European Economic Review 161 (2024) 104620

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

On the coevolution of cooperation and social institutions

Verónica Salazar [a], Balázs Szentes [b,a,*]

[a] Department of Economics, London School of Economics, United Kingdom
[b] HKU Business School, The University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Cooperation
Prisoner's Dilemma
Evolutionary dynamics

ABSTRACT

This paper examines an environment inhabited by self-interested individuals and unconditional cooperators. The individuals are randomly paired and engage in the Prisoner's Dilemma Game. Cooperation among players is incentivized by institutional capital, and selfish individuals incur a cost to identify situations where defection goes unpunished. In this environment, we explore the coevolution of types and institutional capital, with both the distribution of types and capital evolving through myopic best-response dynamics. The equilibria are shown to be Pareto-ranked. The main finding is that any equilibrium level of institutional capital exceeds the optimal amount in the long run. Thus, forward-looking optimal institutions not only foster a more cooperative culture but are also more cost-effective compared to the myopically optimal ones.

1. Introduction

In human societies, cooperation is often facilitated through institutions that monitor and punish anti-social behavior. For example, police forces prevent criminal acts, while judicial systems resolve disputes and enforce contracts. However, developing and maintaining these institutions is costly, leading to variations in their quality across societies. Individuals respond to these institutions by either cooperating or exerting an effort to avoid punishment while engaging in non-cooperative behavior. The relative appeal of these choices depends on the strength and quality of the social institutions, which, in turn, are influenced by the behavioral culture in the society. The goal of this paper is to examine the coevolution of cooperative behavior and social institutions and understand the welfare implications of their mutual dependency.

We propose a stylized model where individuals play the Prisoner's Dilemma repeatedly against random opponents. The game's payoffs are designed so that cooperation becomes more beneficial as the number of cooperators increases. A fraction of these interactions is monitored and enforced, and this fraction is determined by the accumulated amount of *institutional capital* in the society. Individuals fall into one of two types: *cooperators* always cooperate, while *strategic* individuals cooperate if and only if the interaction is monitored. Strategic individuals incur a fitness cost to distinguish between monitored and unmonitored interactions. The distribution of types evolves based on a best-response dynamic, where a type's frequency increases if its payoff exceeds that of the other type. The level of institutional capital adjusts myopically and sluggishly towards the level that would be socially optimal given the current distribution of types. This adjustment is myopic, because it does not take into account how a change in institutions will further modify behavior through affecting the type-composition, and sluggish, because it only allows for gradual changes to the current level of institutions. The state of the society is determined by two variables: the fraction of strategic types and the amount of institutional capital. A state is an *equilibrium* when neither the type distribution nor the institutional capital changes. Coexistence of the two types in equilibrium is only possible if their payoffs are equal. Furthermore, the equilibrium amount of institutional capital is myopically optimal given the composition of types.

* Corresponding author at: Department of Economics, London School of Economics, United Kingdom.
E-mail addresses: v.salazar-restrepo@lse.ac.uk (V. Salazar), szentes@hku.hk (B. Szentes).

https://doi.org/10.1016/j.euroecorev.2023.104620
Received 4 October 2022; Received in revised form 18 October 2023; Accepted 21 October 2023
Available online 28 October 2023

Our first finding is that equilibria always exist and, if multiple, can be ranked in terms of Pareto efficiency. Equilibria with higher levels of institutional capital have more strategic individuals and lower social welfare. To elaborate, consider a society in equilibrium where a surge of strategic individuals occurs, shifting the type distribution to another equilibrium. Due to the complementarity of cooperation, the payoff of a strategic individual exceeds that of a cooperator after the surge. To balance this payoff gain, social institutions must make it harder for strategic individuals to exploit cooperators. Consequently, the other equilibrium requires a higher amount of institutional capital. However, both the surge of strategic individuals and the increase in capital reduce the payoff of both types, thereby lowering welfare. This occurs because cooperative behavior becomes less frequent and strengthening institutions incurs costs.

We then focus on determining the long-run optimal amount of institutional capital. This quantity aims to maximize social welfare while considering that the type distribution evolves in response to the strength of social institutions. The main result of the paper is that the long-run optimal level of institutional capital is smaller than any equilibrium level. In fact, the optimal amount is the minimum required to prevent strategic types from entering the society. Interestingly, setting the institutional capital to the long-run optimum not only leads to a steady state populated entirely by cooperators but also reduces spending on institutional capital.

Optimal social institutions are designed to uphold a cooperative culture and prevent its decline at minimal costs. However, these institutions do not align with the preferences of the current population. In the long-run optimal steady state where only cooperators exist, there is no need to monitor their interactions, resulting in a myopically optimal capital level of zero. When investment in social institutions is driven by short-term goals, these institutions weaken, leading to an influx of strategic individuals and reduced cooperation. Institutional capital continues to deplete until the presence of numerous strategic individuals makes further capital reduction no longer myopically optimal. At this point, the payoff of a strategic individual still surpasses that of a cooperator, attracting more strategic individuals while increasing the level of institutional capital. Ultimately, the fraction of strategic types becomes significant, and the amount of institutional capital exceeds the long-run optimum. However, this argument does not imply that disinvesting institutional capital improves welfare once society reaches an equilibrium. Strengthening institutions to eliminate anti-social behavior is necessary to transition to the long-run optimum. Only after establishing a cooperative culture can capital be reduced.

