The social leverage effect: Institutions transform weak reputation effects into strong incentives for cooperation

Julien Lie-Panis*[a], Léo Fitouchi†[b], Nicolas Baumard[c], and Jean-Baptiste André[c]

[a] *Max Planck Research Group Dynamics of Social Behavior, Max Planck Institute for Evolutionary Biology, 24306, Plön, Germany*

[b] *Institute for Advanced Studies in Toulouse, Université Toulouse 1 Capitole, 31080 Toulouse, France*

[c] *Institut Jean Nicod, Département d'études cognitives, Ecole normale supérieure, Université PSL, EHESS, CNRS, 75005 Paris, France*

April 8, 2024

Abstract

Institutions allow cooperation to persist when reciprocity and reputation provide insufficient incentives. Yet how they do so remains unclear, especially given that institutions are themselves a form of cooperation. To solve this puzzle, we develop a mathematical model of reputation-based cooperation in which two social dilemmas are nested within one another. The first dilemma, characterized by high individual costs or insufficient monitoring, cannot be solved by reputation alone. The second dilemma, an institutional collective action, involves individuals contributing to change the parameters of the first dilemma in a way that incentivizes cooperation. Our model demonstrates that this nested architecture creates a leverage effect. While insufficient on its own to incentivize cooperation in the first dilemma, reputation incentivizes contributions to the institutional collective action which, in turn, strengthen the initially weak incentives for cooperation in the first dilemma. Just as a pulley system transforms minimal muscular strength into significant lifting capability, institutions act as cooperative pulleys, transforming initially weak reputational incentives into powerful drivers of cooperative behavior. Based on these results, we suggest that institutions have developed as social technologies, designed by humans to exploit this social leverage effect, just as material technologies are designed to exploit physical laws.

*Corresponding author. Email: `jliep@protonmail.com`

†Corresponding author. Email: `leo.fitouchi@gmail.com`

Large-scale cooperation is central to the success of the human species (Henrich & Muthukrishna, 2021). Yet its origins remain poorly understood. Canonical explanations, such as kin altruism (Hamilton, 1963; Ohtsuki et al., 2006), reciprocity (Axelrod & Hamilton, 1981; Barclay, 2020; Trivers, 1971), and reputation (Barclay et al., 2021; Giardini & Vilone, 2016; Lie-Panis & André, 2022; Nowak & Sigmund, 1998; Panchanathan & Boyd, 2003; Quillien, 2020), seem insufficient to explain the scale and intensity of human cooperation. In large human societies, more often than not, partners are unrelated, interactions are one-shot, and reputational information is narrowly disseminated (Lehmann et al., 2022; Powers et al., 2021).

The social sciences have long recognized that institutions play a crucial role in surmounting these challenges. Humans have designed social organizations such as clans (Schulz et al., 2019), age sets (Lienard, 2016), merchant guilds (Greif et al., 1994), assemblies (Hadfield & Weingast, 2013), governments (Fukuyama, 2011), and justice systems (Fitouchi & Singh, 2023; Milgrom et al., 1990; Sznycer & Patrick, 2020), that make rules of good behavior explicit, specify role-specific obligations, and organize the monitoring and punishment of free-riders (Currie et al., 2021; Gavrilets & Currie, 2022). Essentially, these organizations solve the free-rider problem by instituting new incentives for cooperation (North, 1990; Powers et al., 2016).

Institutions, however, are themselves cooperative enterprises, and as such they face a second-order free-rider problem (Dixit, 2018; Ostrom, 1990; Persson et al., 2013; Yamagishi, 1986). People must devote time and resources to create new rules and pay institutional operatives. These operatives, in turn, must resist corruption; they must, for instance, rebuff bribes (Muthukrishna et al., 2017) and avoid abuses of power (Acemoglu & Robinson, 2013). In other words, saying that institutions stabilize cooperation seems to only push the problem one step further: What stabilizes institutions?

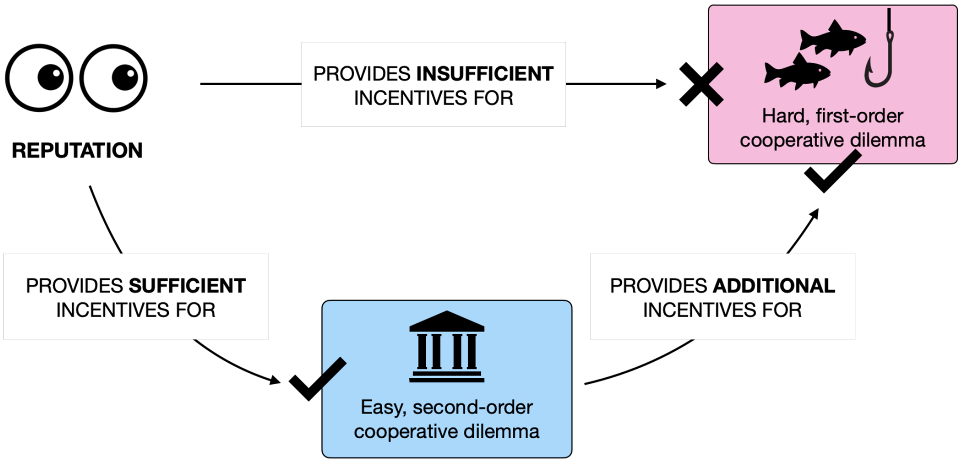
REPUTATION
PROVIDES **INSUFFICIENT** INCENTIVES FOR
Hard, first-order cooperative dilemma
PROVIDES **SUFFICIENT** INCENTIVES FOR
Easy, second-order cooperative dilemma
PROVIDES **ADDITIONAL** INCENTIVES FOR

Figure 1: **Institutions allow reputation to solve hard cooperation problems indirectly**. Reputation can solve hard cooperation problems indirectly, by incentivizing an easier form of second-order cooperation, which in turn increases the incentive to cooperate at the first order. By engineering an institution based on such a form of second-order cooperation, humans engineer a technological solution to a hard cooperation problem, using only the limited reputational incentives at their disposal.

In this paper, we present a mathematical model of institutions, that sheds light on how they can stabilize cooperation while themselves relying on cooperation. We show that institutions do more than just push the problem one step further; they effectively solve the problem. This solution is achieved through a social leverage effect that arises from the nesting of multiple collective actions within one another.

Our premise is that cooperative dilemmas vary in difficulty. Some cooperative dilemmas are hard; because the temptation to cheat is high, because cheaters are unlikely to be observed, or because the dilemma involves many unrelated individuals. Other cooperative dilemmas are easy; because cooperation is cheap, behaviors are observable, and interactions occur within small groups of kith and kin.

Humans need not tackle hard cooperation problems head on. Instead, they can design another cooperative interaction that is easier to solve (e.g., because behaviors are more observable), and that generates new incentives for cooperation in the hard dilemma (e.g., by organizing the monitoring of free-riding). Institutions, we argue, consist of these easy dilemmas, within which hard cooperation problems are embedded. If the cost of institutional cooperation is low enough to be driven solely by reputational concerns, and the institution generates enough new incentives to solve the initial hard dilemma, cooperation becomes indirectly solvable through reputation (see Figure 1). By creating a nested architecture of dilemmas, institutions create a leverage effect that can amplify the power of reputation, analogous to how levers amplify physical forces.

Take a historical example. In rural Japan, villagers needed to cooperate to preserve communal forests from overuse (McKean, 1986; Ostrom, 1990, pp. 65-69). This cooperation problem was hard: it was strongly in each villager's interest to overuse the communal forest, and it was difficult to check that no one was doing so. To solve this hard problem, villages hired specialized monitors called detectives, whose job was to spot and impose fines on free-riders, thereby generating new incentives for cooperation. This institution was itself a cooperative enterprise: for the whole thing to work, detectives had to do their job faithfully, instead of soliciting bribes or exacting unfair penalties. Yet the underlying cooperation problem was easier: if they abused their power, detectives were likely to be spotted, and, thus, to lose their hard-earned reputation. By hiring detectives, the villagers had found a way to solve their hard problem indirectly, using only the limited reputational incentives at their disposal.

We formalize this idea using the mathematical model below. Our model focuses on individuals who can cooperate in two different ways: sometimes they can pay to help an individual partner (first-order cooperation), and sometimes they can pay to contribute to an institution (second-order cooperation). In both cases, the only benefit they gain from cooperation is reputational. Each time individuals are observed cooperating, whether at the first- or second-order, they enhance their reputation, and become more likely to be trusted by partners in the future.

