AUTHORS: Rahul Bhui[1], Maciej Chudek[2], Joseph Henrich[3,4,*]

TITLE: How Exploitation Launched Human Cooperation

AFFILIATION:

[1]Department of Psychology and Center for Brain Science, Harvard University

[2] Unaffiliated

[3] Department of Human Evolutionary Biology, Harvard University

[4] Canadian Institute for Advanced Research

* Corresponding Author: Department of Human Evolutionary Biology, Divinity Avenue, Cambridge, MA; email: henrich@fas.harvard.edu

ABSTRACT

The evolution of large-scale human cooperation from the cognitive fundamentals found in other primates remains an evolutionary puzzle. Most theoretical work focuses on positive reciprocity (helping) or coordinated punishment by assuming well-defined social roles (e.g. donor) or institutions (e.g., punishment pools), sophisticated cognitive abilities for navigating these, and sufficiently harmonious communities to allow for mutual aid. Here we explore the evolutionary and developmental origins of these assumed preconditions by showing how Negative Indirect Reciprocity (NIR)—tolerated exploitation of those with bad reputations—can suppress misbehavior to foster harmonious communities, favor the cognitive abilities often assumed by other models, and support costly adherence to norms (including contributing to public goods). With minimal cognitive prerequisites, NIR sustains cooperation when exploitation is inefficient (victims suffer greatly; exploiters gain little), which is more plausible earlier in human evolutionary history than the efficient helping found in many models. Moreover, as auxiliary opportunities to improve one's reputation become more frequent, the communal benefits provided in equilibrium grow, although NIR becomes harder to maintain. This suggests that NIR sets the stage for the evolution of more complex strategies to support positive cooperation.

KEYWORDS: Evolution, Negative indirect reciprocity, Cooperation, Reputation, Social norms

SIGNIFICANCE STATEMENT

The evolutionary origins of human cooperation our species' prosociality remain an evolutionary puzzle. Theoretical models exploring the dynamics which shaped our ancestors' interactions stimulate empirical investigations by anthropologists, primatologists, psychologists, archaeologists and others, whose results in turn refine and direct theoretical inquiry. Common experience has focused this scholarly synergy on positive cooperation (cooperating by helping) and largely neglected the distinct and important challenge of negative cooperation (cooperating by not exploiting). Our contribution puts negative cooperation back in the spotlight. We outline what makes negative cooperation, especially negative indirect reciprocity, different and potentially more potent than positive cooperation, and present a simple model of how it emerges, shapes interactions, and can form a dynamic foundation that catalyzes more sophisticated forms of cooperation.

INTRODUCTION

On a small island in the northwestern corner of the Fijian archipelago, subsistence-oriented farmers and fishers cooperate intensely in many domains of life. The villagers on Yasawa Island reliably show up to work on communal projects such as cleaning up the village, constructing communal buildings, and preparing for public feasts. Such collective activities happen at least weekly, and Yasawans work hard with good cheer and laughter. Yasawan geniality is evident even in experimental paradigms used to measure prosociality; they make equitable offers in dictator, ultimatum, and third-party punishment games, approaching those of Western populations (Henrich and Henrich 2014); yet, unlike Westerners, they generally won't pay to punish or sanction in these experiments. This way of life stands in stark contrast to many other small-scale populations—like the Matsigenka of Peru or the Mapuche of Chile—where folks are wary of communal work and collective action in large groups, making it virtually impossible to assemble labor forces to perform tasks similar to those routinely performed in Yasawan villages; not surprisingly, people in these populations are far less equitable to their fellow villagers in experiments compared to Yasawans (Henrich et al. 2001, 2005; Henrich and Smith 2004).

How is Yasawan cooperation maintained? Some classic theories about the evolution of cooperation imply that prosociality can be driven by direct reciprocity or costly punishment, that is, by overt retaliation in the same kind of economic interaction or by individually costly actions taken by observers. But while this behavior is systematically observed in experiments with Western participants (Ensminger and Henrich 2014), it is far less common or non-existent among the Yasawans (Henrich and Henrich 2014). Instead, systemic interviews and vignette studies reveal that in rare instances where an individual consistently does not contribute to village affairs, their reputation is damaged by gossip and they are sanctioned by anonymous punishment such as the theft of their crops, often carried out by those with preexisting grudges. Such acts, which provide benefits to the punishers, would normally be investigated by the community—but when the targeted individual has a bad reputation, the community looks the other way. In this world, it's only bad to do bad things to good (or well-reputed) people. In this paper, we formally explore how this mechanism of *negative indirect reciprocity* can simultaneously control harmful exploitative behaviors and sustain norm adherence (including socially beneficial cooperation) in other domains.

From a wider perspective, human cooperation is peculiar in several ways. Unlike other species, humans not only cooperate more broadly and intensively than other species, but the extent of this cooperation varies dramatically across diverse domains (e.g., in fishing, house building, and war) as well as among societies, including those inhabiting identical environments. Moreover, the scale of human cooperation has expanded dramatically over the last twelve millennia in patterns and at speeds that cannot be accounted for by genetic evolution (Henrich and Henrich 2007; Chudek and Henrich 2011). Consequently, a proper evolutionary approach to human cooperation must seat our species within the natural world, subject to both natural selection and phylogenetic constraints, while at the same time proposing evolutionary hypotheses that account for the unique evolutionary, developmental, psychological, and historical features of human cooperation.

Aiming to address the puzzle of human ultra-sociality, many formal evolutionary models of cooperation make assumptions about the cognitive abilities of potential cooperators. Some, such as kinship (Hamilton 1964) and direct reciprocity (Trivers 1971; Axelrod and Hamilton 1981), presuppose few cognitive prerequisites but only explain cooperation under special conditions—among kin, or in very small groups (Boyd and Richerson 1988a,b). Other models tackle the challenge of explaining distinctly human forms of cooperation, but do so by presupposing a cognitively sophisticated, highly cultural species. For instance, important models assume that people can establish sophisticated institutions (Sigmund et al. 2010), interpret one another's signals of cooperative intent (Boyd et al. 2010), or coordinate their community-wide definitions of deserving "recipients" and responsible "donors" (Leimar and Hammerstein 2001; Panchanathan and Boyd 2004; Boyd et al. 2010). By emphasizing the evolution of positive cooperation (reciprocal helping) these models also presuppose relatively harmonious communities where the benefits of mutual aid can accumulate and shape long term fitness without being rapidly undermined by opportunistic exploitation, such as theft or rape.

Though they demonstrate how human cooperation may have rapidly escalated, these models gloss over the critical earliest stages of the emergence of human cooperation, since harmonious communities which coordinate complex cognitive representations (e.g., who is a "donor"), establish institutions and dynamically signal their behavioral intentions in novel domains are themselves impressive cooperative accomplishments. Explaining the origins of such communities while assuming only minimal cognitive prerequisites (consistent with what is known about primate cognition) remains an outstanding challenge. To address this challenge, we detail an evolutionary mechanism that rapidly coordinates expectations and behavior in arbitrary domains (e.g., hunting, sharing information, trade) and yet can arise without preexisting capacities for coordinating complex institutions or socially prescribed roles.