It is well-documented that social institutions foster cooperative societies, for reviews in the contexts of crime deterrence by police and tax enforcement, see Klick and Tabarrok (2010) and Slemrod (2007), respectively. Moreover, it is evident that short-sighted decision-making, driven by factors like political concerns and myopic voter behavior, can lead to inefficient investment in these institutions. The main takeaway from our paper is that these inefficiencies not only deteriorate a cooperative culture but also result in excessively strong and costly social institutions.

Literature review. The ubiquity of cooperation in strategic interactions between unrelated individuals is the focus of a large game-theoretic literature, which seeks to explain behavior through the lens of methodological individualism (Dawes and Thaler, 1988; Boyd and Richerson, 2009).[1] A prominent approach to explain cooperation in the Prisoner's Dilemma played by self-interested and genetically unrelated individuals is to consider a repeated version of the game, which opens the door to reciprocity and reputation-building.[2] In efficient equilibria of such games, players punish defectors and opponents who failed to execute punishment strategies in the past. In our model, such strategies are not available because players do not observe the history of their opponents' play. Instead of relying on equilibrium strategies, we introduce social institutions which are built to enforce cooperation. We then examine the evolutionary stability of equilibrium strategies in the spirit of Maynard Smith and Price (1973). The seminal work of Axelrod (1981) also suggested to consider evolutionary stable strategies to study cooperation.

Our model deals specifically with instances in which cooperation occurs in the absence of kinship, repeated encounters, reputation formation, and assortative matching. Such cooperation appears to be unique to human species (Fehr and Fischbacher, 2003). Research in evolutionary biology suggests that gene-based evolutionary theories are not enough to explain many patterns of human altruism, thus gene-culture coevolution must be considered (Gintis, 2003) . Along these lines, we postulate that individual strategies and social institutions are evolving jointly. In line with North's definition of institutions as "the rules of the game", institutions are a parameter that changes the payoff structure of the game (Davis and North, 1970). They can be interpreted literally to be the strength of centralized enforcement of behaviors that align with the common good but not with private interest, such as legal and judiciary systems. Less literally, institutions can be seen as a reduced-form representation of all social forces that encourage cooperation beyond individuals' propensities to cooperate.

Despite the large literature looking at cooperation through an evolutionary lens, there are few theoretical models that explore the simultaneous evolution of cooperative behavior and institutions. Some exceptions include Tabellini (2008), Bisin and Verdier (2017), and Migliaccio and Verdier (2018). Tabellini (2008) focuses on a model where there is a notion of distance between individuals, and individuals value cooperating with others who are closely related. Thus, the incentives to cooperate or defect are driven by how close the players are, the extent of external enforcement, and an individual's intrinsic type. Its key finding is that strong legal enforcement between unrelated individuals breeds more "good values", whereas more localized external enforcement is likely to undermine the transmission of cooperative values. The former acts as a complement to value-based cooperation, while the latter substitutes, or crowds-out, "good values"in favor of a closeness-based cooperation. Bisin and Verdier (2017) model institutions as Pareto weights in the objective function of the policy-designer, as opposed to the extent to which cooperative behavior is enforced.

[1] For reviews of the literature in the field of theoretical biology, see Nowak (2012), Perc and Szolnoki (2010).

[2] In other models, players can choose to punish defectors at a private cost, see, for example, Sethi and Somanathan (1996), Fehr and Gächter (2002), Fehr and Fischbacher (2003), Sasaki et al. (2017).

Monitoring and enforcement in their framework would be the outcome of a policy that is determined by a political equilibrium in each period. In our model, in contrast, Pareto weights perfectly coincide with the distribution of types and it is the "policy" itself that evolves. Migliaccio and Verdier's (2018) model considers a different kind of evolving institutions, where the degree to which matches are assortative is the evolving parameter. That is, what coevolves with cooperation is how much cooperators are able to interact with other cooperators. This model is closer to the literature in theoretical biology on endogenous network structures (see Perc and Szolnoki, 2010 for an overview).

Perhaps the most related work to ours is Lee et al. (2019) who also examine how evolving institutions can facilitate cooperation in the Prisoner's Dilemma. The authors model social institution by a set of designated agents, called *umpires*, who are specialized in monitoring and enforcing cooperation. These umpires may be honest or corrupt and the frequency of their types changes over time endogenously. We view our work as complementary to that of Lee et al. (2019). Lee et al. (2019) focus on the incentives of employees of social institutions but take the structure of the social institutions as given exogenously. Indeed, the number of umpires, the size of their fee and bribe are all fixed.[3] While we recognize that a well-functioning social institutions are founded on the basis of reliable employees, our model abstracts from their incentives. Instead, we take the view that institutions can be built to enforce arbitrary amount of cooperation but they are costly. In sharp contrast to Lee et al. (2019), we allow the structure of social institutions to change and reflect the preferences of the population. Modeling social institutions as costly accumulation of capital enables us to characterize the social cost of myopic decisions regarding organizing a society.

2. The model

Consider a population of individuals, normalized to have unit mass. Time is continuous and each individual lives forever. Individuals are randomly receive opportunities to play the Prisoner's Dilemma Game with the utility function $u : \{C, D\}^2 \to \mathbb{R}$ described by the following matrix

	C	D
C	$1, 1$	$-l, d$
D	$d, -l$	$0, 0$

(1)

where $l > 0$ and $d > 1$. As is standard in the literature, we assume strategic complementarities in cooperation, that is, $d < 1 + l$. This assumption implies that the loss from cooperating is smaller when the opponent cooperates. Note that from this inequality it follows that the efficient outcome is (C, C), that is, $d - l < 2$. The opportunities to play arrive independently across agents and time according to a Poisson distribution with arrival rate one. Agents with opportunities are matched into pairs instantaneously. If the amount of *institutional capital* is $k \left(\in \mathbb{R}_+\right)$ then the players in the match are forced to cooperate with probability k.[4] Otherwise, the payoffs are determined according to the action profile chosen by the players. Each agent has one of two possible types: she is either an unconditional cooperator, γ, or strategic, σ. Unconditional cooperators always choose action C. On the other hand, a strategic individual plays action C only if she is forced to do so and chooses action D otherwise.