The institution collects individual contributions, and transforms them into incentives for first-order cooperation. We show that the institution extends the domain of reputation-based cooperation, to include hard cooperation problems. What's more, we show that the amount of additional cooperation generated by the institution varies with its *efficiency*—the amount of incentives the institution produces for every resource unit it receives. This underscores the idea that institutions should be viewed as a social technology. Just as a pulley system helps lift heavy loads with minimal effort, institutions maximize the potential of reputational incentives, helping humans address hard cooperation problems that reputation couldn't solve directly. Institutions appear as social engineering tools that humans have invented and gradually refined to build the most mutually beneficial social organizations that can be sustained by reputation alone.

Model

Life of an individual actor

We consider a repeated game with two types of individuals, actors and choosers. Actors are long-run players: they play all infinite rounds of the repeated game. Choosers are short-run players: they play only one round. For mathematical convenience, actors and choosers are members of two separate populations of infinite size.

Our model focuses on actors (see Figure 2). Actors can cooperate in two ways: sometimes they can pay to help an individual chooser (first-order cooperation), and sometimes they can pay to contribute to an institution (second-order cooperation). More precisely, in each round, actors either play a trust game, with probability q, or they play the institution game, with probability $1 - q$. A trust game is played with a chooser. The institution game is played with the fraction $1 - q$ of the actor population that draws that game in that round. Both games are described below.

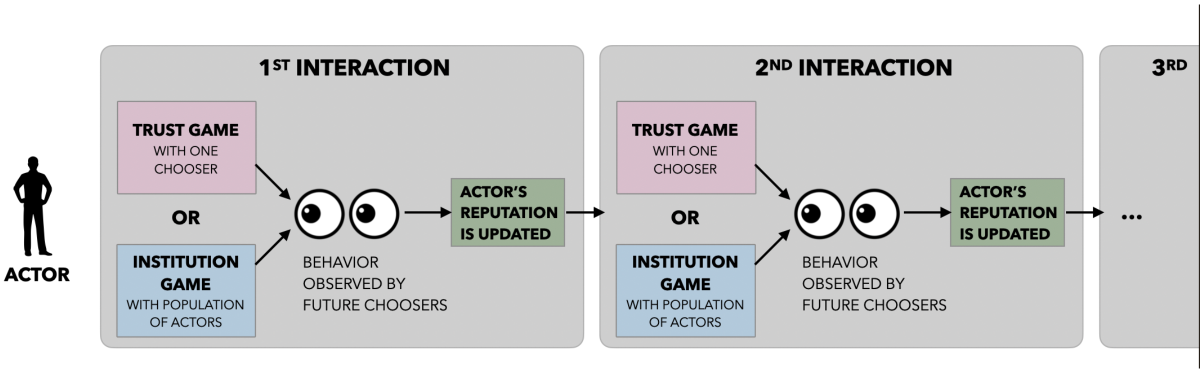
ACTOR
1ST INTERACTION
TRUST GAME
WITH ONE CHOOSER
OR
INSTITUTION GAME
WITH POPULATION OF ACTORS
BEHAVIOR OBSERVED BY FUTURE CHOOSERS
ACTOR'S REPUTATION IS UPDATED
2ND INTERACTION
TRUST GAME
WITH ONE CHOOSER
OR
INSTITUTION GAME
WITH POPULATION OF ACTORS
BEHAVIOR OBSERVED BY FUTURE CHOOSERS
ACTOR'S REPUTATION IS UPDATED
...
3RD

Figure 2: **Life of an individual actor**. Throughout her[1] life, an actor engages in infinitely many interactions. These interactions either involve a chooser and follow the logic of a trust game, or involve the population of actors and follow the logic of the institution game (both games are described below). After each interaction, the actor's behavior may be observed by future choosers. Her reputation is updated accordingly.

Every actor begins with an empty reputation. At the end of each round, an actor's behavior is observed by all choosers with a certain probability. The value of this observation probability depends both on the type of game the actor played—either a trust game or the institution game—and on the incentives produced by the institution (as will be further detailed below). With the complementary probability, the actor's behavior remains invisible to choosers. We assume that choosers only have access to information from the previous round, and not from those before. For instance, consider an actor who plays the institution game in round 1, decides to contribute, and happens to be observed by choosers. Entering round 2, her reputation is 'contributed'. If this actor then plays a trust game without being observed, choosers receive no information in that round, and the actor's reputation entering round 3 reverts to being empty.

We vary actors' ability to invest in their future reputation by varying their time preferences. Each actor is characterized by a discount factor δ ($0 < \delta < 1$), which remains hidden from other players. Payoffs throughout an actor's life are calculated following a discounted utility model, whereby the present value of a payoff unit that will be received in t rounds is δ^t. When δ is high, the actor is patient. Individual values of δ are drawn at birth, depending on the population distribution of discount factors. We consider a normal distribution of mode μ and standard deviation σ, truncated over the interval $[0, 1]$ ($0 < \mu < 1$, $0 < \sigma < 1$). When μ is high, most individual actors are patient. We refer to μ as the *patience of the population.*

Trust Game (first-order cooperation)

A trust game is a two-step process. In the first step, a chooser decides whether or not to trust an actor, depending on her reputation. Trust costs $k > 0$ to the chooser, and brings reward $r > 0$ to the actor. If trusted by the chooser, the actor decides whether or not to reciprocate, in the second step. Reciprocation costs $c_1 > 0$ to the actor, and brings benefit $b > 0$ to the chooser.

We assume $b > k$. Choosers obtain a net benefit when they partner with a trustworthy actor. When trusted by their partner, actors are observed with baseline probability p_1 by future choosers ($0 < p_1 < 1$; actors who are not trusted do not exhibit any behavior). The probability of observation in the trust game may be increased through the action of the institution (see below).

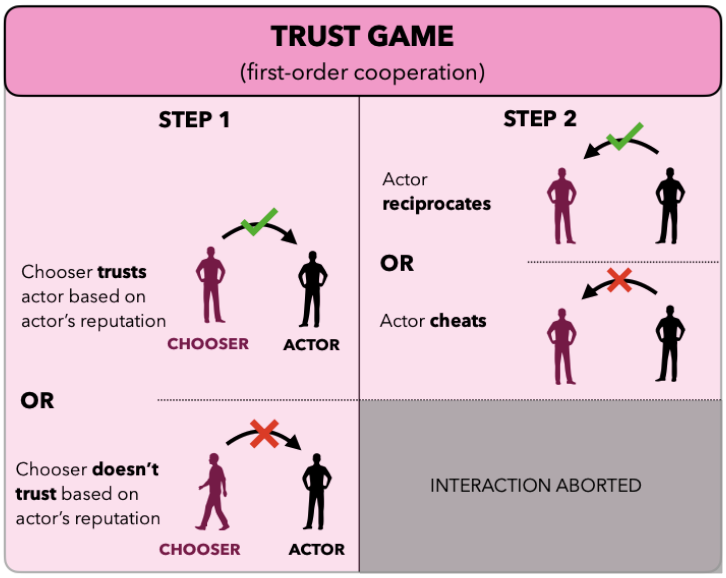
TRUST GAME
(first-order cooperation)
STEP 1
Chooser **trusts** actor based on actor's reputation
CHOOSER
ACTOR
OR
Chooser **doesn't trust** based on actor's reputation
CHOOSER
ACTOR
STEP 2
Actor **reciprocates**
OR
Actor **cheats**
INTERACTION ABORTED

Figure 3: **Trust Game**. In a trust game, one actor interacts with one chooser. The chooser acts first: on the basis of the actor's reputation, the chooser may either trust the actor or put an early end to the interaction. If trusted, the actor may, second, either reciprocate the chooser's trust, or cheat.

Institution Game (second-order cooperation)

The institution game consists in a collective action involving all the actors who draw that game. In any given round, it involves infinitely many individuals: a fraction $1 - q$ of the infinite population of actors. Each of them decides whether or not to pay c_2 in order to contribute to the institution. Their behavior is observed by choosers with fixed probability p_2 $(0 < p_2 < 1)$.