Of these approaches to human cooperation, one important class of models is based on "Indirect Reciprocity" (IR; e.g., Nowak and Sigmund 1998; Leimar and Hammerstein 2001; Panchanathan et al. 2003; Nowak and Sigmund 2005). *Prima facie* IR models assume only that (a) individuals have opinions of one another and that these opinions (b) influence how individuals treat each other and (c) can be culturally transmitted. Since many primates form coalitions with non-kin (Silk 2002; Watts 2002; Langergraber et al. 2007; Perry and Manson 2008; Higham and

Maestripieri 2010), the first two assumptions are plausible socio-cognitive pre-adaptations in our Pliocene ancestors. The third assumption is also plausible if our early cognitive adaptations for cultural learning (e.g., for acquiring food preferences) spilled over into other domains, producing individuals who sometimes culturally acquired their opinions of one another (Henrich and Gil-White 2001). The cultural transmission of social opinions can transform pairwise coalitional affiliations into community-wide “reputations”. Once reputations had fitness consequences, they could begin shaping behavior in any reputation-relevant domain (Panchanathan and Boyd 2004), stabilizing conformity to arbitrary community norms and providing the substrate for the more complex cooperation-sustaining mechanisms that presuppose coordinated communities (Chudek and Henrich 2011; Henrich 2016, Chapter 11). Crucially, such culture-driven forms of genetic evolution do not emerge in most species due to the barriers to evolving cumulative cultural evolution (Boyd and Richerson 1996; Henrich 2016, Chapter 16).

However, existing IR models make substantially stronger assumptions about the cognitive sophistication and social coordination capacities of our ancestors. Framed in the context of reciprocal helping, these models assume that sometimes someone has an opportunity to help but does nothing, and that their reputation worsens as a consequence of their *inaction*. This seemingly innocuous assumption implies that their peers cognitively represent, and coordinate their representations of both the abstract opportunity to act, and the significance of inaction. This is a sophisticated cognitive feat. Noting this issue, Leimar and Hammerstein (2001) write that IR models assume “a reasonably fair and efficient mechanism of assigning donors and recipients […] a well-organized society, with a fair amount of agreement between its members as to which circumstances define [these] roles”. Most IR models implicitly mirror these assumptions (Nowak and Sigmund 1998; Panchanathan et al. 2003).

Here we ask whether IR is plausible *without* assuming coordinated reactions to "inaction". We develop a general model of IR, which incorporates the possibility that reputations are regularly buffeted by random external influences, but inaction *never* changes reputations. Our results show that IR is nevertheless plausible under these circumstances and can support adherence to community norms in other domains. We demonstrate how early proto-reputations (by-products of cultural learning and coalitional psychology) can escalate in importance until they form the substrate of more complex forms of cooperation.

Since we are interested in modelling the earliest forms of distinctly human cooperation, we focus on "Negative Indirect Reciprocity" (hereafter, NIR), which has rarely been the focus of study. "Negative reciprocity" broadly denotes retaliation in response to another's uncooperative behavior (e.g., Fehr and Gächter 2000). NIR extends this retaliation to depend on the other person's reputation, and hence indirectly on their behavior. Such punitive interactions take place in negative cooperative dilemmas, where "defecting" means gainfully exploiting someone and "cooperating" means seeing such an opportunity to exploit someone but passing it up ("doing nothing")—though note that reputations (and hence retribution) are allowed to be contingent on behavior in other, positive dilemmas in addition to the focal negative one. Typical models treat negative dilemmas as merely the symmetrical flip-side of standard (positive) cooperative dilemmas due to their equivalent payoff matrices. However, there are both theoretical and empirical reasons to think that negative dilemmas are psychologically distinct scenarios that were particularly potent early in the evolution of human cooperation:

1. **Substantial positive cooperation presupposes harmonious communities**: Before more complex forms of mutual aid, defense, and helping can emerge, the ubiquitous opportunities to exploit each other (particularly the old, weak, and injured) must be

brought under control. Otherwise, exploitation and cycles of revenge will undermine positive cooperation. A degree of harmony must come first.

2. **Positive cooperation creates or exacerbates negative dilemmas (but not the reverse):** Positive cooperation will often create an abundance of exploitable resources, both tangible (e.g., food caches) and intangible (e.g., trust). If cooperation has not first been stabilized in negative dilemmas, escalating opportunities for exploitation can quickly sap these benefits, sabotaging the viability of positive cooperation. For example, our band might cooperate to create a community store of food for the winter. But, then, over several wintery months, nightly thieves might slowly pilfer it away.
3. **Escalating returns**: Prior to the emergence of complex institutions like debt, money or police, if a well-reputed individual is helped multiple times (i.e., by multiple peers) they are likely to experience diminishing marginal returns. A little food when you are starving provides a huge benefit, whereas a lot of food when you are full provides only incremental benefits. On the other hand, repeated exploitation (e.g., stealing someone's resources) can put victims in ever more dire situations with escalating fitness consequences (e.g., the repeated theft of food from the hungry and weak). This suggests that in the IR context, where many community members respond to a focal well- or ill-reputed individual, negative dilemmas likely generate steeper selection gradients. This was likely most relevant earlier in our evolutionary history, before widespread food-sharing norms emerged (likely an early form of positive cooperation).
4. **No chicken and egg problem:** In a positive cooperative dilemma, when inaction is unobservable or there is a lack of sufficient agreement about what constitutes

"inaction", an individual's reputation can endogenously rise (by helping) but it cannot effectively fall through inaction. Though an individual's reputation might fall accidentally, selection will not favor individuals who take deliberate costly actions to worsen their reputation. Clearly, reputation has little value until it can fall as well as rise; but without complex culturally-evolved institutions or cognitive abilities to establish agreement about what constitutes "inaction", it is not clear how positive indirect reciprocity gets off the ground—there is a chicken and egg situation. Negative dilemmas lack the chicken and egg quality because "defections" (e.g., stealing food from the injured) are salient and observable actions.