We assume that being strategic entails a flow fitness cost of τ for the individual and having an amount k of institutional capital requires a flow cost of $c(k)$ per capita. The function c is strictly convex, $c'(0) = 0$ and $c'(1) = 1$[5]

Payoffs and welfare.— First note that, since the arrival rate of opportunities is one, an individual's expected payoff from playing the game within a small dt-long time period is $\approx \overline{u}_t dt$, where $\overline{u}_t$ is the individual's expected utility from playing the game in (1) at time t. Consequently, an individual's expected utility, $\overline{u}_t$, is part of her flow payoff at time t. The individual's total flow payoff has two more additively separable components. The first one is the fitness cost τ which must be incurred only by strategic types and the second one is the cost of institutional capital. In what follows, we characterize the flow payoffs for both types as a function of the distribution of types and the amount of institutional capital.

Suppose that the fraction of strategic individuals is $\mu \in (0, 1)$ and the amount of institutional capital is k. Then when the match is monitored, with probability k, a strategic individual and her opponent will cooperate and get a payoff of one. When the match is not monitored, with probability $(1 - k)$, her opponent is another strategic individual with probability μ and is an unconditional cooperator with probability $(1 - \mu)$. In the former case, the payoff of the strategic individual is zero and in the latter one it is d. To summarize, the expected payoff flow of a strategic individual is

$$\pi^{\sigma}(k, \mu) = k + (1 - \mu)(1 - k)\, d - \tau - c(k).$$

Similarly, the match of a cooperator is monitored with probability k, in which case her payoff from the game is one. If her match is not monitored, her opponent defects with probability μ and cooperates with probability $(1 - \mu)$. Therefore, a cooperator's expected payoff flow is

$$\pi^{\gamma}(k, \mu) = k + (1 - \mu)(1 - k) - \mu(1 - k)\, l - c(k).$$

[3] In addition, there are no umpires who monitor other umpires.

[4] We interpret the institutional capital k as the strength of the legal environment which enables contracts to be enforced and facilitate cooperation among agents.

[5] This latter assumption is not innocuous and implies that inducing full cooperation is not prohibitively expensive. However, all our results remain valid if $c'(0) \leq 0$.

Note that the expected payoff flow of each type is determined by the fraction of strategic individuals, μ, and the amount of institutional capital, k. In what follows, we refer to the pair (k,μ) as the *state* of the environment. We denote the total payoff of the agents by $W(k,\mu)$ if the state of the environment is (μ,k), that is,

$$W(k,\mu)=\mu\pi^{\sigma}(k,\mu)+(1-\mu)\pi^{\gamma}(k,\mu). \tag{2}$$

We refer to $W(k,\mu)$ as the *welfare* of the society.[6]

The state of the environment at time t is denoted by $\left(k_t,\mu_t\right)$ and $\left(k_0,\mu_0\right)$ is referred to as the initial state.

Evolution of types and institutional capital.— We assume that the fraction of types depend on the relative payoffs of the two types. In particular, if the payoff of a certain type is larger than that of another type then the frequency of the more successful type increases in the population. Formally, the evolution of types is described by the following differential equation:

$$\dot{\mu}_t=\begin{cases}0 & \text{if } h\left(\pi^{\sigma}\left(k_t,1\right)-\pi^{\gamma}\left(k_t,1\right),1\right)>0 \text{ and } \mu_t=1\\ 0 & \text{if } h\left(\pi^{\sigma}\left(k_t,0\right)-\pi^{\gamma}\left(k_t,0\right),0\right)<0 \text{ and } \mu_t=0\\ h\left(\pi^{\sigma}\left(k_t,\mu_t\right)-\pi^{\gamma}\left(k_t,\mu_t\right),\mu_t\right) & \text{otherwise,}\end{cases} \tag{3}$$

where $h:\mathbb{R}\times[0,1]\to\mathbb{R}$ is continuously differentiable, and $h(x,\mu)>(<)\,0$ if $x>(<)\,0$. The first two cases of the definition above ensure that the distribution of types is bounded by $[0,1]$. Note that standard evolutionary dynamics such as myopic best-response and replicator dynamics are special cases of our general formulation.

We assume that the adjustment of the amount of institutional capital reflects the average preferences of individuals. In particular, the socially optimal level of institutional capital, $\widetilde{k}(\mu)$, is defined as

$$\arg\max_{k} W(k,\mu).$$

Then, if at a certain state (k,μ), $\widetilde{k}(\mu)>k$, the amount of institutional capital increases and it weakly decreases otherwise. Formally,

$$\dot{k}_t=H\left(\widetilde{k}\left(\mu_t\right)-k_t,k_t\right), \tag{4}$$

where $H:\mathbb{R}^2\to\mathbb{R}$ is continuously differentiable and $H(x,k)>(<)\,0$ if $x>(<)\,0$.

Equilibrium and Stability.— We consider a state an equilibrium if it is stationary, that is, neither the composition of types nor the amount of institutional capital changes over time. Formally, we call a state (k^*,μ^*) an *equilibrium* if $\left(k_t,\mu_t\right)=(k^*,\mu^*)$ for all $t>0$ whenever the initial state $\left(k_0,\mu_0\right)=(k^*,\mu^*)$.