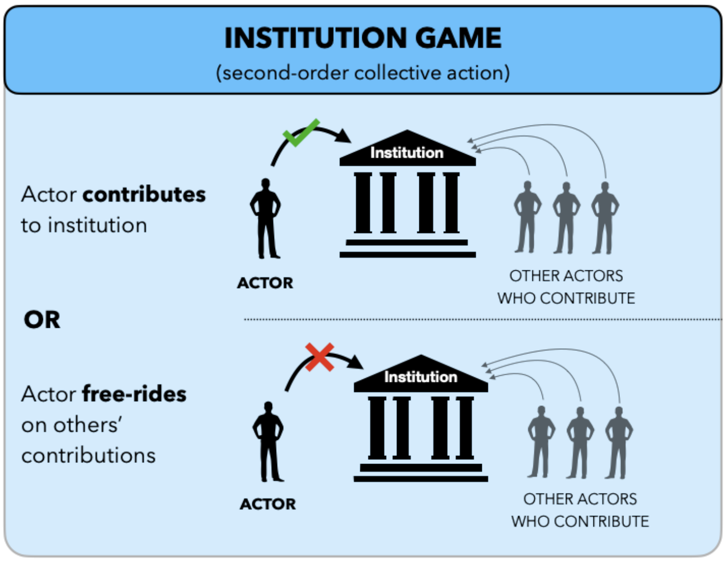
INSTITUTION GAME
(second-order collective action)
Actor **contributes** to institution
Institution
ACTOR
OTHER ACTORS WHO CONTRIBUTE
OR
Actor **free-rides** on others' contributions
Institution
ACTOR
OTHER ACTORS WHO CONTRIBUTE

Figure 4: **Institution Game**. All actors who face the institution game in a given round take part in a collective action. They can each either contribute to the institution or free-ride on others' contributions.

The institution collects actors' contributions. In a given round, we define f_2 as the fraction of potential contributors; that is, the fraction of actors who would contribute if given the opportunity to do so. Since every actor faces the institution game with probability $1 - q$, and since contributions are worth c_2, the total amount of contributions received by the institution is proportional to $(1 - q)f_2c_2$. Note that since the actor population is infinite, the total amount of contributions is infinite as well. Each individual contribution has a negligible effect (in contrast to e.g., Schoenmakers et al., 2014): in our model, the only reason to contribute to the collective action is reputational.

Mechanism of the institution

The institution transforms these contributions into incentives for first-order cooperation. One portion is allocated to rewarding cooperators, another portion is used for punishing cheaters, and the remaining portion is dedicated to monitoring. These incentives are uniformly applied to every trust game played that round; that is, the trust games played by the fraction q of actors that interact with a chooser that round. Every actor who reciprocated a partner's trust earns reward $\beta \geq 0$, every actor who cheated is punished by $\gamma \geq 0$, and the probability of observation in every trust game is increased by $\pi_1 \geq 0$. In other words, the total amount of incentives generated by the institution is proportional to $q(\beta + \gamma + c_1\pi_1)$ (again, this quantity is infinite). Note that we apply a factor of conversion c_1 to convert the probability increase π_1 into an amount in resource units.

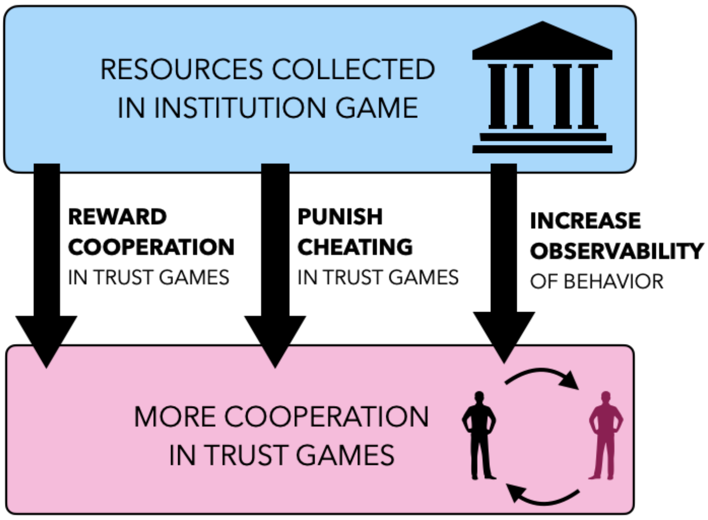

Figure 5: **Mechanism of the institution**. The institution transforms contributions made in the institution game into incentives for cooperation in trust games.

We define the *efficiency of the institution* ρ as the ratio between output and input; that is, the ratio between the incentives the institution generates and the contributions it receives. Mathematically:

$$\rho = \frac{\text{Incentives generated by the institution}}{\text{Contributions received by the institution}} = \frac{q(\beta + \gamma + c_1\pi_1)}{(1-q)f_2c_2}. \tag{1}$$

With this general model, we can consider different types of institutions by choosing different parameter values. For instance, a purely punishing institution is obtained by taking $\beta = \pi_1 = 0$. In that case, every unit of resources collected by the institution is converted into a penalty for defectors in the trust game, equal to $\gamma = \rho f_2 c_2(1-q)/q$. A purely monitoring institution is obtained by taking $\beta = \gamma = 0$; in that case, observability in the trust game increases by $\pi_1 = \rho f_2(c_2/c_1)(1-q)/q$. Finally, a (purely) rewarding institution is obtained by taking $\gamma = \pi_1 = 0$.

Taking into account the effect of the institution, we calculate the net cost of cooperation by subtracting the total payoff of cooperators from the total payoff of defectors, and obtain: $(r-\gamma)-(r-c_1+\beta) = c_1-\beta-\gamma$. The total observability of cooperation is equal to $p_1 + \pi_1$. Here, we assume that, even after accounting for the institution, first-order cooperation remains costlier and less observable than second-order cooperation, that is, that: $c_2 \leq c_1 - \beta - \gamma$ and $p_1 + \pi_1 \leq p_2$.

Results

Equilibrium analysis

We analyze our model by characterizing all possible endpoints of an evolutionary process. To do so, we use the concept of a Perfect Bayesian Equilibrium. An equilibrium is perfect when it is stable given a small likelihood of perturbing mistakes (Selten, 1983). In addition, it is Bayesian when players make decisions under incomplete information—as choosers do here, given that actors' discount factors are hidden.

Baseline: cooperation in the absence of an institution

To establish a baseline, we turn off the institution, by assuming that choosers do not observe second-order cooperation ($p_2 = 0$). In such a situation, the institution is moot. Actors never contribute to the collective action, since doing so is costly and cannot lead to reputational benefits.

We show that there then exists a unique Perfect Bayesian Equilibrium in which cooperation occurs, which we call the baseline equilibrium. In this equilibrium, reputation incentivizes first-order cooperation only. We characterize the baseline equilibrium in the Methods section at the end of this document.

The baseline equilibrium is characterized by two values: the probability that choosers trust actors whose reputation is empty, and a threshold discount factor $\hat{\delta}^b$, which separates trustworthy actors from untrustworthy actors. Sufficiently future-oriented actors ($\delta \geq \hat{\delta}^b$) always reciprocate their partners' trust, and present-oriented actors ($\delta < \hat{\delta}^b$) always cheat.

In the most favorable case, the threshold discount factor satisfies $\hat{\delta}^b = c_1/(p_1 q r)$ (in other cases, $\hat{\delta}^b > c_1/(p_1 q r)$). We refer to this minimum value as the *intrinsic difficulty of cooperation*; that is, the difficulty of cooperation in the absence of an institution. We denote it by δ^b (without a hat). Re-arranging, $\delta \geq \delta^b$ is equivalent to $(p_1 q) \times (\delta \times r) \geq c_1$. Actors cooperate when they can afford to pay c_1 in order to obtain r in the future with probability $p_1 q$—the probability of being observed in the current trust game *and* facing another chooser in the next interaction. When cooperation is costlier or less observable, its intrinsic difficulty δ^b increases, and fewer actors are able to cooperate.

Institution equilibrium

When choosers do observe second-order cooperation ($p_2 > 0$), another Perfect Bayesian Equilibrium becomes possible. We call this equilibrium the institution equilibrium. In this equilibrium, reputation incentives both first- and second-order cooperation. As with the baseline equilibrium, we characterize the institution equilibrium in the Methods section at the end of this document.

The institution equilibrium is characterized by three values: the probability that choosers trust actors whose reputation is empty, and two threshold discount factors, $\hat{\delta}_1$ and $\hat{\delta}_2$. These discount factors respectively separate trustworthy actors from cheaters, and contributors from free-riders. An actor whose discount factor is δ reciprocates her partners' trust if $\delta \geq \hat{\delta}_1$ (otherwise, she cheats on them), and contributes to the institution if $\delta \geq \hat{\delta}_2$ (otherwise, she free-rides).

In the most favorable case, the threshold discount factors satisfy $\hat{\delta}_1 = [c_1 - (\beta + \gamma)]/[(p_1 + \pi_1)q(r - \gamma)]$, and $\hat{\delta}_2 = c_2/[p_2 q(r - \gamma)]$ (in other cases, these threshold take larger values). We refer to these minimum values as the difficulty of first- and second-order cooperation in the institution equilibrium, and denote them by δ_1 and δ_2, respectively.

