On the Coevolution of Cooperation and Social Institutions

Verónica Salazar* and Balázs Szentes†

June 22, 2021

Abstract

This paper considers an environment cohabited by selfish individuals and unconditional cooperators. Individuals are randomly matched and play the Prisoner's Dilemma Game. Institutional capital facilitates cooperation among players and selfish individuals have to pay a cost in order to identify situations where defection is unpunished. This paper considers the coevolution of types and institutional capital. Both the type distribution and capital evolve according to a myopic best-response dynamics. The set of equilibria are shown to be Pareto ranked. The main result is that any equilibrium level of institutional capital exceeds the long-run optimal amount. So, the long-run optimal institutions don't only maintain a more cooperative culture but are also cheaper than equilibrium ones.

1 Introduction

In human societies, cooperation is often promoted by the establishment of institutions that monitor and punish anti-social behaviour. Examples include police forces that identify and prevent criminal acts and judicial systems that resolve disputes and enforce contracts. Such institutions are costly to develop and maintain and there is a large variation in their quality across societies. Individuals may respond to existing institutions either by cooperating or by exerting an effort to avoid being punished while engaging in non-cooperative behaviour. Of course, the profitability of these choices depends on the strength and quality of social institutions. In turn, the society's optimal institutional investment is largely determined by the behavioural culture in the society, e.g. an increase in criminal activity justifies expanding the police force. The goal of this paper is to study how cooperative behaviour and social institutions coevolve and to understand the welfare consequences of their mutual dependency.

We consider a stylized model where individuals play the Prisoner's Dilemma repeatedly against random opponents. The payoffs in the game are defined so that there is complementarity in cooperation in the sense that the expected gain from switching from defection to cooperation is increasing in the number of cooperators. A fraction of these interactions is monitored and the cooperative action is enforced. Crucially, this fraction is determined by the amount of *institutional capital* accumulated in the society. Individuals have one of two possible types: *cooperator* or *strategic*. Cooperators always cooperate while strategic individuals are able to condition their actions on whether their play is monitored. Strategic individuals must pay a fitness cost in order to be able to distinguish between monitored and unmonitored interactions. The type

*Department of Economics, London School of Economics. E-mail: v.salazar-restrepo@lse.ac.uk.
†Department of Economics, London School of Economics. E-mail: b.szentes@lse.ac.uk.

distribution evolves according to a standard best-response dynamic: at any point in time, the frequency of a type increases continuously if its payoff is higher than that of the other type. Similarly, institutional capital adjusts myopically and sluggishly towards the level that would be socially optimal given the type distribution. This adjustment is myopic, because it does not take into account how a change in institutions will further modify behaviour through affecting the type-composition, and sluggish, because it only allows for gradual changes to the current level of institutions. Thus, the dynamics of the economy are fully determined by a two-dimensional state: the fraction of strategic types and the amount of institutional capital. A state is an *equilibrium* when neither the distribution of types nor the institutional capital changes. In particular, the two types can coexist in equilibrium only if their payoffs are the same. Furthermore, the equilibrium amount of institutional capital is myopically optimal given the equilibrium composition of types.

Our first result is that equilibria always exist, and, if there are multiple, they can be Pareto-ranked. Equilibria with a higher level of institutional capital will also have a higher fraction of strategic individuals and smaller social welfare. To give an explanation, suppose that the economy is in an equilibrium and consider a surge of strategic individuals so that the population's type-composition corresponds to another equilibrium distribution. Since there is complementarity in cooperation, the payoff of a strategic individual exceeds that of a cooperator after this surge. In order to offset this payoff gain, social institutions must make it more difficult for strategic individuals to exploit cooperators. So, to guarantee that the two types are indifferent, the amount of institutional capital must also be larger in the other equilibrium. In the meantime, notice that both the surge of strategic individuals and the increase in capital lower the payoff of both types, and hence welfare. The reason is that cooperative play becomes less frequent and strengthening institutions is costly.

We then turn our attention to the long-run optimal amount of institutional capital. This quantity is defined to maximise social welfare while taking into account that the distribution of types 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 of capital. In fact, we show that the long-run optimum is the smallest amount of capital which prevents strategic types entering the economy. It turns out that this amount is smaller than any equilibrium level of capital. Paradoxically, if the institutional capital is set to be the long-run optimum, the society is not only entirely populated by cooperators in steady state but also save cost of accumulating institutional capital.