5. **Relevance to culture:** The cooperative dilemma of cultural learning (whether to trust information shared by others, and whether to share information honestly) is a major hurdle to more sophisticated institutional forms of cooperation and is a fundamentally negative dilemma. Individuals must pass up opportunities to gainfully deceive their credulous conspecifics. This dilemma is all the worse for more culture-dependent species. Negative dilemmas related to sharing cultural information must be solved to unleash powerful forms of cumulative cultural evolution (Henrich 2016).
6. **Preadaptations are more plausible**: The cognitive capacities for navigating negative dilemmas (noticing and responding to opportunities to gain benefits by exploiting others) yield individual advantages and so were likely better honed by selection earlier than those for navigating positive dilemmas (noticing opportunities to pay costs for others' welfare).
7. **Supported by psychological evidence:** Much contemporary psychological evidence points to the relevance of negative dilemmas. People today are more sensitive to harm than helping (negativity bias), and to harm by commission than by omission. Harmful

or aversive actions, events, or stimuli have more and stronger effects on contemporary humans than their positive or beneficial counterparts (for reviews, see Cacioppo and Berntson 1994; Baumeister et al. 2001; Rozin and Royzman 2001; for antecedents in three-month-olds, see Hamlin et al. 2010). Of particular relevance, negative information (i.e., about others' harmful acts) has a far more potent effect on reputations than positive information (Fiske 1980; Skowronski and Carlston 1987; Rozin and Royzman 2001), and people judge that others caused negative outcomes more intentionally than positive ones (Knobe 2003, 2010). If our ancestors were as negativity-biased as we are, negative cooperative dilemmas would have dwarfed positive ones in determining the long-run distribution of reputations. People condemn others' moral transgressions more severely when they are the result of deliberate actions, compared to equal but intentional inactions (Spranca et al. 1991; Baron and Ritov 2004; Cushman et al. 2006). Correspondingly, people seem less disposed to transgress by commission than omission (Ritov and Baron 1999), especially if they might be punished by others (DeScioli et al. 2011). These effects, which seem peculiar to negative commissions (Spranca et al. 1991) not positive ones, support our model's emphasis on negative cooperation by commission alone.

MODEL AND RESULTS

We are interested in whether detrimental exploitation can be curbed with a simple form of reputation that demands only limited cognitive capacities, and whether this can be used to sustain communal contributions and adherence to norms in other interactions. To tackle this puzzle, we construct a model of Negative Indirect Reciprocity (NIR) where we analyze interactions between very different kinds of individuals, such as reputation-contingent cooperators who always cooperate with well-reputed individuals or obligate defectors who exploit at every turn. We can thus reason formally about what kinds of strategies would be favored by selective evolutionary processes, whether via genetic or cultural evolution. Figure 1 lays out the basic elements of our NIR model. We first solve the model and describe its properties, and then discuss the degree of public goods provisioning that NIR supports.

To begin, imagine a single, large population of individuals who each have a "reputation"—a community-wide opinion of them that can influence others' behavior—which can be either "good" or "bad". We represent this reputation as a binary stochastic variable whose stationary distribution (denoted G) is the probability of being "good" on average. Reputations are determined by a person's actions in two kinds of social situations: with probability (1 - ρ), chance furnishes each individual in the population with an opportunity to gainfully exploit (and potentially be exploited by) a random peer; with probability ρ, individuals instead face an opportunity to improve their reputation by paying a cost. We refer to the former as the "theft game" and the latter as the "contribution game". The parameter ρ expresses the relative frequency with which each scenario occurs.

In the theft game, people can choose either to exploit their peers ($X = 1$) to accrue a personal gain (the takings, t) at the expense of the victim who suffers harm (damage, d), or do nothing ($X = 0$). Important reputational implications follow in each case. If an individual chooses exploitation, we assume that the thief's reputation declines only if the victim has a good reputation in the community—people do not care about what happens to poorly regarded victims. Thus, in this model and under IR more generally, individuals with "good" reputations are defined as those publicly well-liked enough, with enough friends, allies, or social connections, that actions directed towards them carry reputational consequences. If you exploit someone with a good reputation you acquire a bad reputation. If an individual chooses instead not to exploit a potential target, we assume that no one notices their inaction and nothing changes (assuming their propriety is correctly perceived). This novel assumption lessens the cognitive sophistication assumed by our model relative to existing IR models. With probability η, an individual's reputation is misperceived such that someone who refrains from exploitation is mistakenly thought to have defected.

In the contribution game, people can choose to either pay to improve their reputation (*Y = 1*) by contributing a public benefit *b* at personal cost *c*, or do nothing (*Y = 0*). To deliberately improve your peers' opinion of you, you need to know what pleases them as a group. This naturally suggests provisioning public goods (providing for a public feast, communal defense, or chasing away pests or predators) but could also include conformity to others' preferred behavioral standards and imitation of the best-reputed individuals (and so b need not be positive). Here, to better understand how the socio-ecology of NIR unfolds once norms have become established we consider the possibility that forfeiting an opportunity to improve one's reputation (e.g., by not sharing a fortunate day's catch), whether deliberately or by accident, actually worsens one's reputation (with probability ζ). As ζ increases, voluntary cooperative contributions become mandatory or normatively cooperative actions—think about giving to charity versus paying taxes. This parameter also nests the possibility that inaction is ignored as before (when $\zeta = 0$). Additionally, following Panchanathan and Boyd (2004) we allow for positive assortment in group formation with strength r, so that the probability of encountering another person of the same type (equivalently, the expected fraction of individuals of the same type in the group) is $r + (1 - r)p$ where p is the frequency of that strategy in the population (and the complementary probability is $(1 - r)(1 - p)$). Finally, we assume that individuals who try to improve their reputation can accidentally be misperceived with probability ε as having made no such attempt, though the cost is still exacted and the benefit still produced.

We consider four different strategies defined by their behavior in each game:

1) Obligate Defectors (*D*) who exploit everyone and never contribute ($X = 1$; $Y = 0$),
2) Reputational Cooperators (*R*) who never exploit the well-reputed and always contribute ($X = 0$; $Y = 1$),
3) Stingy Types (*S*) who never exploit the well-reputed but also do not contribute ($X = 0$; $Y = 0$), and
4) Mafiosos (*M*) who exploit everyone but contribute to the public good ($X = 1$; $Y = 1$).

Since Obligate Cooperators (who play $X = 0$ and $Y = 1$ regardless of reputation) are dominated by Reputational Cooperators (see supplemental materials section 4), we do not consider them further. Our main analysis establishes conditions under which a population of reputational cooperators is stable against rare invaders of each type (stability conditions for all other strategies are provided in the supplemental materials section 5).

Stability of *Reputational Cooperator* population against *Defector* invasion

In a population of common *R* with rare *D* playing the contribution game, an individual with strategy *R* gains benefit *b* from interaction with other *R*s, and always pays the contribution cost *c*. In the theft game, they gain takings *t* when encountering another individual who is in bad standing and suffer damage *d* when they are themselves in bad standing (since that is the only time other *R*s will exploit them). The (long-run mean) fitness of *R* here is thus

$$w_R = \rho\{b(r + (1 - r)p_R) - c\} + (1 - \rho)\{t(1 - G_R) - d(1 - G_R)\},$$

where $p_R \approx 1$ is the population frequency of *R*, and G_R is the (steady state) probability that an *R* strategist is in good standing. An individual with strategy *D* also gains *b* when they interact with

*R*s, but never pays *c* in the contribution game. They always exploit others in the theft game and hence always gain *t*, but lose *d* when they are in bad standing. The fitness of *D* is thus

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t - d(1-G_D)\},$$

where G_D is the probability that someone playing *D* is in good standing.