We call an equilibrium stable if, after perturbing the equilibrium state locally, the state of the environment does not diverge away from the equilibrium. Formally, the equilibrium (k^*,μ^*) is *stable* if for all $\varepsilon>0$ there exists a $\delta>0$ such that if $\left\|\left(k_0,\mu_0\right)-(k^*,\mu^*)\right\|<\delta$ then

$$\left\|\left(k_t,\mu_t\right)-\left(k^*,\mu^*\right)\right\|<\varepsilon$$

for all $t>0$.

3. Results

This section states our main results. We first demonstrate that equilibria exist and they are Pareto-ranked. Then we prove our main result, that any equilibrium amount of institutional capital is larger than the long-run optimal one. The long-run optimal amount of capital is defined to be the welfare-maximizing level subject to the constraint that the type distribution in the population is in steady state, as determined by the myopic best-response dynamics described in Eq. (3).

3.1. Equilibrium existence and pareto ranking

We are ready to state our first result.

Proposition 1. *Generically, there are finitely many equilibria,* $\left\{\left(k_i^*,\mu_i^*\right)\right\}_1^n\subset\mathbb{R}^2_{++}$, *with* $k_{i+1}^*>k_i^*$ *for* $i=1,\ldots,n$. *Furthermore, for all* $i\in\{1,\ldots,n-1\}$,

(i) $\mu_i^*<\mu_{i+1}^*$ *and*

(ii) $W\left(k_i^*,\mu_i^*\right)>W\left(k_{i+1}^*,\mu_{i+1}^*\right)$.

In what follows, we present a geometric argument for the proof of this proposition. We explain that each equilibrium is at an intersection of two curves in the (k,μ) plane. We then show that these curves must have at least one intersection. Let us described the aforementioned two curves. First, for each k, we define $\widetilde{\mu}(k)$ to be the steady state fraction of strategic individuals if the amount of institutional capital is k forever. Second, recall that for each μ, the myopically optimal amount of capital is $\widetilde{k}(\mu)$. So if the fraction

[6] Defining welfare more generally by weighing different types differently would have no qualitative impact on our result.

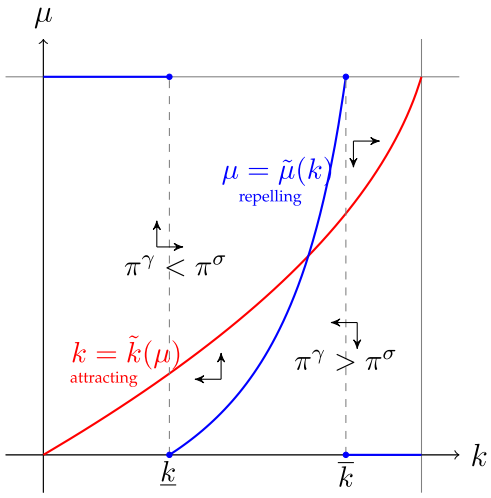

Fig. 1. The blue curve depicts the states in which the type-composition is stable, the red curve depicts the states in which institutional capital is myopically optimal. The arrows indicate the direction of the evolutionary dynamics in each region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of strategic types was μ forever then the steady state amount of capital is $\widetilde{k}(\mu)$. Observe that a state is stationary if and only if it is an intersections of the curves $\widetilde{\mu}$ and $\widetilde{k}$. That is, the set of equilibria coincides with the set of intersections of $\widetilde{\mu}$ and $\widetilde{k}$.

Steady-state Composition of Types.— Next, we characterize the correspondence $\tilde{\mu}$ which maps each k to the set of steady-state fraction of strategic types, $\tilde{\mu}(k)$. We show that the value of $\tilde{\mu}$ is a singleton except at two elements of its domain. We abuse notation and denote the value of $\tilde{\mu}$ by μ instead of $\{\mu\}$ whenever it is a singleton. Recall that if both types coexist in steady-state, then strategic and cooperator individuals must earn the same payoff. The following equality guarantees that $\pi^\sigma(k,\mu) = \pi^\gamma(k,\mu)$:

$$k + (1-\mu)(1-k)d - \tau - c(k) = k + (1-\mu)(1-k) - \mu(1-k)l - c(k).$$

Note that there is a threshold $\underline{k} \in (0,1)$ such that this equation has no solution in μ if $k < \underline{k}$. The reason is that if the amount of institutional capital is too small strategic individuals earn a higher payoff than cooperators irrespective of μ. Consequently, $\tilde{\mu}(k) = 1$ if $k < \underline{k}$. Similarly, there is another threshold $\overline{k} \in (\underline{k}, 1)$ such that if $k > \overline{k}$ the payoff of cooperators exceeds that of strategic individuals, so $\tilde{\mu}(k) = 0$ for $k > \overline{k}$. Otherwise, for each, $k \in \left[\underline{k}, \overline{k}\right]$, let $\tilde{\mu}(k)$ denote the solution for the previous displayed equation, that is,

$$\tilde{\mu}(k) = \frac{\frac{\tau}{1-k} - d + 1}{1 + l - d}. \tag{5}$$

Observe that $\tilde{\mu}$ is strictly increasing on $\left(\underline{k}, \overline{k}\right)$. Finally, if $k = \underline{k}$ or $\overline{k}$ then $\tilde{\mu}(k) = \{0,1\}$, that is, both zero and one are steady-state fractions.