Optimal social institutions are built to maintain a cooperative culture and prevent its deterioration at minimal costs. However, these institutions do not reflect the preferences of the current population. Indeed, since there are only cooperators in the long-run optimal steady state, there is no reason to monitor their interactions and the myopically optimal amount of capital is zero. When investment in social institutions are motivated by short-term goals, these institutions are weakened which leads to an inflow of strategic individuals and less cooperation. Deaccumulation of institutional capital continues until there are so many strategic individuals in the society that it is no longer myopically optimal to further decrease capital. At that point, the payoff of a strategic individual is still larger than that of a cooperator, so more strategic individuals enter but the level of institutional capital increases. At the end, the fraction of strategic type is going to be large and the amount of institutional capital will be larger than the long-run optimal amount. By no means does this argument imply that, in an equilibrium, welfare can be improved by disinvesting institutional capital. Indeed, in order to transit to the long-run optimum, institutions must first be strengthened to eliminate anti-social behaviour. Only after a cooperative culture is established, capital can be disinvested.

It is well-understood and studied that social institutions promote 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 also clear that short-sighted decision making (e.g. politician with re-election concerns and myopic voters) may result in inefficient investment in these institutions. The main takeaway from our paper is that these inefficiencies do not only lead to a deterioration of a cooperative culture but also to social institutions which are too strong and too expensive to maintain.

Literature Review. The ubiquity of cooperation in strategic interactions between unrelated individuals is the focus of a large game-theoretic literature, which seek to explain behaviour 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 co-evolution 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 centralised enforcement of behaviours 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 behaviour and institutions. Some exceptions include Tabellini (2008), Bisin and Verdier (2017), and Migliaccio and Verdier (2018). Tabellini (2008) assigns a central role to the spatial patterns of individual values and legal enforcement. Its key finding with respect institutions 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". 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 behaviour is enforced. Monitoring and enforcement in their framework would be the outcome of a policy that is determined by

[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).

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 (2018)'s 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).

Some empirical studies have looked at the endogeneity of institutions to particular environments as well as the endogeneity of pro-social behaviours to institutional histories[3]. Examples include Acemoglu et al. (2012) and Lowes et al. (2017). The former shows that lower settler mortality at the times of colonisation led to the establishment of more inclusive institutions that persisted over centuries. Lowes et al. (2017) look at the long-lasting impacts of the pre-colonial Kuba Kingdom in Central Africa on the current culture and values in the contemporary descendants of its inhabitants. Their result is that people whose ancestors lived under a society with more centralized enforcement exhibit less pro-social preferences.

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 Poison distribution with arrival rate one. Agents with opportunities are matched into pairs instantaneously. If the amount of *institutional capital* is $k\,(\in \mathbb{R}_+)$ 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 k amount 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 opportinities 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

[3]For a review of the empirically-oriented literature exploring the coevolution of institutions and cooperative behaviour, and culture more generally, see Alesina and Giuliano (2015).

[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.

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 in a given moment. Then a strategic individual cooperates with probability k in which case her opponent also cooperates and receives a payoff of one. If her 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 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 is

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

Note that the expected payoff 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 (k_t,μ_t) and (k_0,μ_0) 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:

$$\dot{\mu_t} = h\left(\pi^{\sigma}(k_t,\mu_t), \pi^{\gamma}(k_t,\mu_t), \mu_t\right), \tag{3}$$

where $h : \mathbb{R}^2 \times [0,1] \to \mathbb{R}$ is continuously differentiable, $h(x,y,\mu) > (<)\, 0$ if $x > (<)\, y$. Note that standard evolutionary dynamics such as myopic best-response and replication 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

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

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

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

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

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

Equilibrium and Stability.— We consider a state an equilibrium if it is absorbing, 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 $(k_t,\mu_t) = (k^*,\mu^*)$ for all $t > 0$ whenever the initial state $(k_0,\mu_0) = (k^*,\mu^*)$.

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 $\|(k_0,\mu_0) - (k^*,\mu^*)\| < \delta$ then

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

for all $t > 0$.

3 Results

This section states our three 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.

3.1 Equilibrium Existence and Pareto Ranking

We are ready to state our first result.