In the long run, the probability of an agent having a good reputation is well approximated by the mean of its stationary distribution; that is, $G = \frac{P_g}{P_g + P_b}$ where P_g and P_b are the probabilities of good and bad reputational transitions. An individual arrives at good standing only by paying for reputation and being correctly perceived as such, so $P_g = \rho Y(1-\varepsilon)$. They fall to bad standing by failing to pay when the community cares or by stealing from someone in good standing (or being misperceived as having committed either transgression), so in a population of *R*s, $P_b = \rho[(1-Y) + Y\varepsilon]\zeta + (1-\rho)G_R[X + (1-X)\eta]$. Thus,

$$G_i = \frac{\rho Y_i(1-\varepsilon)}{\rho[Y_i(1-\varepsilon)(1-\zeta) + \zeta] + (1-\rho)G_R[X_i + (1-X_i)\eta]},$$

so $G_D = 0$, and $G_R = \frac{\rho(1-\varepsilon)}{\rho\left(1-\varepsilon(1-\zeta)\right) + (1-\rho)G_R\eta}$ is the solution to the quadratic equation $(1-\rho)\eta G_R^2 + \rho\left(1-\varepsilon(1-\zeta)\right)G_R - \rho(1-\varepsilon) = 0$. This solution is opaque and hard to interpret analytically (though written out in the supplemental materials section 3)—so, in what follows, we will develop bounds that approximate the solution and depict its properties more clearly. Note that when errors are small ($\varepsilon, \eta \to 0$), $G_R \to 1$. Intuitively, this happens because *R*s never intentionally do anything that would place them in bad standing, and always pay to improve their reputation.

R is stable against invasion by *D* ($w_R > w_D$) when

$$\rho\{rb - c\} + (1-\rho)\{t(1 - G_R - 1) - d(1 - G_R - (1 - G_D))\} > 0$$

$$(1-\rho)\{d - t\}G_R > \rho\{c - rb\}$$

$$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right). \qquad (1)$$

This holds assuming that *c* > *rb*. If *rb* > *c,* cooperation will evolve simply via the non-random association captured in *r*. So, this formulation show how NIR can expand the conditions favorable to cooperation beyond *r*. This expression is closely related to the basin of attraction for the *R* regime, $p_R > \frac{c-rb}{d-t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right)$ as shown in section 2 of the supplemental materials, which also includes basins of attraction for strategy trios. To obtain a refined approximation of $\left(\frac{1}{G_R}\right)$, we first expand out its expression and subsequently assume that errors are small. By the preceding computations we have that

$$\frac{1}{G_R} = \frac{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}{\rho(1-\varepsilon)} = \left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{1-\rho}{\rho}\frac{\eta}{1-\varepsilon}G_R,$$

meaning that the right-hand side of the stability condition is

$$\frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right) = \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{\eta}{1-\varepsilon}G_R.$$

When errors are small, so $G_R \to 1$, the stability condition for *R* to resist *D* is approximately

$$\underbrace{\frac{d-t}{c-rb}}_{\substack{\textit{Ratio of net costs}\\ \textit{from two games}}} > \underbrace{\frac{\rho}{1-\rho}}_{\substack{\textit{Odds of}\\ \textit{contribution}\\ \textit{relative to}\\ \textit{theft game}}} \underbrace{\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}}_{\substack{\textit{Impact of the errors}\\ \textit{and judgements}}}. \qquad (2)$$

This reflects an upper bound on the right-hand side since G_R is bounded above by 1, therefore whenever our approximation (2) is satisfied, the exact condition (1) is always also satisfied; the two conditions coincide exactly when $\eta = 0$. The simulations depicted in Figure 2 illustrate the accuracy and conservative nature of the approximation, especially when errors are small (see section 1 of the supplemental materials for extensive simulations).

This stability condition (2) holds a number of meaningful implications. First, defectors will struggle to invade when exploitation is more *inefficient*—yielding relatively less benefit to the exploiter (t) than the harm it does their victim (d). Intuitively, $d > t$ when the strong and healthy steal from or injure the weak, old, and sick. Second, with positive assortment ($r > 0$), the most stable arrangements are those in which the contributed public benefits (b) are sufficiently large relative to the cost of provision (c), as will be discussed later. That said, even neutral or harmful norms (where $b \leq 0$) can be maintained under certain (more stringent) conditions. For example, both b and r can be zero and R can still be stable. Third, public contributions can only be sustained by the disciplining force of the theft game. Hence, the latter must occur sufficiently often relative to the former, meaning ρ cannot be too large. If $\rho = 0$, the condition holds and R cannot be invaded as long as $d > t$. Fourth, errors are always detrimental to stability, as the right-hand side terms are increasing in ε and η. Their multiplicative relationship also implies the errors compound each other, as the effect of η (doing nothing misperceived as exploitation) is increasing in ε (contribution misperceived as inaction). Finally, intriguingly, the propensity for the community to frown on non-contribution has an adverse effect on the stability of R. Intuitively, this happens because defectors never have good reputations in the long-run, so punishment for non-contribution harms mostly cooperators that are erroneously perceived to have shirked their communal duties; this is made clear by observing that the effect of ζ relies entirely on its interaction with ε. Thus NIR appears most effective at staving off defectors in early societies, before more complex cognitive faculties have developed—but as we will see later, selection pressures entail that when people are strongly expected to contribute, the public benefits produced in equilibrium tend to be more highly valued.

Stability of *Reputational Cooperator* population against *Stingy* invasion

In a population of common R with rare S, an individual with strategy R again has fitness

$$w_R = \rho\{b(r + (1 - r)p_R) - c\} + (1 - \rho)\{t(1 - G_R) - d(1 - G_R)\}.$$

An *S* does not pay in the contribution game, and so earns *b* only when meeting *R*s. They exploit only those in bad standing in the theft game and are exploited when they are themselves in bad standing. The fitness of strategy *S* is thus

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)\{t(1-G_R) - d(1-G_S)\}.$$

Since *S*s never pay for reputational improvements, they have no other way to achieve good standing and hence $G_S = 0$. Thus, assuming that *c* > *rb*, *R* is stable against invasion by *S* ($w_R > w_S$) when

$$\frac{d}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right). \tag{3}$$

Since *t* > *0*, this is a less stringent version of the stability condition against defectors. Therefore, when a population of *R* is stable against *D*, it is also stable against *S*, and the results of the previous section apply here equivalently.

Stability of *Reputational Cooperator* population against *Mafioso* Invasion

In a population of common *R* with rare *Mafiosos*, an individual with strategy *R* always gains *b* and pays *c* in any contribution event, exploits only the ill reputed in the theft game, and is exploited only when in ill repute. The fitness of *R* here is thus

$$w_R = \rho\{b-c\} + (1-\rho)\{t(1-G_R) - d(1-G_R)\}.$$

An *M* also gains *b* and pays *c* in the contribution game, but exploits everyone in the theft game, and hence has fitness

$$w_M = \rho\{b-c\} + (1-\rho)\{t - d(1-G_M)\}.$$

Thus, *R* is stable against invasion by *M* ($w_R > w_M$) when

$$t(1-G_R-1) - d\big(1-G_R-(1-G_M)\big) > 0$$

$$d(G_R - G_M) > tG_R$$

$$\frac{d}{t} > \frac{G_R}{G_R - G_M}. \tag{4}$$

This expression is closely tied to the basin of attraction for the *R* regime, $p_R > \left(\frac{t}{d-t}\right)\left(\frac{G_M}{G_R-G_M}\right)$ as shown in section 2 of the supplemental materials.