All types of institution lower the difficulty of cooperation: we verify that $\delta_1 \leq \delta^b$ whatever the balance operated between rewards, punishment and monitoring (i.e. the value given to the parameters β, γ and

π_1). In addition, under our assumptions, second-order cooperation is always more difficult than first-order cooperation: $\delta_2 \leq \delta_1$.

Numerical solution

To illustrate our results, we fix the institution type. We consider a monitoring-punishing institution, which allocates incentives equally between increasing the observability of cooperation and punishing defectors ($\beta = 0$, $\gamma = 1/2(\rho f_2 c_2)(1-q)/q$, $\pi_1 = 1/2(\rho f_2 c_2/c_1)(1-q)/q$). In the Supplementary Information, we consider other types of institution, and obtain similar results.

We consider three cases: (a) the baseline equilibrium obtained when choosers do not observe second-order cooperation ($p_2 = 0$), (b) the institution equilibrium obtained when the institution is inefficient ($\rho = 1/3$), and (c) the institution equilibrium obtained when the institution is efficient ($\rho = 3$). Figure 6 shows the rate of cooperation in each of these three cases, as a function of the patience of the population μ on the x-axis, and the intrinsic difficulty of cooperation δ^b on the y-axis.

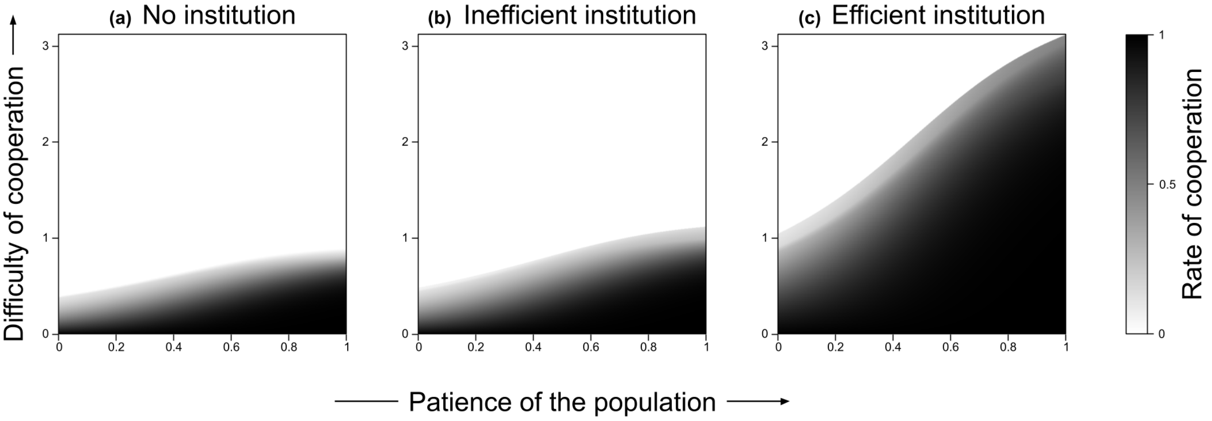

Figure 6: **Rate of cooperation**. The rate of cooperation is defined as the long-run probability of cooperation between a randomly selected actor and chooser; that is, the probability that, after many rounds of the game have already been played, a random chooser trusts a random actor based on her reputation, and the actor then reciprocates that trust. It is computed as a function of the patience of the population μ (x-axis), and the intrinsic difficulty of cooperation δ^b (y-axis), in three cases: **(a)** the baseline equilibrium obtained when $p_2 = 0$ (or no institution), **(b)** the institution equilibrium obtained when $\rho = 1/3$ (inefficient institution), and **(c)** the institution equilibrium obtained when $\rho = 3$ (efficient institution). The shade of gray indicates the rate of cooperation at a given point, with black indicating the maximum rate of 1. We consider a monitoring-punishing institution ($\beta = 0$, $\gamma = 1/2(\rho f_2 c_2)(1-q)/q$, $\pi_1 = 1/2(\rho f_2 c_2/c_1)(1-q)/q$). We fix $q = 0.5$, $r = 2$, $p_1 = 0.25$, $b = 1$, $\sigma = 0.25$, $p_2 = 3p_1 = 0.75$. We vary the intrinsic difficulty of cooperation δ^b between 0 and 3.25, and take $c_1 = k = (p_1 q r)\delta^b = \delta^b/4$, and $c_2 = c_1/3$. Without accounting for the incentives produced by the institution, actors and choosers face similar costs in trust games. Compared to first-order cooperation, second-order cooperation is initially three times cheaper, and three times as likely to be observed.

Efficient institutions extend the domain of cooperation

In the absence of an institution, hard cooperation problems cannot be solved by reputation. On panel (a) of Figure 6, null cooperation rates are obtained as soon as the intrinsic difficulty of cooperation δ^b exceeds 1.

Efficient institutions extend the domain of reputation-based cooperation, to include hard problems. On panel (c) of Figure 6, positive cooperation rates are obtained even when the intrinsic difficulty of cooperation exceeds 1—in fact, even for $\delta^b > 3$. Efficient institutions allow reputation to stabilize hard cooperation problems, by amplifying its limited effects. In contrast, an inefficient institution does not make much of a dent, as visible on panel (b) of Figure 6.

Institutions are stable when the population is patient

This beneficial effect of efficient institutions is confined to large values of μ. All other things being equal, the institution equilibrium is more likely to appear when the population is patient. Since institutions are a form of cooperation, they require that individuals pay immediate costs to invest in their long-term reputation.

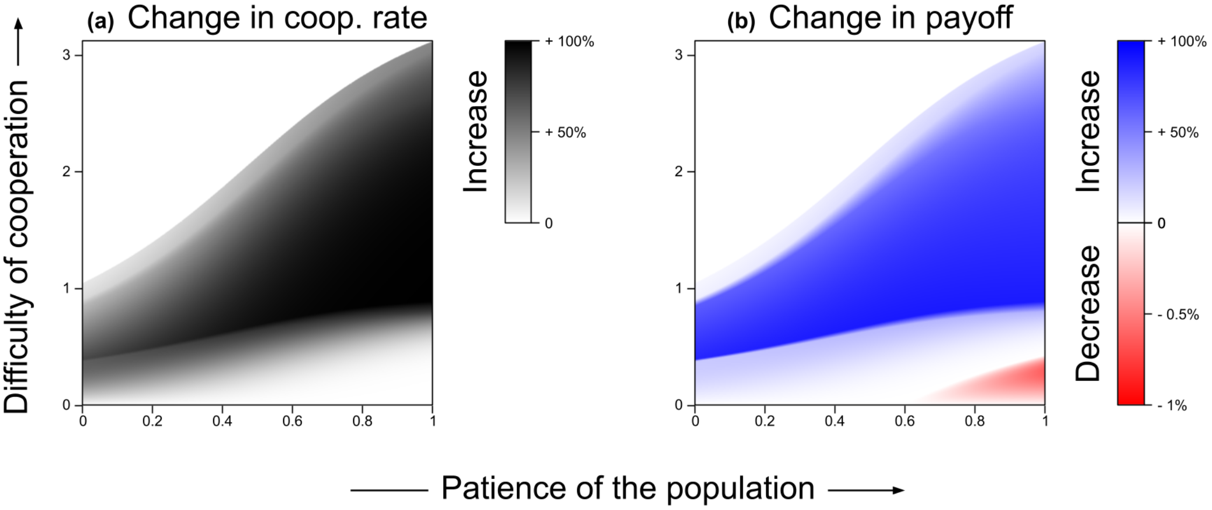

Figure 7: **Comparison between an efficient institution and no institution**. We subtract the value of **(a)** the rate of cooperation and **(b)** the expected payoff in the baseline equilibrium with $p_2 = 0$, to those same values in the institution equilibrium with $\rho = 3$. The expected payoff is defined as the normalized payoff of an individual drawn at random in both the actor and chooser populations. When an actor is drawn, we compute her expected lifetime payoff; when it's a chooser, who engages in only one interaction, we look at the expected payoff obtained once many rounds of the game have already been played. The rate of cooperation is defined as in Figure 6; we consider the same monitoring-punishing institution, and the same parameter values. The shade of gray indicates the increase in the rate of cooperation at a given point. Black: maximum increase of 100%. Shades of blue indicate an increase in the expected payoff, and shades of red indicate a decrease. Blue: maximum increase of 100%. Red: decrease of 1%. To explain these small decreases, note that with our chosen parameters, an actor who contributes to the institution pays on average $(1-q)c_2 = c_1/6$ throughout her life. In the parameter region in which the institution appears unnecessary, c_1 is small, and $c_1/6$ is very small.