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

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

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

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 of strategic types was μ forever then the steady state amount of capital is $\widetilde{k}(\mu)$. Observe that a state is absorbing 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.— In what follows, for each k we compute the steady-state fraction of strategic types, $\widetilde{\mu}(k)$. Recall that if both types coexist in steady-state, $\widetilde{\mu}(k) \in (0,1)$, then strategic and cooperator individuals must earn the same payoff. The following equality guarantees that $\pi^\sigma(k,\mu) = \pi^\gamma(k,\mu)$:

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

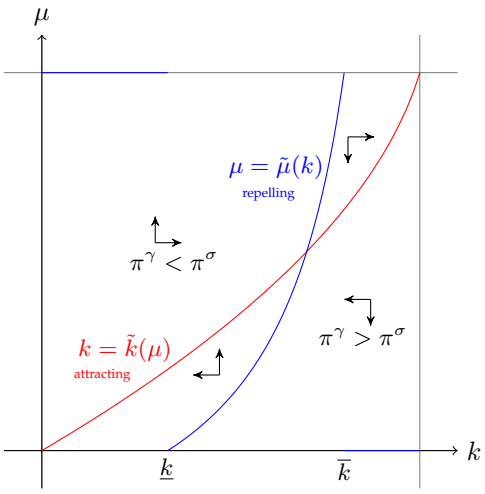

Figure 1: The blue curve depicts the states in which the type-composition is stable, the red curve depicts the states in which institutional capital is myopically optimal. The arrows indicate the direction of the evolutionary dynamics in each region.

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 $\left[\underline{k}, \overline{k}\right]$ because its derivative is $\tau / \left[(1 + l - d)(1-k)^2\right] > 0$.

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 Figure 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 parts (i) and (ii) of the proposition. Since each equilibrium is on the 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 $W(k_{i+1}, \mu_{i+1})$ because the sate (k_{i+1}, μ_{i+1}) 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

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

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

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

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

Stability.— The dynamics that emerge from equations (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 Figure 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. We formalize this result in the Appendix.

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' behaviour as given. In particular, we compare k_i^* with the *long-run optimal* amount of institutional capital, denoted by k_l. We define the k_l to be the welfare-maximising 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 equation (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:

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

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 $i = 1, ..., n$,*
(i) $k_l < k_i^*$ and
(ii) $\widetilde{\mu}(k_l) = 0 < \mu_i^*$.

Let us now explain the proof of this proposition. First, we argue that $\underline{k}$ maximizes $W(k, \widetilde{\mu}(k))$ on $k \in [\underline{k}, \overline{k}]$. 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 $\widetilde{\mu}(\underline{k}) = 0$, 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

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

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 $c\left(\widetilde{k}(1)\right) \geq c(\underline{k})$) and that the aggregate payoffs from playing the game is maximized if $\mu = 0$. The last equality is implied by $\widetilde{\mu}(\underline{k}) = 0$. Finally, since $\underline{k} < k_i^*$ and $\widetilde{\mu}(\underline{k}) = 0$, 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 what happens if capital evolves myopically at the initial state $(\underline{k}, 0)$. Since there are no strategic types, the myopically optimal amount of capital is zero. Therefore, capital is myopically adjusted downwards. This, in turn, makes it more profitable for an individual to be strategic and the myopic best-response dynamics result in an increase of strategic types. Initially, the fraction of strategic types is small and the amount of capital further decreases which makes it even more profitable for strategic types to enter. This process continues until there is a significant number of strategic individuals in the society and it is no longer myopically optimal to decrease capital. At that point, the payoff of strategic types is still larger than that of cooperators, so more strategic individuals enter but the level of institutional capital increases. At 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

Our objective in this paper was to examine how cooperative cultures and institutions that maintain those cultures coevolve in societies. We entertain the hypothesis that institutions are adjusted myopically towards those that would be optimal given the current population while ignoring how these institutions affect the future dynamics of cooperative culture. Our main result is that the welfare cost of this myopic adjustment is not only that there is too little cooperation in the society but also that sustaining the social institutions is too expensive. This observation highlights the importance of providing decision-makers with long-term incentives even if they are benevolent.

References

D. Acemoglu, S. Johnson, and J. A. Robinson. The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, 102(6):3077–3110, May 2012. doi: 10.1257/aer.102.6.3077. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.6.3077.

A. Alesina and P. Giuliano. Culture and institutions. *Journal of Economic Literature*, 53(4):898–944, 2015. ISSN 00428736. doi: 10.32609/0042-8736-2016-10-82-111.

R. Axelrod. The Emergence of Cooperation among Egoists. *The American Political Science Review*, 75(2): 306–318, 1981.

A. Bisin and T. Verdier. On the Joint Evolution of Culture and Institutions. Working Paper, 2017.

R. Boyd and P. J. Richerson. Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288, 2009. ISSN 14712970. doi: 10.1098/rstb.2009.0134.