Here, *M*s are in good standing some of the time:

$$G_M = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R},$$

and recall that

$$G_R = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}.$$

Hence,

$$\frac{G_R - G_M}{G_R} = 1 - \frac{G_M}{G_R} = 1 - \frac{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R} = \frac{(1-\rho)G_R(1-\eta)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R},$$

and its reciprocal is

$$\frac{G_R}{G_R - G_M} = \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\left(\frac{1-\varepsilon(1-\zeta)}{G_R}\right)\right].$$

As before, for added insight we expand out G_R, and as shown in the appendix we obtain the approximate (upper bound) stability condition:

$$\frac{d}{t} > \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\right]. \tag{5}$$

The simulations depicted in Figure 3 indicate that this approximation mimics the properties of the exact solution (and it is indeed exact when $\eta = 0$), and several other bounds laid out in section 1 of the supplemental materials converge on similar predictions.

This stability condition (5) has several interesting implications. First, as in the case of the defector invasion, the existence of reputation-based cooperation requires exploitation to be inefficient (*d* > *t*). Second, the costs and benefits in the contribution game are not relevant here because both types pay for reputation. Third, as before, contributions are sustained by the threat of punishment via exploitation in the theft game, so $1-\rho$ must be reasonably large. Fourth, positive expectations of contributing still make cooperation harder to sustain; the derivative of the right-hand side with respect to is ζ is $\frac{\rho}{1-\rho}\left(\frac{2\varepsilon}{1-\eta}\right)\left(1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right)$ which is always positive and crucially dependent on ε.

More surprisingly, in some cases errors can be beneficial for reputational cooperators. While η always has strong adverse effects that magnify the threshold, a higher ε can actually be advantageous. Intuitively, this happens because although errors in the contribution game are bad for both strategies, they can be even worse for Mafiosos because they often fall into disrepute due to their exploitative ways, and are thus more in need of a reliable path back to good standing. This effect turns out to be beneficial on net when non-contribution is not penalized, that is, when ζ is low, so that reputational cooperators are not punished too harshly for others' mistaken perceptions. To illustrate this mathematically, observe that the key middle term $\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}$ reflecting the interaction is $1-\varepsilon$ when $\zeta=0$ (which is decreasing in ε) but $\frac{1}{1-\varepsilon}$ when $\zeta=1$ (which is increasing in ε). More generally, the derivative of the right-hand side with respect to ε is $\frac{1}{1-\eta}\frac{\rho}{1-\rho}\left[\frac{\left(2\zeta-\left(1-\varepsilon(1-\zeta)\right)\right)\left(1-\varepsilon(1-\zeta)\right)}{(1-\varepsilon)^2}\right]$, which is negative when $\zeta<\frac{1-\varepsilon}{2-\varepsilon}$. In the small error limit where $\varepsilon\to 0$, this inequality simplifies to $\zeta<\frac{1}{2}$. Figure S3 in the supplemental materials shows how the minimum stability threshold for *d/t* changes with each parameter when ζ is small, depicting the reversal of ε's effect. This further indicates that conditions are most favorable for NIR when ζ is small.

Stages of NIR and sustainable cooperation

What are the consequences of NIR on cooperative outcomes? Through the lens of our model we envision three progressive stages of socio-cognitive complexity, embodied in special cases of our parameters, which generate different levels of cooperation. Figure 4 presents the logic of our perspective. We begin with a plausible situation, early in our evolutionary history. The cognitive and behavioral prerequisites for reputations are in place: individuals selectively like or dislike their peers, and care or, selectively, do not care about how third parties treat them. The cultural transmission of reputations (opinions about others) is new, on evolutionary timescales. Here, however, second-order strategic responses to the existence of fitness-relevant reputations have not arisen yet: individuals do not actively monitor others' opinions of them or seek out opportunities to improve their reputation. In this earliest, least cognitively demanding stage, reputations were improved only by good fortune, not by deliberate effort. In such an environment, even if inaction is unobservable, selection can sustain harmony. This Stage 1 occurs in our model when $\rho \to 0$; inequality (2) reveals that reputation-based reciprocity is then stable whenever $d > t$. Here, NIR can establish more harmonious communities that limit exploitation of others—the weak, injured, sick and elderly—though no public goods are provided in this first stage.

Deleted:

Even when individuals are unaware of their own reputations, oblivious to inaction and to anything that happens to the ill-reputed, the dynamics of the first stage can coordinate the weighty fitness consequences of community-wide exploitation. This opens up a new selective landscape, where selection favors monitoring one's own reputation and deliberately acting to improve it. We explore the unfolding of NIR dynamics by opening up the possibility that individuals notice costly opportunities to improve their reputation, which happens when ρ increases above zero. We explore what happens if opportunities for reputational improvement can be ignored without adverse consequences ($\zeta \rightarrow 0$). In this socio-ecology of Stage 2, your peers are delighted if you share food with them, but barely notice if you instead keep it for yourself.

Here, expression (2) entails that cooperation can be sustained when $\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}$ (assuming small errors). Then some positive amount of reputational norm adherence occurs, but the resulting public benefits must be large enough to resists defectors. Specifically, rearranging the inequality reveals that we need

Deleted:

$$rb > c - \left(\frac{1-\rho}{\rho}\right)(d-t). \tag{6}$$

This inequality shows how the theft game eases the standard conditions for cooperation created by non-random association ($rb > c$). The larger ρ and less inefficient theft ($d - t$) is, the easier it is to maintain cooperation. The right-hand side of (6) is increasing in ρ (supposing $d > t$) as its derivative with respect to ρ is $\left(\frac{1-\rho}{\rho}\right)^2 (d-t) > 0$, meaning that selection pressures enforce a higher minimum benefit provided in equilibrium as Stage 2 progresses. Figure 5 shows that this property is shared by the exact solution (including both types of errors). Though neutral or even harmful behaviors can potentially be sustained when the right-hand side of the inequality is negative, positive contributions will be particularly favored. We view this voluntary public goods provisioning as a key transitional phase, where selection begins to favor individuals who pay closer attention to their reputation and opportunities to improve it, and therefore to their community's behavioral expectations. To deliberately improve your reputation, you need to know what pleases your peers. Stage 2 provides a plausible cognitive foundation for the emergence of "social norms" (Chudek and Henrich 2011; Henrich 2016).