Institutions are wasteful when cooperation is easy and the population is very patient

When δ^b is small in addition to μ being large, institutions become unnecessary. Large rates of cooperation can already be achieved in the non-institution equilibrium in that region. Since institutions require that individuals pay costs, they are then wasteful.

To make this more apparent, we subtract the rate of cooperation obtained in the baseline equilibrium with $p_2 = 0$ to the rate of cooperation obtained in the institution equilibrium with $\rho = 3$, and plot the difference, in panel (a) of Figure 7. We do the same operation for the expected individual payoff, and plot results in panel (b). When δ^b is small and μ is large, the institution leads only to a marginal increase in cooperation. As a result, individuals are worse off.

Discussion

> You can set up British-style courts of law, and even provide the barristers with wigs, but if the judges are venal and the barristers have no professional pride and if the public disdains them both, then the introduction of such a nice-sounding institution will fail to improve the rule of law. (McCloskey, 2016, chapter 15)

Humans rely on institutions to stabilize cooperation. Yet, as McCloskey vividly illustrates, institutions are no magic bullet. Institutions require more than just sound structures; they hinge on the people within them, whose personal interests will inevitably clash with the common good. They are, in other words, second-order cooperative interactions—cooperative interactions aimed at promoting cooperation—which emerge from the very communities they are supposed to regulate. Any satisfying model of institutions for cooperation, then, should explain both (1) how institutions generate enough incentives for cooperation, and (2) how endogenous social mechanisms within the community allow the formation of institutions in the first place.

Existing models do not provide answers to these questions (they answer other interesting questions instead). Many models assume that institutions can be enforced without individual costs—whether because enforcement is in the interest of powerful leaders (Gavrilets & Duwal Shrestha, 2021; Isakov & Rand, 2012), or because everyone commits to either reward enforcers (Sasaki et al., 2015; Wang et al., 2018) or punish non-enforcers (Sigmund et al., 2010; see also: Schoenmakers et al., 2014). In other models, enforcers pay costs that are never recouped, their selective disadvantage being compensated by selection at the level of the group (Bowles et al., 2003; Bowles & Choi, 2013; Powers & Lehmann, 2013). Enforcement is then either non-cooperative or self-sacrificial—these models do not explain how institutions can create cooperation while themselves relying on cooperation.

To fill this gap, we have introduced a mathematical model with two types of cooperative interactions, that are nested within one another. Individuals can sometimes pay to help a dyadic partner (first-order cooperation), and they can sometimes pay to contribute to an institution (second-order cooperation). The institution pools all individual contributions, and transforms them into incentives for first-order cooperation (e.g., by punishing individuals who do not help their partners).

We show that this nested architecture creates a social leverage effect. By nesting a hard dilemma within an easy one, the institution offers an indirect solution to the hard cooperation problem. For this solution to be effective, the institution must also be efficient. Efficient institutions allow reputation to indirectly stabilize cooperation in hard dilemmas, by embedding them within an easier, solvable dilemma, and still generating enough incentives based on the easy dilemma.

Our model generates distinctive predictions for how institutions form and stabilize in human societies. In the following, we detail the model's predictions, and show that they are supported by evidence from across the psychological and social sciences.

Institutions require intrinsic honesty and social capital

A distinctive feature of our model is that contributions to the institution are always costly for individuals. The more individuals are willing to pay costs to contribute to the institution, the more incentives for cooperation it can produce. This is consistent with a large body of evidence from psychology (Muthukrishna et al., 2017; Spadaro et al., 2023), economics (Beekman et al., 2014; Rose-Ackerman & Palifka, 2016), and political science (Bersch, 2019; McCloskey, 2016): well-functioning institutions require the costly cooperation of individuals, who must resist second-order free-riding (e.g., corruption) for the institution to successfully promote cooperation.

Consequently, our model predicts that institutions will be most successful in promoting cooperation in populations that are already predisposed towards cooperation. In line with this prediction, across 23 societies, institutional quality is associated with people's intrinsic honesty—people's propensity to cooperate even when they are not incentivized by institutions to do so (Gächter & Schulz, 2016). Further supporting this, in a famous study, Putnam et al. (1994) showed that the best predictor of institutional performance across Italian regions was people's propensity to engage in grassroots cooperative interactions such as sports clubs, literary guilds, or choral societies. Putnam explained this association in terms of social capital—the social networks and norms of reciprocity that emerge from a long history of grassroots cooperation. The importance of social capital for institutional functioning replicates in other geographic areas and historical periods (Andrews & Brewer, 2014; Coffé & Geys, 2005; Cusack, 1999; Gutiérrez et al., 2011; Knack, 2002; Nannicini et al., 2013; Pierce et al., 2016).

Institutional honesty depends on reputational incentives

If institutional quality depends on intrinsic honesty, what compels agents to be honest in the first place? In line with previous models (Jordan & Rand, 2017; Pal & Hilbe, 2022; Panchanathan & Boyd, 2004) and experimental evidence (Barclay, 2006; Dhaliwal et al., 2021; Jordan et al., 2016), our model suggests that reputational incentives can drive institutional cooperation. By contributing to the institution, individuals in our model enhance their reputation, and increase their chances of being rewarded by a partner's trust.

In the real world, individuals who take on an institutional role are indeed motivated by reputation and social rewards. In her famous review, Ostrom underlines how, in communities that create long-lasting institutions for common-pool resources, monitors are incentivized through reputation: "The individual who finds a rule-infractor gains status and prestige for being a good protector of the commons" (Ostrom, 1990, p. 96). Similar dynamics can be found in nonindustrial societies. Among the Enga of Papua New Guinea, for example, mediators who resolve conflicts in customary courts gain a good reputation (Wiessner, 2020). Among the Amazonian Tsimane, similarly, men who mediate more conflicts are more frequently cited as cooperation partners (Glowacki & von Rueden, 2015). More largely, across nonindustrial societies, informal leaders tend to resolve conflicts on the one hand, and enjoy high status on the other (Garfield et al., 2020).

Reputation-based institutions develop in future-oriented populations

If reputation drives institutional cooperation, what drives variation in institutional quality? Why does reputation, in many societies, manifestly fail to limit corruption? Our model suggests that this variability can be analyzed in terms of time preferences.

In the model, both first- and second-order cooperation involve a present-future trade off—cooperative individuals pay to acquire a good reputation today, and increase their chances of being trusted tomorrow. As a result, future-oriented individuals are more likely to engage in either form of cooperation, and future-oriented populations are more likely to sustain the institution.

Consequently, our model predicts that better functioning institutions should emerge in more future-oriented populations. This allows us to put two stylized facts in perspective. First, time preferences allow us to revisit the importance of social capital for institutional functioning (Putnam et al., 1994). As Putnam explains, a long history of cooperation makes social capital. It also makes the future loom large. In communities with strong social networks and norms of reciprocity, individuals can expect more from their cooperative future. With respect to their reputation, they can be characterized as patient.

Our model also explains why material circumstances matter for institutional quality. In more affluent environments, individuals' most pressing needs are already met, allowing them to explore other opportunities, like investing in their reputation or social network (Boon-Falleur et al., 2024; Mell et al., 2021). Thus, all other things being equal, individuals in more affluent environments should be more patient, and more able to trust that others will also invest in their cooperative reputation. Supporting this, experimental evidence shows that political leaders are more corrupt when their voters are poor (Denly & Gautam, n.d.), and that poorer individuals more often have to pay bribes to government officials (Justesen & Bjørnskov, 2014). At the macroscopic level, a country's level of corruption is negatively associated with its wealth (Montinola & Jackman, 2002; Ortiz-Ospina & Roser, 2016; Serra, 2006). It should be noted, however, that the relationship is bidirectional (Apergis et al., 2010; Dimant & Tosato, 2018). While economic hardship paves the way for enduring corruption (Paldam & Gundlach, 2008), corrupt institutions can also lead to economic hardship (Acemoglu & Robinson, 2013).

Social engineering and the cultural evolution of institutions

Lastly, our model contributes to understanding the cultural evolution of institutions. It illustrates how institutions can harness the social leverage effect—by nesting a hard dilemma within an easy one, the institution in our model offers an indirect solution to the hard dilemma. Put differently, it creates the possibility of stabilizing costly forms of cooperation with only weak reputational incentives. Thus, institutions appear as technologies that exploit social laws, just as material technologies exploit physical laws. They have been likely designed, and gradually refined, to build the most mutually beneficial social organizations that can be sustained by reputation alone.