L. Davis and D. North. Institutional Change and American Economic Growth: A First Step Towards a Theory of Institutional Innovation. *The Journal of Economic History*, 30(1):131–149, 1970.

R. M. Dawes and R. H. Thaler. Anomalies: Cooperation. *Journal of Economic Perspectives*, 2(3):187–197, 1988. ISSN 0895-3309. doi: 10.1257/jep.2.3.187. URL http://pubs.aeaweb.org/doi/10.1257/jep.2.3.187.

E. Fehr and U. Fischbacher. The nature of human altruism. *Nature*, 425(6960):785–791, 2003. ISSN 1476-4687. doi: 10.1038/nature02043. URL https://doi.org/10.1038/nature02043.

E. Fehr and S. Gächter. Altruistic punishment in humans. *Nature.*, 415(6868):137–140, 2002. ISSN 0028-0836.

H. Gintis. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. *Journal of Theoretical Biology*, 220(4):407–418, 2003. ISSN 00225193. doi: 10.1006/jtbi.2003.3104.

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

S. Lowes, N. Nunn, J. A. Robinson, and J. L. Weigel. The Evolution of Culture and Institutions: Evidence From the Kuba Kingdom. *Econometrica*, 85(4):1065–1091, 2017. ISSN 0012-9682. doi: 10.3982/ecta14139.

E. Migliaccio and T. Verdier. On the Spatial Diffusion of Cooperation with Endogenous Matching Institutions. *Games*, 9(3):58, 2018. ISSN 2073-4336. doi: 10.3390/g9030058. URL http://www.mdpi.com/2073-4336/9/3/58.

M. A. Nowak. Evolving cooperation. *Journal of Theoretical Biology*, 299:1–8, 2012. ISSN 0022-5193. doi: 10.1016/j.jtbi.2012.01.014. URL http://dx.doi.org/10.1016/j.jtbi.2012.01.014.

M. Perc and A. Szolnoki. Coevolutionary games - A mini review. *BioSystems*, 99(2):109–125, 2010. ISSN 03032647. doi: 10.1016/j.biosystems.2009.10.003.

T. Sasaki, H. Yamamoto, I. Okada, and S. Uchida. The Evolution of Reputation-Based Cooperation in Regular Networks. *Games*, 8(1):8, 2017. ISSN 2073-4336. doi: 10.3390/g8010008. URL http://www.mdpi.com/2073-4336/8/1/8.

R. Sethi and E. Somanathan. The Evolution of Social Norms in Common Property Resource Use. *American Economic Review*, 86(4):766–788, 1996. ISSN 00028282. doi: 10.2307/2118304.

J. Slemrod. Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives*, 21(1):25–48, 2007. ISSN 08953309. doi: 10.1257/jep.21.1.25.

G. Tabellini. The Scope of Cooperation : Values and Incentives. *Quarterly Journal of Economics*, 123(3): 905–950, 2008.

Appendix

Stability

In order to characterise the stable steady states, let us add a further assumption, that is, that the function $H : \mathbb{R}^2 \to \mathbb{R}$ and $h : \mathbb{R}^2 \times [0,1] \to \mathbb{R}$ are symmetric with respect to the first two arguments around the region where the first two arguments are equal. Formally,

Assumption (Symmetry of dynamics) For all $x \in \mathbb{R}$ and $\mu \in [0,1]$

$$\begin{aligned} H_1(x,x) &= -H_2(x,x) \\ h_1(x,x,\mu) &= -h_2(x,x,\mu). \end{aligned} \tag{7}$$

For the population dynamics, this means that increasing the payoff of the cooperators respect to the is equivalent to decreasing the payoff of the strategists. For the institutional dynamics, it means that what determines the speed of institutional adjustment after a deviation from the myopic optimum is the difference from the myopic target and not whether it is the result of a higher target or a lower current level.
Now we are ready to characterise the local stability of steady states.