Once the evolutionary processes in Stage 2 have selected for individuals who attend carefully to their own reputations and opportunities to improve it, it is natural to ask what would happen if individuals also began attending to other's reputations and opportunities. Once an evolutionary mechanism has led people to regularly contribute to others' welfare (e.g., sharing their surplus forage to improve their reputation), it is more plausible that individuals would begin to notice others' opportunities to do this, and have a reputation-relevant reaction to their inaction. Here, we ask what would happen if failing to act on reputation improvement opportunity actually worsened one's reputation, characteristic of Stage 3. The ζ parameter describes a continuous transition from voluntary to mandatory norm-following ($\zeta \to 1$), including public goods provisioning, as individuals become more conscious of other individual's reputations and failures to conform to normative expectations. Rearranging expression (2), supposing $\eta \to 0$ for clarity, implies that $rb > c - \left(\frac{1-\rho}{\rho}\right)\left(\frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}\right)(d-t)$. The right-hand side is increasing in ζ (supposing $d > t$) as its derivative with respect to ζ is $\left(\frac{1-\rho}{\rho}\right)\left(\frac{\varepsilon(1-\varepsilon)}{\left(1-\varepsilon(1-\zeta)\right)^2}\right)(d-t) > 0$, and Figure 5 demonstrates that the exact solution shares this property. Hence the minimum benefit provided in equilibrium must grow even larger in Stage 3. Of course, a costly and mandatory reputation-improving norm behavior can still be sustained even if it delivers no benefit at all ($b = 0$) as long as that $c < \left(\frac{1-\rho}{\rho}\right)(1-\varepsilon)(d-t)$.

Deleted:

The overarching trend is thus for public goods provision to improve throughout the progression of stages. However, this comes at a price: stable states are harder to come by, as the requirements for cooperative equilibria become stricter (unless selection has also been acting to reduce people's inclination to make errors or misperceive others' actions). This means that NIR is most capable of limiting exploitation early on, but is also capable of supporting the production of communal benefits especially under conditions when errors and misperception are high, such as in large groups. As opportunities for reputational improvement via norm adherence rise in prevalence, exploitation becomes harder to control but higher-value public goods are reaped in compensation (indeed, the latter is the reason for the former). In the extreme case, stable equilibria may become sufficiently rare that NIR is no longer viable at large scale. This raises the intriguing possibility that NIR could render itself obsolete; it might be a transitional step along the path to widespread cooperation bolstered by other mechanisms. While NIR may not vanish completely, such an analysis suggests that it would naturally set the stage for, and then give way to the more cognitively complex reputation systems that have been previously proposed. So, despite the modern prominence of positive indirect reciprocity, it may have been mid-wifed into existence by NIR.

DISCUSSION

Building from minimal cognitive prerequisites, plausibly found in our Pliocene ancestors, we have mapped a path to larger-scale forms of human cooperation by first suppressing within-group exploitation (such as theft or rape), and then harnessing exploitation to sustain arbitrary, costly reputation-raising acts. Crucially, these reputation-raising acts may include cooperative contributions to others' welfare, such as meat sharing or communal defense.

Stage 1 describes dynamics when reputational systems first emerge: if community members are sufficiently reluctant to exploit their well-reputed peers, selective forces will sustain and enhance this reluctance, perpetuating harmonious (i.e., non-exploiting) communities. This is particularly likely if there are many opportunities to exploit others that benefit perpetrators little relative to the harm they cause their victims. Such circumstances minimize benefits to indiscriminate exploiters and maximize the value of a good reputation. Our postulated reputational system imposes only minimal cognitive demands on early reputational cooperators, since they can ignore (1) anything that happens to people in bad standing, (2) all "non-events" (like not exploiting), and (3) the exploiter's previous reputation. By contrast, the stable cooperative equilibrium in positive indirect reciprocity models require communities to converge on a single reputational system that specifies up to eight (2^3) possible events, defined by the target's reputation (good/bad), the actor's reputation (good/bad), and their action (help/inaction) (Ohtsuki and Iwasa 2004, 2006). Even the simplest strategy ("image-scoring"; Nowak and Sigmund 1998), which is not evolutionarily stable (Panchanathan et al. 2003), requires individuals to track non-events or notice inactions (failure to "help").

The conditions explored in our Stage 1 model may have been particularly likely in ancestral human societies. When individuals fell sick, were injured, or faced emergencies requiring them to rapidly leave camp, exploiters had opportunities to steal food, mating opportunities, allies, beads, and raw materials (like skins, flint, ochre, and obsidian) with little chance of direct retribution, either because the victim could not pinpoint the perpetrator or was in no position to enact revenge. In times of distress (illness or injuries) exploitation is particularly easy and the loss of valuable resources is particularly damaging (Wrangham 2009).

Once harmonious communities develop in Stage 1 and reputations carry fitness consequences, selection can favor individuals disposed to act in costly ways that improve their reputation. Achieving this requires an awareness of others' expectations, favoring cognitive adaptations for noticing and navigating social norms (Chudek and Henrich 2011; Henrich 2016). These norms, which themselves can become the object of evolutionary dynamics, potentially include contributions to others' welfare and to larger scale cooperative endeavors. This puts a community's normative behavioral expectations on the culture-gene co-evolutionary landscape that shapes its members' behavior, cognitive abilities and motivations in the long-run.

Interestingly, it is the central challenge of NIR (that "negative cooperation", i.e., not exploiting others), is typically unobservable and so cannot reliably improve reputations that leads to pressure for the cognitive abilities assumed by many existing models of human cooperation—that individuals can indeed recognize and rapidly coordinate on arbitrary shared norms. This includes nearly all models based on reputations or indirect reciprocity as well as costly punishment models. Once NIR's evolutionary dynamics create fitness consequences for shared expectations and cause individuals to sometimes (when is not too costly) do whatever it takes to satisfy those expectations, these dynamics can push communities even closer to full-blown social norms and a psychology for navigating them. If individuals are sensitive to others' opportunities for reputation-raising acts and are disappointed by their absence, counter-normative behavior can actually lower one's reputation and invite opportunistic exploitation from one's peers. The more frequent is this kind of disappointment at counter-normative actions (or even inactions), the more strongly selection favors adherence to community norms.

To thrive in the social ecologies enabled by NIR, individuals must be quick to perceive their community's norms (the behaviors that please others on average, which could include generosity in times of plenty, sharing adaptive knowledge or resting on the Sabbath) and be disposed to adhere to them. Communities meanwhile come to wield a powerful means of enforcing compliance to these norms. This distributed mechanism for norm enforcement can emerge without any individuals necessarily intending it; they merely selfishly exploit friendless, low-status victims when the opportunity arises because they know they can get away with it. Indeed, it is possible we still witness these dynamics today, as the recurrent emergence of schoolyard bullying recapitulates the socio-ecological dynamics of early, pre-institutional human societies (Card et al. 2008; Merrell et al. 2008; Rodkin and Berger 2008). Or, as with the Yasawans, individual grudges can be transformed into an instrument for societal harmony.

In some cases, NIR can sustain costly adherence to nearly any community standard, which means that it can potentially sustain both cooperative norms (public goods) as well as maladaptive norms (public bads). We see this as an advantage of our models since the ethnographic record is replete with examples of social norms that are costly for the individual (reputation effects aside) and maladaptive at the group level. Classic examples include female infibulation and mortuary consumption of dead relatives, which promotes the spread of prion diseases like Kuru (Glasse 1963; Edgerton 1992).