For the social leverage effect to function, however, the institution must be sufficiently efficient—it needs to generate enough incentives for the hard dilemma using resources coming from the easy dilemma. The cultural evolution of institutions may have unfolded as humans discovered more efficient institutional arrangements, allowing them to exploit higher leverage, and expand the scope of cooperation. One way to maximize leverage, for example, is to assign monitoring and punishing duties to only a small group of specialized individuals, to ensure that deviations are easy to spot and identify (Lie-Panis & André, 2024). Accordingly, many real-world institutions rely on specialized monitors (Ostrom, 1990), and experimental evidence suggests that people prefer to delegate punishment decisions (Traulsen et al., 2012). Another way to maximize leverage is to rely on an increasingly nested architecture. Many institutional arrangements, particularly in large-scale societies, group individuals into lower level units, ensuring that reputation can continue to act as a strong incentive even as the number of total individuals increases (Lehmann et al., 2022).

Methods

To analyze our model, we assume that choosers can trust probabilistically an actor whose reputation is empty, and that choosers behave in a deterministic manner when faced with an actor whose reputation is non-empty—for instance, they may trust actors whose reputation is 'reciprocated' or 'contributed', and distrust actors whose reputation is 'cheated' or 'free-rode'.

Throughout this section, we denote by θ the probability that choosers trust an actor whose reputation is empty. We establish the equilibrium value of θ in the baseline and the institution equilibria below. We allow θ to vary between 0 and 1 in order to capture all situations in which cooperation is possible. In some cases, the equilibrium value of θ belongs to $(0, 1)$—considering only pure chooser strategies would lead us to miss these cases, and the plots shown in Figures 6 and 7 would have holes.

An issue with our model is that the predictive value of an empty reputation changes with time. Initially, trustworthy and untrustworthy actors are equally likely to have an empty reputation. However, as the game progresses, untrustworthy actors cheat, and are subsequently more likely to be distrusted, and to acquire an empty reputation in the next round (this is not an issue with non-empty reputations). To get around this issue, and establish the equilibrium value of θ, we look at choosers' long-run payoffs, once many rounds of the game have already been played, and each actor's reputation has reached a steady state.

We fully analyze our model in the Supplementary Information. Here, we describe the main steps that we follow in order to characterize the baseline equilibrium, as well as the institution equilibrium (in particular, we leave the characterization of choosers' posterior beliefs for the Supplementary Information).

Baseline equilibrium

To establish a baseline, we turn off information coming from the institution game, by taking $p_2 = 0$. In such a situation, we can restrict reputation to three possibilities: 'reciprocated', 'cheated', or empty. Recall that θ denotes the probability that choosers trust actors whose reputation is empty.

The baseline equilibrium is established when reputation incentivizes first-order cooperation—that is, when choosers trust actors whose reputation is 'reciprocated', and do not trust actors whose reputation is 'cheated'. Since $p_2 = 0$, reputation cannot incentivize second-order cooperation; actors never contribute to the institution in this case.

We consider a generic actor, whose discount factor is denoted by δ. Let U_δ^G be her continuation payoff when in good standing. This is the continuation payoff of the actor when her current reputation is 'reciprocated', in which case a chooser would trust her if given the opportunity to do so (i.e., if the actor draws the trust game this round). Let U_δ^B be the actor's continuation payoff when in bad standing. This is the continuation payoff of the actor when her current reputation is 'cheated'.

We show that, whatever the actor's current reputation, she will reciprocate a chooser's trust in a Perfect Bayesian Equilibrium (from here on: PBE) if and only if:

$$c_1 \leq \delta \times p_1 \times (U_\delta^G - U_\delta^B).$$

In other words, the actor is trustworthy if and only if the immediate cost of cooperation c_1 is smaller than her future (i.e. multiplied by her δ) benefit of achieving good rather than bad reputation (U_δ^G rather than U_δ^B), if observed (with probability p_1).

By calculating U_δ^G and U_δ^B in each possible case, we show that patient actors ($\delta \geq \hat{\delta}^b(\theta)$) always reciprocate their partner's trust, and that impatient actors ($\delta < \hat{\delta}^b(\theta)$) always cheat on their partners. The threshold separating trustworthy actors from untrustworthy ones is given by the equation:

$$\hat{\delta}^b(\theta) = \frac{c_1}{p_1 q (r - \theta c_1)}. \tag{B.1}$$

When θ varies between 0 and 1, $\hat{\delta}^b(\theta)$ strictly increases from $c_1/(p_1 q r)$ to $c_1/(p_1 q r(r - c_1))$. As we have defined it, the intrinsic difficulty of cooperation $\delta^b = c_1/(p_1 q r)$ is the minimum value taken by the threshold discount factor $\hat{\delta}^b(\theta)$—when actors are not trusted by default, the incentive to cooperate is the highest, and we obtain the lowest value for this threshold.

To determine the equilibrium value of θ, we look at the long-run predictive value of the empty reputation $P_\theta^\infty(\text{reciprocates} \mid \text{empty})$. This is the probability that an actor whose reputation is empty will reciprocate if trusted, once many rounds of the game have been played, and each actor's reputation has reached a steady state. We calculate this probability as a function of θ.

With our assumptions, we show that choosers trust actors who have empty reputation if and only if:

$$k \leq b \times P_\theta^\infty(\text{reciprocates} \mid \text{empty}).$$

Choosers trust actors who have empty reputation if it is beneficial to do so in expectation; that is, if the cost of trust k is smaller than b times the probability that a random actor with empty reputation can be trusted to reciprocate.

We show that in a PBE, θ must equal $\theta^{*,b}$, as given by the following algorithm:

$$\theta^{*,b} = \begin{cases} 0 & \text{if } P^{\infty}_{\theta=0}(\text{reciprocates} \mid \text{empty}) \leq \frac{k}{b}, \\ 1 & \text{if } P^{\infty}_{\theta=1}(\text{reciprocates} \mid \text{empty}) \geq \frac{k}{b}, \\ t & \text{such that } P^{\infty}_{\theta=t}(\text{reciprocates} \mid \text{empty}) = \frac{k}{b}. \end{cases} \tag{B.2}$$

The equilibrium value of θ is 0 when $P^{\infty}_{\theta}(\text{reciprocates} \mid \text{empty})$ is smaller than the relative cost of trust k/b *even* in the best case scenario for actors; that is, when $\theta = 0$ and $\hat{\delta}^b(\theta)$ is at its lowest possible value, δ^b. Conversely, the equilibrium value of θ is 1 when the probability that an actor of empty reputation reciprocates is larger than the relative cost of trust k/b *even* in the worst case scenario for actors. In all other cases, we find a unique value $0 < \theta < 1$. We allow choosers to mix given empty reputation in order to include these cases.

When reputation incentivizes first-order cooperation and $p_2 = 0$, θ is then determined by (B.2), and actor strategy is determined by θ and the threshold given by (B.1). We then obtain a PBE as long as choosers benefit from trusting actors in good standing, and from distrusting actors in bad standing. We show that a necessary and sufficient condition is that the equilibrium value of the difficulty of cooperation $\hat{\delta}^b_1(\theta^{*,b})$ be smaller than 1, in order to guarantee a positive fraction of trustworthy actors, whose discount factor exceeds the difficulty of cooperation.

In other words, we show that the baseline equilibrium is a PBE if and only if:

$$\hat{\delta}^b_1(\theta^{*,b}) < 1. \tag{B.3}$$

Institution equilibrium

The institution equilibrium is established when reputation incenvitizes both first- and second-order cooperation—that is, when choosers trust actors whose reputation is 'reciprocated' as well as actors whose reputation is 'contributed', and distrust actors whose reputation is 'cheated' or 'free-rode'. Again, θ denotes the probability that choosers trust actors whose reputation is empty.

We once again consider a generic actor, whose discount factor is denoted by δ. Let U^G_δ be her continuation payoff when in good standing. This time, U^G_δ is the continuation payoff of the actor when her current reputation is 'reciprocated' or 'contributed'—in both of these cases, a chooser would trust her if given the opportunity. Let U^B_δ be the actor's continuation payoff when in bad standing. This time, U^B_δ is the continuation payoff of the actor when her current reputation is 'cheated' or 'free-rode'.