Proposition 3 (Conditions for stability of steady states)**.** *The steady state (μ_i^*, k_i^*) is stable if and only if i is odd and*

$$\frac{\rho_k(\mu_i^*)}{\rho_\mu(k_i^*)} \geq \frac{\partial(\pi^\sigma - \pi^\gamma)}{\partial \mu}\Big|_{(k_i^*, \mu_i^*)}. \tag{8}$$

where

$$\rho_k(\mu) \equiv H_1(\tilde{k}(\mu), \tilde{k}(\mu)) = H_2(\tilde{k}(\mu), \tilde{k}(\mu))$$

and

$$\rho_\mu(k) \equiv h_1(\pi^\sigma(k, \tilde{\mu}(k)), \pi^\gamma(k, \tilde{\mu}(k)), \tilde{\mu}(k)) = h_2(\pi^\sigma(k, \tilde{\mu}(k)), \pi^\gamma(k, \tilde{\mu}(k)), \tilde{\mu}(k)).$$

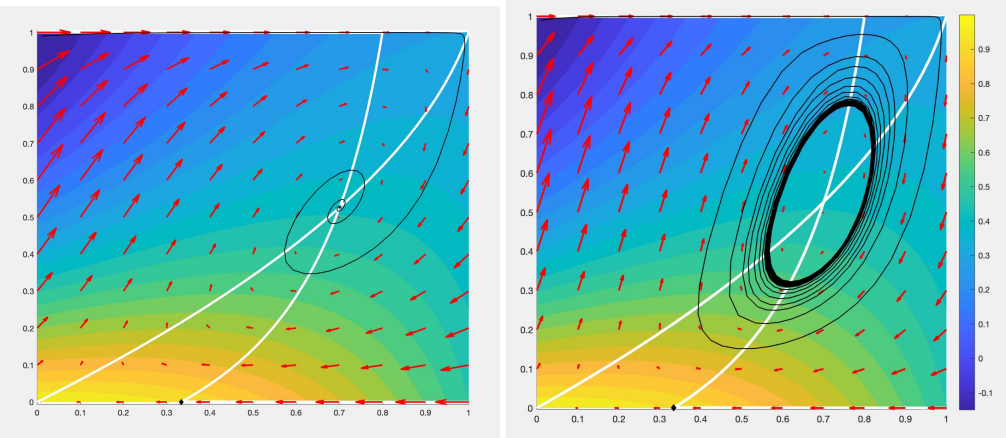

Figure 2: In both of these graphs the population dynamic is a replicator equation $h(x,y,\mu) = \rho_\mu \mu(1-\mu)(x-y)$ and the instutional dynamics is the myopic best response dynamic represented by $H(x,y) = \rho_k(x-y)$. Left: Trajectories when $d = 1.3$, $l = 5$, $\tau = 0.2$, $\rho_k/\rho_\mu = 0.1$. Right: Trajectories when $d = 1.3$, $l = 5$, $\tau = 0.2$, $\rho_k/\rho_\mu = 0.05$.

In order to analyse 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

$$J(k_i^*, \mu_i^*) = \begin{pmatrix} -\rho_k & \rho_k \dfrac{\partial \tilde{k}}{\partial \mu} \\ \\ \rho_\mu \dfrac{\partial(\pi^\sigma - \pi^\gamma)}{\partial k} & -\rho_\mu \dfrac{\partial(\pi^\sigma - \pi^\gamma)}{\partial \mu} \end{pmatrix}_{(k_i^*, \mu_i^*)}$$

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) $\text{tr} J_i^* < 0$

(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

$$\begin{aligned} \det J_i^* &= -\rho_\mu \rho_k \Big\{ \frac{\partial(\pi^\sigma - \pi^\gamma)}{\partial \mu}\Big|_{(k_i^*, \mu_i^*)} + \frac{\partial(\pi^\sigma - \pi^\gamma)}{\partial k}\Big|_{(k_i^*, \mu_i^*)} \frac{\partial \tilde{k}}{\partial \mu}\Big|_{(k_i^*, \mu_i^*)} \Big\} \\ &= -\rho_\mu \rho_k D'(\mu_i^*), \end{aligned}$$

where $D(\mu) \equiv \pi^\sigma(\tilde{k}(\mu), \mu) - \pi^\gamma(\tilde{k}(\mu), \mu)$ 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. Now let us see that $D'(\mu_i^*) < 0$ if and only if i is odd. That is to say, when the the equilibrium is such that the curve $\tilde{\mu}$ intersects the curve $\tilde{k}$ from above.

Now let us see that (T) is equivalent to inequality (8) because $\text{tr} J_i^* = -\rho_k - \rho_\mu \mu_i^* (1 - \mu_i^*) \dfrac{\partial(\pi^\sigma - \pi^\gamma)}{\partial \mu}\Big|_{(k_i^*, \mu_i^*)}$ and that is always negative by the assumption of strategic complementarities.