Steady-state Amount of Capital.— Recall that $\widetilde{k}(\mu)$ is defined to be $\arg\max_k W(\mu,k)$, or equivalently, by the solution of the following first-order condition:

$$c'\left(\widetilde{k}(\mu)\right) = 1 - \mu(1-\mu)(d-l) - (1-\mu)^2.$$

Note that $\widetilde{k}(0) = 0$ because the right-hand side evaluated at zero is zero and $c'(0) = 0$. In addition, $\widetilde{k}(1) = 1$ because the right-hand side evaluated at one is one and $c'(1) = 1$. Moreover, $\widetilde{k}$ is a continuous function.

Existence and Ranking.— As explained above, a state (k^*, μ^*) is an equilibrium if and only if is an intersection of the curves $\tilde{\mu}$ and $\widetilde{k}$, that is, $\widetilde{k}(\mu^*) = k^*$ and $\tilde{\mu}(k^*) = \mu^*$. To argue that these curves intersect note that at $\underline{k}$, $\widetilde{k}^{-1} > \tilde{\mu}$ and that at $\overline{k}$, $\widetilde{k}^{-1} < \tilde{\mu}$, see Fig. 1 . Since both $\widetilde{k}^{-1}$ and $\tilde{\mu}$ are continuous, the Intermediate Value Theorem implies that the two curves intersect. Finally, note that, generically, there are only finitely many such intersections.

Let us turn our attention to the proof of parts (i) and (ii) of the proposition. Since each equilibrium is on the strictly increasing curve $\tilde{\mu}$, part (i) immediately follows. To see part (ii), consider two equilibria, (k_i, μ_i) and (k_{i+1}, μ_{i+1}) and recall that part (i) implies $\mu_i < \mu_{i+1}$. Note that the payoffs of both types are decreasing in the fraction of strategic individuals, μ, and hence,

$$\pi^\sigma(k_{i+1}, \mu_i) > \pi^\sigma(k_{i+1}, \mu_{i+1}) \text{ and } \pi^\gamma(k_{i+1}, \mu_i) > \pi^\gamma(k_{i+1}, \mu_{i+1}).$$

Also observe that the right-hand side of each of these inequalities is $W\left(k_{i+1}, \mu_{i+1}\right)$ because the sate $\left(k_{i+1}, \mu_{i+1}\right)$ is an equilibrium, so the payoffs of both types are the same and represent social welfare. Therefore, we conclude that the previous displayed inequalities

imply

$$\mu_i \pi^{\sigma}(k_{i+1}, \mu_i) + \left(1 - \mu_i\right) \pi^{\gamma}(k_{i+1}, \mu_i) > W\left(k_{i+1}, \mu_{i+1}\right). \tag{6}$$

Finally, part (ii) follows from the following inequality chain:

$$W\left(k_i, \mu_i\right) \geq W(k_{i+1}, \mu_i) = \mu_i \pi^{\sigma}(k_{i+1}, \mu_i) + \left(1 - \mu_i\right) \pi^{\gamma}(k_{i+1}, \mu_i) > W\left(k_{i+1}, \mu_{i+1}\right),$$

where the first inequality follows from $k_i = \arg\max_k W\left(k, \mu_i\right)$, the equality follows from the definition of W, and the second inequality is just (6).

Stability.— The dynamics that emerge from Eqs. (3) and (4) are such that the force driving institutional change is stabilizing whereas the force driving change in types is destabilizing. On one hand, the amount of institutional capital is being constantly driven towards the level which is optimal for the current composition of types. This means that institutional capital tends to revert back to a steady state level after a perturbation away from it, as illustrated in Fig. 1. On the other hand, the more strategic individuals there are, the higher their payoffs are relative to that of unconditional cooperators. This implies that a small increase in their fraction of strategic types away from the steady state would be amplified over time. Therefore, a steady-state is stable only if the former effect dominates the latter; that is, the institutional capital adjusts faster than the type distribution (see the Appendix for formal proofs). It is not unreasonable to expect the conditions for stability to hold in applications. Indeed, we interpret an individual's type in our model as a cultural trait and believe that the culture in a society, defined as the composition of types, is a slowly moving object relative to social institutions.

3.2. Institutional inefficiency

The primary goal of this section is to understand the inefficiency generated by the myopic adjustment of institutional capital while taking the dynamics of individuals' behavior as given. In particular, we compare k_i^* with the *long-run optimal* amount of institutional capital, denoted by k_l. We define k_l to be the welfare-maximizing amount of institutional capital subject to the constraint that the steady-state type distribution in the population is determined by the myopic best-response dynamics described in Eq. (3). Recall that if the institutional capital is k forever, the steady-state composition of types is given by $\widetilde{\mu}(k)$. Therefore, the long-run optimal capital is defined as follows:

$$k_l \in \arg\max_{k \in [0,1]} W\left(k, \widetilde{\mu}(k)\right).$$

The next proposition states the main result of our paper.

Proposition 2. *The long-run optimal social capital, k_l, is $\underline{k}$. Moreover, for all $i = 1, \ldots, n$,*
(i) $k_l < k_i^*$ and
(ii) $\widetilde{\mu}\left(k_l\right) = 0 < \mu_i^*$.

Let us now explain the proof of this proposition. First, we argue that $\underline{k}$ maximizes $W\left(k, \widetilde{\mu}(k)\right)$ on $k \in \left[\underline{k}, \overline{k}\right]$. Of course, on this interval, the social cost of accumulating institutional capital is minimized at $\underline{k}$. So, it is enough to argue that the aggregate payoffs from playing the Prisoner's Dilemma is maximized at $\underline{k}$. This follows from the observations that $\widetilde{\mu}\left(\underline{k}\right) = 0$, that is, the society is entirely inhibited by cooperators, and that the sum of the payoffs in the game is largest if both players cooperate.

