0$ (DC binds, deterrence), and $\lambda_i^* > 0$ ($i_1^* = 0$) Because $i_1^* = 0$, BC binding implies that $a_i^* = \kappa_1 v_1$, but DC binding implies that $a_i^* = \kappa_c v_1$, so $\kappa_1 = \kappa_c$ which is non-generic.

6. Case $\lambda_{BC}^* > 0$ (BC binds), $\lambda_{DC}^* = 0$ (DC does not bind, conflict), and $\lambda_i^* > 0$ ($i_1^* = 0$) The BC binding and $i_1^* = 0$ yield $a_i^* = v_1 \kappa_1$. The BC binds iff $\kappa_1 > 4\kappa_c$. The DC not binding, however, implies $v_1 \kappa_c - v_1 \kappa_1 > 0 \iff \kappa_c > \kappa_1$ which violates the condition for BC to bind $\kappa_1 > 4\kappa_c$, making this case infeasible.

7. Case $\lambda_{BC}^* = 0$ (BC does not bind), $\lambda_{DC}^* = 0$ (DC does not bind, conflict), and $\lambda_i^* > 0$ ($i_1^* > 0$) Non-generic, since it is consistent for subset of the space $(\rho, \kappa_1, v_1)$ that has (under a suitable measure) measure zero. This follows from (15) and (16), so when $\lambda_{BC}$, $\lambda_{DC}$, $\lambda_i = 0$, FOCs read $a_1 : \frac{1}{2\sqrt{\kappa_c}} \sqrt{\frac{v_1 + \rho \kappa_1}{\kappa_1}} - \frac{1}{\kappa_1} = 0$ and $i_1 : \frac{\rho}{2\sqrt{\kappa_c}} \sqrt{\frac{v_1 + \rho \kappa_1}{\kappa_1}} - 1 = 0$. The first FOC implies $\sqrt{\frac{v_1 + \rho \kappa_1}{\kappa_1}} = 2\sqrt{\kappa_c}/\kappa_1$ and substituting into the second FOC, we get $\frac{\rho}{2\sqrt{\kappa_c}} \frac{\kappa_1}{2\sqrt{\kappa_c}} = 1$ or $\frac{\rho \kappa_1}{4\kappa_c} = 1$, which implies this holds for a non-generic parameter set.

8. Case $\lambda_{BC}^* = 0$ (BC does not bind), $\lambda_{DC}^* > 0$ (DC binds, consolidation), and
\( \lambda_1^* = 0 \) \((i_1^* > 0)\) The FOCs are,

\[
a_1 : \frac{1}{2\sqrt{\kappa_1}} \sqrt{\frac{v_1 + \rho i_1}{a_1}} - \frac{1}{\kappa_1} - \lambda_{DC} = 0; \quad i_1 : \frac{\rho}{2\sqrt{\kappa_1}} \sqrt{\frac{a_1}{v_1 + \rho i_1}} - 1 + \kappa_1 \rho \lambda_{DC} = 0,
\]

where \( a_1 = \kappa_c(v_1 + \rho i_1) \) indicating that \( \lambda_{DC} \) must simultaneously equal \( \frac{1}{2\kappa_c} - \frac{1}{\kappa_1} \) and \( \left( \frac{\kappa_1}{\kappa_c\rho} \right) \), which forces the equality \( \rho = \frac{\kappa_1}{\kappa_c} \), which is non-generic.

Combining these cases yields that optimal behavior by the incumbent yields a partition of the parameter space \((\rho, \kappa_1, v_1) \in (1, \infty) \times (0, \infty) \times \mathbb{R}_+\) into four regions:

**Region 1 (R1):** \(\{(\kappa_1, \rho, v_1) | \rho < \kappa_1/\kappa_c, \rho > \kappa_1/(\kappa_1 - \kappa_c), \kappa_1 > \kappa_c\}\) Security and prosperity

In **R1** the solution is: \(\left\{ a_1^* = v_1 \frac{\kappa_1 \rho_1 (1 + \rho)}{(\kappa_1 + \kappa_c \rho)}, i_1^* = v_1 \frac{(\kappa_1 - \kappa_c)}{(\kappa_1 + \kappa_c \rho)}, V_1^* = v_1 \frac{\kappa_1 (1 + \rho)}{(\kappa_1 + \kappa_c \rho)} \right\}\)

**Region 2 (R2):** \(\{(\kappa_1, \rho, v_1) | \rho > \kappa_1/\kappa_c, \rho > 4\kappa_c/\kappa_1 \text{ and } \rho > 1\}\) Prosperity without security