Nevertheless, there are two reasons to suspect that over time reputationally-enforced norms will tend to become increasingly prosocial. First, actions that improve others' welfare may be especially likely to raise people's opinion of an actor. This creates what cultural evolutionists have termed a "content bias" that favors bestowing good reputations for highly-salient acts that generate benefits for others (Henrich and McElreath 2007). Second, by making deviations from community expectations costly, NIR favors migrants who adopt the norms of their new community rather than maintaining their old behaviors. This decreases behavioral variability within groups relative to variation between communities, which increases the strength of the between-group component of selection in cultural evolution. Thus, intergroup competition can favor contributions to communal defense, raiding, economic productivity, alliance building, trading and information sharing (Chudek and Henrich 2011; Henrich 2016). Such logic is partly reflected in our supposition of positive assortment, which ties equilibrium outcomes to the value of contributions and enables the rise in public benefits provided across the stages of NIR. Note, in this proposal, the between-group selective process operates through cultural evolution while the within-group selective processes can be either cultural or genetic. Purely genetic group selection is unlikely to play a large role in human cooperation due to the substantial rates of gene flow among groups (Henrich 2016); these same concerns do not apply to cultural evolution (Boyd and Richerson 2002; Henrich and Henrich 2007; Boyd et al. 2011).

We suspect that NIR's dynamics might be particularly important for the evolution of the human capacity for culture. Our species' capacity for cumulative cultural evolution was likely fostered by the dissemination of cultural know-how, about things such as tool-making and food processing, across communities and through broad social networks (Henrich 2016). However, apparently knowledgeable individuals could actively exploit others by spreading false information. NIR dynamics may have helped cumulative cultural evolution get off the ground by suppressing people's inclinations to spread false information to those with a good reputation. Those with bad reputations could be fed misinformation or given no cultural information.

Overall, the cognitive and socio-ecological conditions fostered by NIR should make it easier for more potent, coordinated or institutional forms of cooperation to emerge. The more common such norms or institutions become, in which non-prosocial behavior is punished, the stronger the selection pressure on individuals to (1) default toward prosociality and (2) rapidly acquire prosocial norms relative to antisocial norms. This process may explain both the unusually high-levels of prosociality found in infants and children as well as their inclinations towards learning prosocial norms (Warneken 2015; McAuliffe et al. 2017a, b).

Future research can test these models in at least three ways. First, both field and experimental work in non-human primates can explore the extent to which the most basic cognitive abilities and motivational inclinations we have assumed in our model exist in related species (e.g., Herrmann et al. 2013). This could better ground our assumptions about our early ancestors or else jeopardize our starting point. Such non-human research might also explore if some species are already implementing NIR, effectively suppressing exploitation through some form of shared judgment. Second, cross-cultural developmental psychologists should continue to examine the ontogeny of the cognitive abilities and motivational inclinations looking for the biases we predict. What is the developing structure of children's strategies for judging others? Do infants and children more readily observe, evaluate, and track actions related to “harming” (exploitation) compared to inactions related to “helping” (e.g., Hamlin et al. 2011; Hamlin 2013)? How do infants and children evaluate individuals who exploit other exploiters vs. those who exploit non-exploiters? Can positive helping norms develop in an environment in which exploitation is common? And finally, anthropological work in diverse societies, especially small-scale societies lacking formal institutions, can explore whether negative indirect reciprocity underpins common forms of cooperation and public goods (e.g., Henrich and Henrich 2014).

ETHICAL STATEMENT

Funding: No special funding was available to write this article

Conflict of interest: None

Ethical approval: Our study did not involve animal subjects

Informed consent: Not applicable

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article, as no datasets were generated or analyzed during the current study.

REFERENCES

Axelrod R, Hamilton W (1981) The evolution of cooperation. Science 211:1390-1396.

Baron J, Ritov I (2004) Omission bias, individual differences, and normality. Org Behav Human Dec Proc 94:74-85.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. Rev Gen Psychol 5:323.

Boyd R, Gintis H, Bowles S (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328:617-620.

Boyd R, Richerson PJ (1988a) Culture and the evolutionary process. University of Chicago Press, Chicago.

Boyd R, Richerson PJ (1988b) The evolution of reciprocity in sizable groups. J Theor Biol 132:337-356.

Boyd R, Richerson PJ (1996) Why culture is common, but cultural evolution is rare. Proc Brit Acad 88:77-94.

Boyd R, Richerson PJ (2002) Group beneficial norms can spread rapidly in a structured population. J Theor Biol 215:287-296.

Boyd R, Richerson PJ, Henrich J (2011) Rapid cultural adaptation can facilitate the evolution of large scale cooperation. Behav Ecol Sociobiol 65:431-444.

Cacioppo JT, Berntson GG (1994) Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. Psychol Bull 115:401.

Card NA, Stucky BD, Sawalani GM, Little TD (2008) Direct and indirect aggression during childhood and adolescence: a meta-analytic review of gender differences, inter-correlations, and relations to maladjustment. Child Devel 79:1185-1229.

Chudek M, Henrich J (2011) Culture-gene coevolution, norm-psychology and the emergence of human prosociality. Trends Cogn Sci 15:218-226.

Cushman F, Young L, Hauser M (2006) The role of conscious reasoning and intuition in moral judgment testing three principles of harm. Psychol Sci 17:1082-1089.

DeScioli P, Christner J, Kurzban R (2011) The omission strategy. Psychol Sci 22:442-446.

Edgerton RB (1992) Sick societies: challenging the myth of primitive harmony. Free Press, New York.

Fehr E, Gächter S (2000) Fairness and retaliation: The economics of reciprocity. J Econ Pers 14:159-181.

Fiske ST (1980) Attention and weight in person perception: the impact of negative and extreme behavior. J Pers Soc Psychol 38:889.

Glasse RM (1963) Cannibalism in the kuru region. Trans N Y Acad Sci 29:748-754.

Hamilton W (1964) The genetical evolution of social behavior. I. J Theor Biol 7:1-16.

Hamlin JK (2013) Moral judgment and action in preverbal infants and toddlers: evidence for an innate moral core. Curr Dir Psychol Sci 22:186-193.

Hamlin JK, Wynn K, Bloom P, Mahajan N (2011) How infants and toddlers react to antisocial others. Proc Nat Acad Sci 108:19931-19936.

Hamlin KJ, Wynn K, Bloom P, Hamlin JK, Wynn K, Bloom P (2010) Three-month-olds show a negativity bias in their social evaluations. Devel Sci 13:923-929.

Henrich J (2016) The secret of our success: how culture is driving human evolution, domesticating our species, and making us smart. Princeton University Press, Princeton.

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R (2001) In search of *Homo economicus*: behavioral experiments in 15 small-scale societies. Amer Econ Rev 91:73-78.

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R, Alvard M, Barr A, Ensminger J, Henrich NS (2005) "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. Behav Brain Sci 28:795-815.

Henrich J, Gil-White FJ (2001) The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evol Hum Behav 22:165-196.

Henrich J, Henrich N (2007) Why humans cooperate: a cultural and evolutionary explanation. Oxford University Press, Oxford.