To conclude that $k_l = \underline{k}$, we need to explain that it is suboptimal to set k to be outside of the interval $\left[\underline{k}, \overline{k}\right]$. Recall that if $k > \overline{k}$, the society is populated by only cooperators in the long-run, that is, $\widetilde{\mu}(k) = 0$ if $k > \overline{k}$. Since $\underline{k}$ results the same composition of types at a lower cost of capital, any $k > \overline{k}$ is dominated by $\underline{k}$. If $k < \underline{k}$, then

$$W\left(k, \widetilde{\mu}(k)\right) = W(k, 1) \leq W\left(\widetilde{k}(1), 1\right) \leq W\left(\underline{k}, 0\right) = W\left(\underline{k}, \widetilde{\mu}\left(\underline{k}\right)\right),$$

where the first equality follows from $\widetilde{\mu}(k) = 1$ for $k < \underline{k}$ and the first inequality from $\widetilde{k}(1)$ being the myopic welfare-maximizing amount of capital if $\mu = 1$. The second inequality follows from $\widetilde{k}(1) \geq \underline{k}$ (so $c\left(\widetilde{k}(1)\right) \geq c\left(\underline{k}\right)$) and that the aggregate payoffs from playing the game is maximized if $\mu = 0$. The last equality is implied by $\widetilde{\mu}\left(\underline{k}\right) = 0$. Finally, since $\underline{k} < k_i^*$ and $\widetilde{\mu}\left(\underline{k}\right) = 0$, parts (i) and (ii) immediately follow.

Paradoxically, in the long-run optimal steady state, the society is not only entirely populated by cooperators, but it also pays a lower cost to fund its social institutions. Indeed, k_l is smaller than any equilibrium level of institutional capital and it is also the smallest amount of capital that deters strategic types from entering the population. To resolve this paradox, let us explain the dynamics starting from the initial state $\left(\underline{k}, 0\right)$ in the context of an example which has a unique and globally stable equilibrium.

As Fig. 2 illustrates, the transition path cycles around and converges to the equilibrium. Each cycle consists of four phases. Initially, since there are no strategic types at $\left(\underline{k}, 0\right)$, the myopically optimal amount of capital is zero, so there will be disinvestment in institutional capital. This makes it more profitable for an individual to be strategic and the myopic best-response dynamics result in an increase of strategic types. This will continue until it is no longer myopically optimal to decrease capital because there is a sufficiently high fraction of strategic types. This period of transition corresponds to the segment of the transition path starting $\left(\underline{k}, 0\right)$ until the first intersection with the curve $\widetilde{k}$, see Fig. 2. At that point, the payoff of strategic types is still larger than that of cooperators, so more strategic individuals enter while the level of institutional capital increases. This phase lasts until the transition

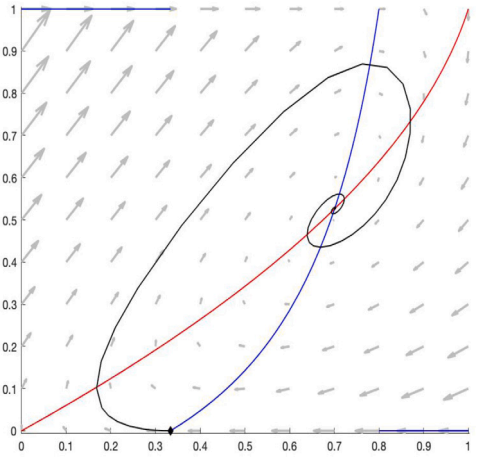

Fig. 2. Trajectories starting at the long run optimum $(k_0, \mu_0) = (\underline{k}, 0)$. In this graphs the population and the institutional capital evolve according to the best response dynamics given by $h(x, y, \mu) = b(x - y)$ and $H(x, y) = a(x - y)$ where $d = 1.3$, $l = 5$, $\tau = 0.2$, $a/b = 0.5$.

path reaches the curve $\widetilde{\mu}$. At that point, the strategic types will be worse off than the cooperators and their fraction will start to decrease. Once the transition path hits the curve $\widetilde{k}$, the fourth phase of transition begins, and both the fraction of strategic individuals and the amount of institutional capital decrease. When the transition path intersects with the $\widetilde{\mu}$, the cycle restarts and repeats. If there is a unique stable equilibrium, as in the example of Fig. 2, the transition path converges to it. In the end, the steady state fraction of strategic type is going to be large and the amount of institutional capital will maximize social welfare taking this fraction as given, and hence, this amount is larger than $\underline{k}$.

An alternative to model the accumulation of institutional capital is to consider a political process through which capital is adjusted. For example, it may move sluggishly towards the level that would maximize the payoff of the median voter. It is not hard to show that in such a model, even though there are two types of equilibria depending on whether the majority of the population is strategic or cooperator, our main result continues to hold in the following sense. When the median voter optimally adjusts capital while taking into account the evolution of types and subject to the constraint that she remains median, there is less capital accumulated than in equilibrium.

4. Conclusion

The objective of this simple model was to study the coevolution of cooperative cultures and the institutions that support them. We hypothesize that institutions tend to adjust myopically, optimizing for the current population without considering their impact on future cooperative dynamics. Our key finding is that this myopic adjustment not only leads to insufficient cooperation within society but also incurs excessive costs for sustaining social institutions. This observation underscores the importance of providing decision-makers, even those with benevolent intentions, with long-term incentives.