We show that, whatever the actor's current reputation, she will reciprocate a chooser's trust in a PBE if and only if:

$$c_1 - (\beta + \gamma) \leq \delta \times (p_1 + \pi_1) \times (U^G_\delta - U^B_\delta).$$

As above, the actor is trustworthy when it is worth paying the immediate cost of cooperation for her future reputation. This time however, the cost of cooperation is $c_1 - (\beta + \gamma)$, and the probability of being observed is $p_1 + \pi_1$, as we must account for institutional rewards, punishing, and/or monitoring.

Similarly, we show that the actor contributes to the institution if and only if:

$$c_2 \leq \delta \times p_2 \times (U^G_\delta - U^B_\delta).$$

By again calculating U^G_δ and U^B_δ in each possible case, we show the existence of two thresholds $\hat{\delta}_1(\theta)$ and $\hat{\delta}_2(\theta)$, separating between trustworthy and untrustworthy actors on the one hand, and contributors and free-riders on the other.

The threshold separating trustworthy actors from untrustworthy ones is given by:

$$\hat{\delta}_1(\theta) = \frac{c_1 - \beta - \gamma}{(p_1 + \pi_1)q[r - \gamma - \theta(c_1 - \gamma - \beta)]}. \tag{I.1}$$

The threshold separating contributors from free-riders is given by:

$$\hat{\delta}_2(\theta) = \frac{c_2}{q[p_2(r - \gamma) - (p_1 + \pi_1)\theta c_2]}. \tag{I.2}$$

We verify that $\hat{\delta}_2(\theta) \leq \hat{\delta}_1(\theta)$. Since we have assumed that second-order cooperation remains less costly ($c_2 \leq c_1 - (\beta + \gamma)$) and more observable ($p_2 \geq p_1 + \pi_1$) than first-order cooperation, even after accounting for the effects of the institution, second-order cooperation requires passing a lower bar (a lower threshold for one's discount factor). In equilibrium, some actors will contribute but not cooperate, but every actor that cooperates will contribute to the institution.

When θ varies between 0 and 1, $\hat{\delta}_1(\theta)$ strictly increases from δ_1 and $\hat{\delta}_2(\theta)$ strictly increases from δ_2—as we have defined them, δ_1 and δ_2 provide lower bounds on the relevant threshold discount factors.

As before, we look at the long-run predictive value of the empty reputation $P^{\infty}_{\theta}(\text{reciprocates} \mid \text{empty})$ to determine the equilibrium value of θ. We calculate this probability as a function of θ.

We show that in a PBE, θ must equal θ^*, as given by the same algorithm as before:

$$\theta^* = \begin{cases} 0 & \text{if } P^{\infty}_{\theta=0}(\text{reciprocates} \mid \text{empty}) \leq \frac{k}{b}, \\ 1 & \text{if } P^{\infty}_{\theta=1}(\text{reciprocates} \mid \text{empty}) \geq \frac{k}{b}, \\ t & \text{such that } P^{\infty}_{\theta=t}(\text{reciprocates} \mid \text{empty}) = \frac{k}{b}. \end{cases} \tag{I.3}$$

We then obtain a PBE, characterized by conditions (I.1-I.3), as long as choosers benefit from trusting actors in good standing, and from distrusting actors in bad standing. We show that two conditions are, together, necessary and sufficient: (1) the equilibrium value of the difficulty of cooperation $\hat{\delta}_1(\theta^*)$ must be smaller than 1, to once again guarantee a positive fraction of trustworthy actors, and (2) second-order cooperation must be a sufficiently good predictor of first-order cooperation—otherwise, trusting actors who contributed in the past is costly in expectation.

In other words, we show that the institution equilibrium is a PBE if and only if:

$$\hat{\delta}_1(\theta^*) < 1, \tag{I.4}$$

$$P(\delta \geq \hat{\delta}_1(\theta^*) \mid \delta \geq \hat{\delta}_2(\theta^*)) \geq \frac{k}{b}. \tag{I.5}$$

Acknowledgements

We would like to thank Mélusine Boon-Falleur, Helena Miton, Jorge Peña, and Manvir Singh for their feedback on earlier versions of this manuscript. This research was funded by Agence Nationale pour la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02).

Notes

[1] We use the pronouns she/her to refer to actors throughout this document.

References

Acemoglu, D., & Robinson, J. A. (2013). *Why nations fail: The origins of power, prosperity, and poverty*. Profile Books.

Andrews, R., & Brewer, G. A. (2014). Social Capital and Public Service Performance: Does Managerial Strategy Matter? *Public Performance & Management Review*, *38*(2), 187–213. Retrieved May 22, 2023, from https://www.jstor.org/stable/24735250

Apergis, N., Dincer, O. C., & Payne, J. E. (2010). The relationship between corruption and income inequality in U.S. states: Evidence from a panel cointegration and error correction model. *Public Choice*, *145*(1), 125–135. https://doi.org/10.1007/s11127-009-9557-1

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2020). Reciprocity creates a stake in one's partner, or why you should cooperate even when anonymous. *Proceedings of the Royal Society B*. https://doi.org/10.1098/rspb.2020.0819

Barclay, P., Bliege Bird, R., Roberts, G., & Számadó, S. (2021). Cooperating to show that you care: Costly helping as an honest signal of fitness interdependence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1838), 20200292. https://doi.org/10.1098/rstb.2020.0292

Beekman, G., Bulte, E., & Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural Liberia. *Journal of Public Economics*, *115*, 37–47. https://doi.org/10.1016/j.jpubeco.2014.04.004

Bersch, K. (2019, January 31). *When Democracies Deliver: Governance Reform in Latin America* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108559638

Boon-Falleur, M., Baumard, N., & André, J. B. (2024). The effect of income and wealth on behavioral strategies, personality traits, and preferences. *Perspectives on Psychological Science*. https://doi.org/17456916231201512

Bowles, S., Choi, J.-K., & Hopfensitz, A. (2003). The co-evolution of individual behaviors and social institutions. *Journal of theoretical biology*. https://doi.org/10.1016/S0022-5193(03)00060-2

Bowles, S., & Choi, J.-K. (2013). Coevolution of farming and private property during the early holocene. *Proceedings of the National Academy of Sciences*, *110*(22), 8830–8835. https://doi.org/10.1073/pnas.1212149110

Coffé, H., & Geys, B. (2005). Institutional Performance and Social Capital: An Application to the Local Government Level. *Journal of Urban Affairs*, *27*(5), 485–501. https://doi.org/10.1111/j.0735-2166.2005.00249.x

Currie, T. E., Campenni, M., Flitton, A., Njagi, T., Ontiri, E., Perret, C., & Walker, L. (2021). The cultural evolution and ecology of institutions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1828), 20200047. https://doi.org/10.1098/rstb.2020.0047

Cusack, T. R. (1999). Social capital, institutional structures, and democratic performance: A comparative study of german local governments. *European Journal of Political Research*, *35*(1), 1–34. https://doi.org/10.1111/1475-6765.00440

Denly, M., & Gautam, A. (n.d.). Poverty, Party Alignment, and Reducing Corruption through Modernization: Evidence from Guatemala, 216.

Dhaliwal, N., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes*, *164*, 27–51. https://doi.org/10.1016/j.obhdp.2021.01.003

Dimant, E., & Tosato, G. (2018). Causes and Effects of Corruption: What Has Past Decade's Empirical Research Taught Us? A Survey. *Journal of Economic Surveys*, *32*(2), 335–356. https://doi.org/10.1111/joes.12198

Dixit, A. (2018). Anti-corruption Institutions: Some History and Theory. In K. Basu & T. Cordella (Eds.), *Institutions, Governance and the Control of Corruption* (pp. 15–49). Springer International Publishing. https://doi.org/10.1007/978-3-319-65684-7_2

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French Revolution*. Profile books.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496–499. https://doi.org/10.1038/nature17160

Garfield, Z. H., Syme, K. L., & Hagen, E. H. (2020). Universal and variable leadership dimensions across human societies. *Evolution and Human Behavior*, *41*(5), 397–414. https://doi.org/10.1016/j.evolhumbehav.2020.07.012

Gavrilets, S., & Currie, T. E. (2022). Mathematical models of the evolution of institutions. https://doi.org/10.31235/osf.io/kuxvd

Gavrilets, S., & Duwal Shrestha, M. (2021). Evolving institutions for collective action by selective imitation and self-interested design. *Evolution and Human Behavior*, *42*(1), 1–11. https://doi.org/10.1016/j.evolhumbehav.2020.05.007

Giardini, F., & Vilone, D. (2016). Evolution of gossip-based indirect reciprocity on a bipartite network. *Scientific Reports*, *6*(1), 37931. https://doi.org/10.1038/srep37931

Glowacki, L., & von Rueden, C. (2015). Leadership solves collective action problems in small-scale societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1683), 20150010. https://doi.org/10.1098/rstb.2015.0010

Greif, A., Milgrom, P., & Weingast, B. R. (1994). Coordination, Commitment, and Enforcement: The Case of the Merchant Guild. *Journal of Political Economy*, *102*(4), 745–776.