In **R2** the solution is: \(\left\{ a_1^* = \frac{\kappa_1 \rho_1}{2} \left( 1 + \frac{1}{\rho} \right), i_1^* = \frac{\rho_1}{2} \left( 1 - \frac{1}{\rho} \right), V_1^* = \frac{\rho_1}{2} \left( 1 + \frac{1}{\rho} \right) \sqrt{\rho_1} \right\}\)

**Region 3 (R3):** \(\{(\kappa_1, \rho, v_1) | 2\kappa_c > \kappa_1 \text{ and } \rho < 4\kappa_c/\kappa_1\}\) Neither prosperity nor security

In **R3** the solution is: \(\left\{ a_1^* = v_1 \left( \frac{\rho_1}{2} \right)^2, i_1^* = 0, V_1^* = v_1 \left( 1 + \frac{\rho_1}{4\kappa_c} \right) \right\}\)

**Region 4 (R4):** \(\{(\kappa_1, \rho, v_1) | \kappa_1 > 2\kappa_c, \rho < \kappa_1/(\kappa_1 - \kappa_c)\}\) Security without prosperity

In **R4** the solution is: \(\left\{ a_1^* = \kappa_c v_1, i_1^* = 0, V_1^* = v_1 \left( 2 - \frac{\kappa_1}{\kappa_1} \right) \right\}\)

A comparison of region by region with the case where \(\kappa_c = 1\) yields the proposition. ■

### A.4 Ancient Cities

We use the data on ancient cities digitized and geocoded by Reba et al. (2016). These authors render spatially the population estimates by Chandler (1987) and Modelski (1999, 2003). Reba et al. spatialized both datasets by using a single, central latitude and longitudinal point with 2 to 8 significant figures for each settlement location. Chandler’s data focus on the largest cities with population over 20,000 inhabitants from AD 800 to AD 1850. This population threshold is very high and yields an under-representation of cities for the period up to and including the Bronze Age, which is crucial to our interest and better covered by Modelski. For this reason, we rely on Modelski’s data. He uses a lower threshold of 10,000 inhabitants, and covers urban settlements from 3700 BC to 1000 BC. The majority of Modelski’s population values come from detailed archeological site reports (Modelski 1999: 384). In a few cases, Modelski uses additional historical evidence and census data, such as
those in Adams (1981), to compute population estimates.

For each century-city, Modelski’s dataset either features a missing value or an estimate of the population (in multiples of 1,000). Our focus is on Sumerian and Egyptian cities, but we also consider cities in China and the Indus Valley. Our proxy for civilization are cities that exhibit at least one non-missing observation between 3700 BC and 1200 BC, the end of the Bronze Age. Table A.1 shows the 30 cities used in our illustrations: 16 correspond to Sumeria, 5 to Egypt and 9 to Indus Valley and China. Reba et al. provide longitude and latitude for two additional Sumerian cities, Akkad and Akshak, but we exclude them from the analysis because no consensus exists about their exact locations (see e.g., Foster 2013).

<table>
<thead>
<tr>
<th>City</th>
<th>1st Observation</th>
<th>City</th>
<th>1st Observation</th>
<th>City</th>
<th>1st Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eridu</td>
<td>3700 BC</td>
<td>Memphis</td>
<td>2500 BC</td>
<td>Harappa</td>
<td>2500 BC</td>
</tr>
<tr>
<td>Larak</td>
<td>3500 BC</td>
<td>Heliopolis</td>
<td>2400 BC</td>
<td>Mohenjodaro</td>
<td>2500 BC</td>
</tr>
<tr>
<td>Uruk</td>
<td>3500 BC</td>
<td>Elephantine</td>
<td>1900 BC</td>
<td>Rakshigarhi</td>
<td>2300 BC</td>
</tr>
<tr>
<td>Nippur</td>
<td>3000 BC</td>
<td>Thebes</td>
<td>1800 BC</td>
<td>Dholavira</td>
<td>2300 BC</td>
</tr>
<tr>
<td>Zabalam</td>
<td>2800 BC</td>
<td>Hermopolis</td>
<td>1300 BC</td>
<td>Erliotu</td>
<td>1700 BC</td>
</tr>
<tr>
<td>Adab</td>
<td>2800 BC</td>
<td>Bo (Yanshi)</td>
<td>1600 BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ur</td>
<td>2800 BC</td>
<td>Ao (Zhengzho)</td>
<td>1400 BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larsa</td>
<td>2800 BC</td>
<td>Yin</td>
<td>1300 BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Umna</td>
<td>2800 BC</td>
<td>Sanxingdui</td>
<td>1200 BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kish</td>
<td>2800 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shuruppak</td>
<td>2800 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lagash</td>
<td>2500 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girsu</td>
<td>2400 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isin</td>
<td>2000 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eshnunna</td>
<td>1800 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Badtibira</td>
<td>1800 BC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Excluded

| Akshak | 2500BC |
| Akkad  | 2200BC |


References