Appendix A

Stability

In order to simplify the conditions for a steady state to be stable, let us define $a : \mathbb{R} \to \mathbb{R}$ and $b : [0, 1] \to \mathbb{R}$ as follows,

$$a(k) \equiv H_1(0, k) \quad \text{and} \quad b(\mu) \equiv h_1(0, \mu).$$

These functions measure, respectively, the responsiveness of behavior to payoff differences between types, and the responsiveness of institutional capital to deviations from the myopically optimal level. Notice that, since the signs of $H(x, k)$ and $h(x, \mu)$ coincide with that of the first argument, x, the functions H and h must be increasing in x at 0. Therefore, $a(k) > 0$ and $b(\mu) > 0$ for all $k \in \mathbb{R}$ and all $\mu \in [0, 1]$ (see Fig. 3).

Proposition 3 (*Conditions for Stability of Steady States*)**.** *The steady state (μ_i^*, k_i^*) is stable if and only if i is odd*[7] *and*

$$\frac{a(k_i^*)}{b(\mu_i^*)} \geq \frac{\partial D}{\partial \mu}\Big|_{(k_i^*, \mu_i^*)}, \tag{7}$$

[7] Recall that there can be multiple steady states $\{(\mu_i^*, k_i^*)\}_{i \in \{1, \ldots, N\}}$ if the two loci (the curve of indifference between the two types and the myopically optimal institutional capital) intersect multiple times. The steady states are ordered so that $k_i^* < k_j^*$ if and only if $i < j$.

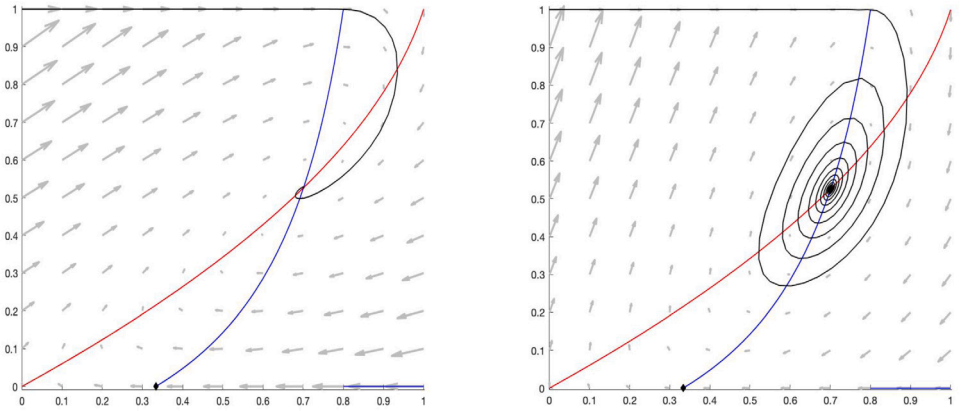

Fig. 3. In both of these graphs the population and the institutional capital evolve, starting at $(k_0, \mu_0) = (0, 1)$ according to the best response dynamics given by $h(x, y, \mu) = b(x - y)$ and $H(x, y) = a(x - y)$. Left: Trajectories when $d = 1.3$, $l = 5$, $\tau = 0.2$, $a/b = 1$. Right: Trajectories when $d = 1.3$, $l = 5$, $\tau = 0.2$, $a/b = 0.25$.

where $D(k, \mu) \equiv \pi^{\sigma}(k, \mu) - \pi^{\gamma}(k, \mu)$.

Proof. In order to analyze the local asymptotic stability of an equilibrium (k_i^*, μ_i^*) of the autonomous differential equation system described by (3), (4), let us consider the Jacobean matrix that approximates it linearly:

$$\begin{pmatrix} \dot{k} \\ \dot{\mu} \end{pmatrix} \approx J(k_i^*, \mu_i^*) \begin{pmatrix} k \\ \mu \end{pmatrix}$$

where

$$J(k, \mu) = \begin{pmatrix} \dfrac{dH}{dk}(\tilde{k}(\mu) - k, k) & \dfrac{dH}{d\mu}(\tilde{k}(\mu) - k, k) \\ \dfrac{dh}{dk}(D(k, \mu), \mu) & \dfrac{dh}{d\mu}(D(k, \mu), \mu) \end{pmatrix}.$$

Evaluating at the steady-state values (k_i^*, μ_i^*), it becomes

$$J_i^* \equiv J(k_i^*, \mu_i^*) = \begin{pmatrix} -a & a\dfrac{\partial \tilde{k}}{\partial \mu} \\ b\dfrac{\partial D}{\partial k} & b\dfrac{\partial D}{\partial \mu} \end{pmatrix}_{(k_i^*, \mu_i^*)}$$

because

$$\begin{aligned}
\frac{dH}{dk}(\tilde{k}(\mu) - k, k)\Big|_{(k_i^*, \mu_i^*)} &= -H_1(0, k_i^*) + H_2(0, k_i^*) = -a(k_i^*); \\
\frac{dH}{d\mu}(\tilde{k}(\mu) - k, k)\Big|_{(k_i^*, \mu_i^*)} &= H_1(0, k_i^*)\frac{\partial \tilde{k}}{\partial \mu}(\mu_i^*) = a(k_i^*)\frac{\partial \tilde{k}}{\partial \mu}(\mu_i^*); \\
\frac{dh}{dk}(D(k, \mu), \mu)\Big|_{(k_i^*, \mu_i^*)} &= h_1(0, \mu_i^*)\frac{\partial D}{\partial k}(k_i^*, \mu_i^*) = b(\mu_i^*)\frac{\partial D}{\partial k}(k_i^*, \mu_i^*); \\
\frac{dh}{d\mu}(D(k, \mu), \mu)\Big|_{(k_i^*, \mu_i^*)} &= h_1(0, \mu_i^*)\frac{\partial D}{\partial \mu}(k_i^*, \mu_i^*) + h_2(0, \mu_i^*) = b(\mu_i^*)\frac{\partial D}{\partial \mu}(k_i^*, \mu_i^*).
\end{aligned}$$