Gutiérrez, N. L., Hilborn, R., & Defeo, O. (2011). Leadership, social capital and incentives promote successful fisheries. *Nature*, *470*(7334), 386–389. https://doi.org/10.1038/nature09689

Hadfield, G. K., & Weingast, B. R. (2013). Law without the State: Legal Attributes and the Coordination of Decentralized Collective Punishment. *Journal of Law and Courts*, *1*(1), 3–34. https://doi.org/10.1086/668604

Hamilton, W. D. (1963). The Evolution of Altruistic Behavior. *The American Naturalist*, *97*(896), 354–356. https://doi.org/10.1086/497114

Henrich, J., & Muthukrishna, M. (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Isakov, A., & Rand, D. G. (2012). The Evolution of Coercive Institutional Punishment. *Dynamic Games and Applications*, *2*(1), 97–109. https://doi.org/10.1007/s13235-011-0020-9

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., & Rand, D. G. (2017). Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, 189–202. https://doi.org/10.1016/j.jtbi.2017.04.004

Justesen, M. K., & Bjørnskov, C. (2014). Exploiting the Poor: Bureaucratic Corruption and Poverty in Africa. *World Development*, *58*, 106–115. https://doi.org/10.1016/j.worlddev.2014.01.002

Knack, S. (2002). Social Capital and the Quality of Government: Evidence from the States. *American Journal of Political Science*, *46*(4), 772–785. https://doi.org/10.2307/3088433

Lehmann, L., Powers, S. T., & van Schaik, C. P. (2022). Four levers of reciprocity across human societies: Concepts, analysis and predictions. *Evolutionary Human Sciences*, *4*, e11. https://doi.org/10.1017/ehs.2022.7

Lienard, P. (2016). Age Grouping and Social Complexity. *Current Anthropology*, *57*(S13), S105–S117. https://doi.org/10.1086/685685

Lie-Panis, J., & André, J.-B. (2024). Peace is a form of cooperation, and so are the cultural technologies which make peace possible. *Behavioral and Brain Sciences*, *47*, e16. https://doi.org/10.1017/S0140525X23002601

Lie-Panis, J., & André, J.-B. (2022). Cooperation as a signal of time preferences. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1973), 20212266. https://doi.org/10.1098/rspb.2021.2266

McCloskey, D. N. (2016). *Bourgeois equality: How ideas, not capital or institutions, enriched the world*. The University of Chicago Press.

McKean, M. A. (1986). Management of traditional common lands (iriaichi) in japan. *Proceedings of the Conference on Common Property Resource Management*, 533–589.

Mell, H., Baumard, N., & André, J.-B. (2021). Time is money. Waiting costs explain why selection favors steeper time discounting in deprived environments. *Evolution and Human Behavior*, *42*(4), 379–387. https://doi.org/10.1016/j.evolhumbehav.2021.02.003

Milgrom, P., North, D., & Weingast, B. (1990). The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics*, *2*, 1–23. https://doi.org/10.1111/j.1468-0343.1990.tb00020.x

Montinola, G. R., & Jackman, R. W. (2002). Sources of Corruption: A Cross-Country Study. *British Journal of Political Science*, *32*(1), 147–170. https://doi.org/10.1017/S0007123402000066

Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, *1*(7), 1–5. https://doi.org/10.1038/s41562-017-0138

Nannicini, T., Stella, A., Tabellini, G., & Troiano, U. (2013). Social Capital and Political Accountability. *American Economic Journal: Economic Policy*, *5*(2), 222–250. https://doi.org/10.1257/pol.5.2.222

North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685), 573–577. https://doi.org/10.1038/31225

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*(7092), 502–505. https://doi.org/10.1038/nature04605

Ortiz-Ospina, E., & Roser, M. (2016). Corruption. *Our World in Data*. Retrieved October 24, 2022, from https://ourworldindata.org/corruption

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Pal, S., & Hilbe, C. (2022). Reputation effects drive the joint evolution of cooperation and social rewarding. *Nature Communications*, *13*(1), 5928. https://doi.org/10.1038/s41467-022-33551-y

Paldam, M., & Gundlach, E. (2008). Two Views on Institutions and Development: The Grand Transition vs the Primacy of Institutions. *Kyklos*, *61*(1), 65–100. https://doi.org/10.1111/j.1467-6435.2008.00393.x

Panchanathan, K., & Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, *224*(1), 115–126. https://doi.org/10.1016/S0022-5193(03)00154-1

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502. https://doi.org/10.1038/nature02978

Persson, A., Rothstein, B., & Teorell, J. (2013). Why Anticorruption Reforms Fail—Systemic Corruption as a Collective Action Problem. *Governance*, *26*(3), 449–471. https://doi.org/10.1111/j.1468-0491.2012.01604.x

_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0491.2012.01604.x

Pierce, J., Lovrich, N., & Budd, W. (2016). Social capital, institutional performance, and sustainability in Italy's regions: Still evidence of enduring historical effects? *The Social Science Journal*, *53*. https://doi.org/10.1016/j.soscij.2016.06.001

Powers, S. T., & Lehmann, L. (2013). The co-evolution of social institutions, demography, and large-scale human cooperation (M. V. Baalen, Ed.). *Ecology Letters*, *16*(11), 1356–1364. https://doi.org/10.1111/ele.12178

Powers, S. T., van Schaik, C. P., & Lehmann, L. (2016). How institutions shaped the last major evolutionary transition to large-scale human societies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687), 20150098. https://doi.org/10.1098/rstb.2015.0098

Powers, S. T., van Schaik, C. P., & Lehmann, L. (2021). Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve? *Evolutionary Anthropology: Issues, News, and Reviews*, *30*(4), 280–293. https://doi.org/10.1002/evan.21909

Putnam, R. D., Leonardi, R., & Nanetti, R. (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton Univ. Press.

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, *492*, 110204. https://doi.org/10.1016/j.jtbi.2020.110204

Rose-Ackerman, S., & Palifka, B. J. (2016). *Corruption and Government: Causes, Consequences, and Reform*. Cambridge University Press.

Sasaki, T., Uchida, S., & Chen, X. (2015). Voluntary rewards mediate the evolution of pool punishment for maintaining public goods in large populations. *Scientific Reports*, *5*(1), 8917. https://doi.org/10.1038/srep08917

Schoenmakers, S., Hilbe, C., Blasius, B., & Traulsen, A. (2014). Sanctions as honest signals – The evolution of pool punishment by public sanctioning institutions. *Journal of Theoretical Biology*, *356*, 36–46. https://doi.org/10.1016/j.jtbi.2014.04.019

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, *366*(6466), eaau5141. https://doi.org/10.1126/science.aau5141

Selten, R. (1983). Evolutionary stability in extensive two-person games. *Mathematical Social Sciences*, *5*(3), 269–363. https://doi.org/10.1016/0165-4896(83)90012-4

Serra, D. (2006). Empirical determinants of corruption: A sensitivity analysis. *Public Choice*, *126*(1), 225–256. https://doi.org/10.1007/s11127-006-0286-4

Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, *466*(7308), 861–863. https://doi.org/10.1038/nature09203

Spadaro, G., Molho, C., Van Prooijen, J.-W., Romano, A., Mosso, C. O., & Van Lange, P. A. M. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nature Human Behaviour*, *7*(1), 46–54. https://doi.org/10.1038/s41562-022-01457-w

Sznycer, D., & Patrick, C. (2020). The origins of criminal law. *Nature Human Behaviour*, *4*(5), 506–516. https://doi.org/10.1038/s41562-020-0827-8

Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1743), 3716–3721. https://doi.org/10.1098/rspb.2012.0937

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Wang, Q., He, N., & Chen, X. (2018). Replicator dynamics for public goods game with resource allocation in large populations. *Applied Mathematics and Computation*, *328*, 162–170. https://doi.org/10.1016/j.amc.2018.01.045

Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the Enga of Papua New Guinea. *Proceedings of the National Academy of Sciences*, *117*(51), 32320–32328. https://doi.org/10.1073/pnas.2014759117

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*, 110–116. https://doi.org/10.1037/0022-3514.51.1.110