By the Hartman-Grobman theorem, a steady state s (i.e. s such that $\dot{s} = 0$) is locally stable if and only if all eigenvalues of $J(s)$ have negative real parts. Let λ_1, λ_2 be the eigenvalues. In the case of a 2×2 matrix that is equivalent to saying that the determinant is positive and the trace is negative.

The two necessary and sufficient conditions for stability are:

(T) $\mathrm{tr} J_i^* < 0$
(D) $\det J_i^* > 0$

First let us see that (D) holds if and only if i is odd. The determinant of that matrix has the opposite sign as the total derivative of the relative benefit of being a strategist at the equilibrium state because

$$\det J_i^* = -b(\mu_i^*)a(k_i^*)\left\{ \frac{\partial D}{\partial \mu}(k_i^*, \mu_i^*) + \frac{\partial D}{\partial k}(k_i^*, \mu_i^*)\frac{\partial \tilde{k}}{\partial \mu}(k_i^*, \mu_i^*) \right\}$$

$$= -b(\mu_i^*)a(k_i^*)\Delta'(\mu_i^*),$$

where $\Delta(\mu) \equiv D(\tilde{k}(\mu), \mu)$ is the relative benefit of being a strategist as a function of the fraction of strategists, when the institution is at its myopic best response to that fraction of strategists. First, let us see that $\Delta'(\mu_i^*) < 0$ if and only if i is odd. That is to say, when the equilibrium is such that the curve $\tilde{\mu}$ intersects the curve $\tilde{k}$ from above.

Second, notice that (T) is equivalent to inequality (7) because $\mathrm{tr}J_i^* = -a(k_i^*) + b(\mu_i^*)\frac{\partial D}{\partial \mu}(k_i^*, \mu_i^*)$. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.euroecorev.2023.104620.

References

Axelrod, R., 1981. The Emergence of Cooperation among Egoists. Amer. Polit. Sci. Rev. 75 (2), 306–318.

Bisin, A., Verdier, T., 2017. On the joint evolution of culture and institutions. NBER Working Paper. Working Paper.

Boyd, R., Richerson, P.J., 2009. Culture and the evolution of human cooperation. Philos. Trans. R. Soc. B 364 (1533), 3281–3288. http://dx.doi.org/10.1098/rstb.2009.0134, arXiv:arXiv:1011.1669v3.

Davis, L., North, D., 1970. Institutional change and American economic growth: A first step towards a theory of institutional innovation. J. Econom. Hist. 30 (1), 131–149.

Dawes, R.M., Thaler, R.H., 1988. Anomalies: Cooperation. J. Econ. Perspect. 2 (3), 187–197. http://dx.doi.org/10.1257/jep.2.3.187, URL http://pubs.aeaweb.org/doi/10.1257/jep.2.3.187.

Fehr, E., Fischbacher, U., 2003. The nature of human altruism. Nature 425 (6960), 785–791. http://dx.doi.org/10.1038/nature02043.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.

Gintis, H., 2003. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. J. Theoret. Biol. 220 (4), 407–418. http://dx.doi.org/10.1006/jtbi.2003.3104.

Klick, J., Tabarrok, A., 2010. Police, prisons, and punishment: the empirical evidence on crime deterrence. In: Handbook on the Economics of Crime. Edward Elgar, Cheltenham, UK.

Lee, J.-H., Iwasa, Y., Dieckmann, U., Sigmund, K., 2019. Social evolution leads to persistent corruption. Proc. Natl. Acad. Sci. 116 (27), 13276–13281. http://dx.doi.org/10.1073/pnas.1900078116, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1900078116.

Migliaccio, E., Verdier, T., 2018. On the spatial diffusion of cooperation with endogenous matching institutions. Games 9 (3), 58. http://dx.doi.org/10.3390/g9030058, URL http://www.mdpi.com/2073-4336/9/3/58.

Nowak, M.A., 2012. Evolving cooperation. J. Theoret. Biol. 299, 1–8. http://dx.doi.org/10.1016/j.jtbi.2012.01.014.

Perc, M., Szolnoki, A., 2010. Coevolutionary games - a mini review. BioSystems 99 (2), 109–125. http://dx.doi.org/10.1016/j.biosystems.2009.10.003, arXiv:0910.0826.

Sasaki, T., Yamamoto, H., Okada, I., Uchida, S., 2017. The evolution of reputation-based cooperation in regular networks. Games 8 (1), 8. http://dx.doi.org/10.3390/g8010008, arXiv:1701.06153.

Sethi, R., Somanathan, E., 1996. The evolution of social norms in common property resource use. Amer. Econ. Rev. 86 (4), 766–788. http://dx.doi.org/10.2307/2118304.

Slemrod, J., 2007. Cheating ourselves: The economics of tax evasion. J. Econ. Perspect. 21 (1), 25–48. http://dx.doi.org/10.1257/jep.21.1.25.

Tabellini, G., 2008. The scope of cooperation : Values and incentives. Q. J. Econ. 123 (3), 905–950.