Henrich J, Henrich N (2014) Fairness without punishment: behavioral experiments in the Yasawa Island, Fiji. In: Henrich J, Ensminger J (eds) Experimenting with social norms: fairness and punishment in cross-cultural perspective. Russell Sage Press, pp 171-218.

Henrich J, McElreath R (2007) Dual-inheritance theory: the evolution of human cultural capacities and cultural evolution. In: Dunbar R, Barrett L (eds) Oxford Handbook of Evolutionary Psychology. Oxford University Press, pp 555-570.

Henrich J, Smith N (2004) Comparative experimental evidence from Machiguenga, Mapuche, Huinca, and American populations. In: Henrich JP, Boyd R, Bowles S, Fehr E, Camerer C, Gintis H (eds) Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford University Press, Oxford, pp 125-167.

Herrmann E, Keupp S, Hare B, Vaish A, Tomasello M (2013) Direct and indirect reputation formation in nonhuman great apes (*Pan paniscus, Pan troglodytes, Gorilla, Pongo pygmaeus*) and human children (*Homo sapiens*). J Comp Psychol 127:63-75.

Higham JP, Maestripieri D (2010) Revolutionary coalitions in male rhesus macaques. Behav 13:1889-1908.

Knobe J (2003) Intentional action and side effects in ordinary language. Anal 63:190-194.

Knobe, J (2010) Person as scientist, person as moralist. Behav Brain Sci 33:315.

Langergraber KE, Mitani JC, Vigilant L (2007) The limited impact of kinship on cooperation in wild chimpanzees. Proc Nat Acad Sci 104:7786-7790.

Leimar O (2009) Multidimensional convergence stability. Evol Ecol Res 11:191-208.

Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. Proc Biol Sci 268:745-753.

McAuliffe K, Blake PR, Steinbeis N, Warneken F (2017a) The developmental foundations of human fairness. Nature Hum Behav 1:0042.

McAuliffe K, Raihani NJ, Dunham Y (2017b) Children are sensitive to norms of giving. Cognition 167: 151-159.

Merrell KW, Gueldner BA, Ross SW, Isava DM (2008) How effective are school bullying intervention programs? A meta-analysis of intervention research. Sch Psychol Q 23:26-42.

Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature 393:573-577.

Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437:1291-1298.

Ohtsuki H, Iwasa Y (2004) How should we define goodness? Reputation dynamics in indirect reciprocity. J Theor Biol 231:107-120.

Ohtsuki H, Iwasa Y (2006) The leading eight: social norms that can maintain cooperation by indirect reciprocity. J Theor Biol 239:435-444.

Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432:499-502.

Panchanathan K, Boyd R (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. J Theor Biol 224:115-126.

Perry S, Manson JH (2008) Manipulative monkeys: the capuchins of Lomas Barbudal. Harvard University Press.

Ritov I, Baron J (1999) Protected values and omission bias. Org Behav Human Dec Proc 79:79-94.

Rodkin PC, Berger C (2008) Who bullies whom? Social status asymmetries by victim gender. Int J Behav Devel 32:473-485.

Rozin P, Royzman EB (2001) Negativity bias, negativity dominance, and contagion. Pers Soc Psychol Rev 5:296-320.

Sigmund K, De Silva H, Traulsen A, Hauert C (2010) Social learning promotes institutions for governing the commons. Nature 466:861-863.

Silk JB (2002) Using the "F"-word in primatology. Behav 139:421-446.

Skowronski JJ, Carlston DE (1987) Social judgment and social memory: the role of cue diagnosticity in negativity, positivity, and extremity biases. J Pers Soc Psychol 52:689-699.

Spranca M, Minsk E, Baron J (1991) Omission and commission in judgment and choice. J Exp Social Psychol 27:76-105.

Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35-57.

Warneken F (2015) Are social norms and reciprocity necessary for early helping? Proc Nat Acad Sci 112:201423750.

Watts DP (2002) Reciprocity and interchange in the social relationships of wild male chimpanzees. Behav 139:343-370.

Wrangham R (2009) Catching fire: how cooking made us human. Basic Books.

FIGURE LEGENDS

Figure 1. The NIR decision tree. The probability of each branch is described by blue parameters, and evolving dispositions are represented by green variables (*Y*: disposition to pa*y* reputation improvement costs; and *X*: disposition to e*x*ploit well-reputed victims). Red text at terminal nodes describes the consequences of each outcome.

Figure 2. Minimum threshold values of *d-t/c-rb* required for reputational cooperation to be stable against rare defectors. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

Figure 3. Minimum threshold values of *d/t* required for reputational cooperation to be stable against rare Mafiosos. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

Figure 4. Socio-cognitive stages of NIR.

Figure 5. Increase in the minimum stability threshold for public benefit provision *b* across stages of NIR. Parameters are set at $d = 1$, $t = \frac{1}{2}$, $c = 1$, $r = \frac{1}{10}$, $\varepsilon = \eta = \frac{1}{10}$, $\zeta = \frac{1}{2}$ (left), and $\rho = \frac{3}{10}$ (right).

FIGURE 1

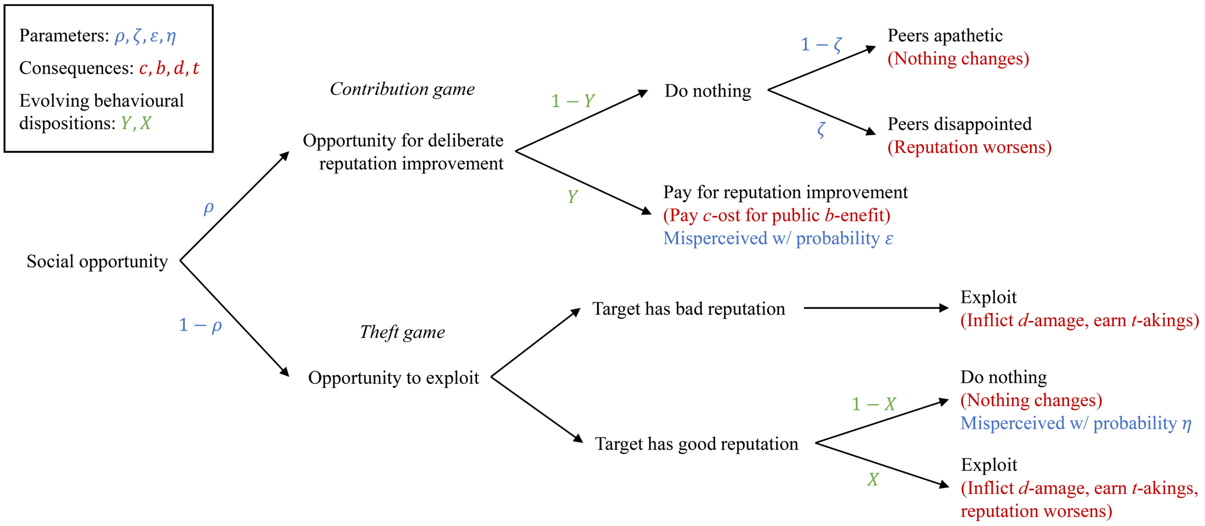
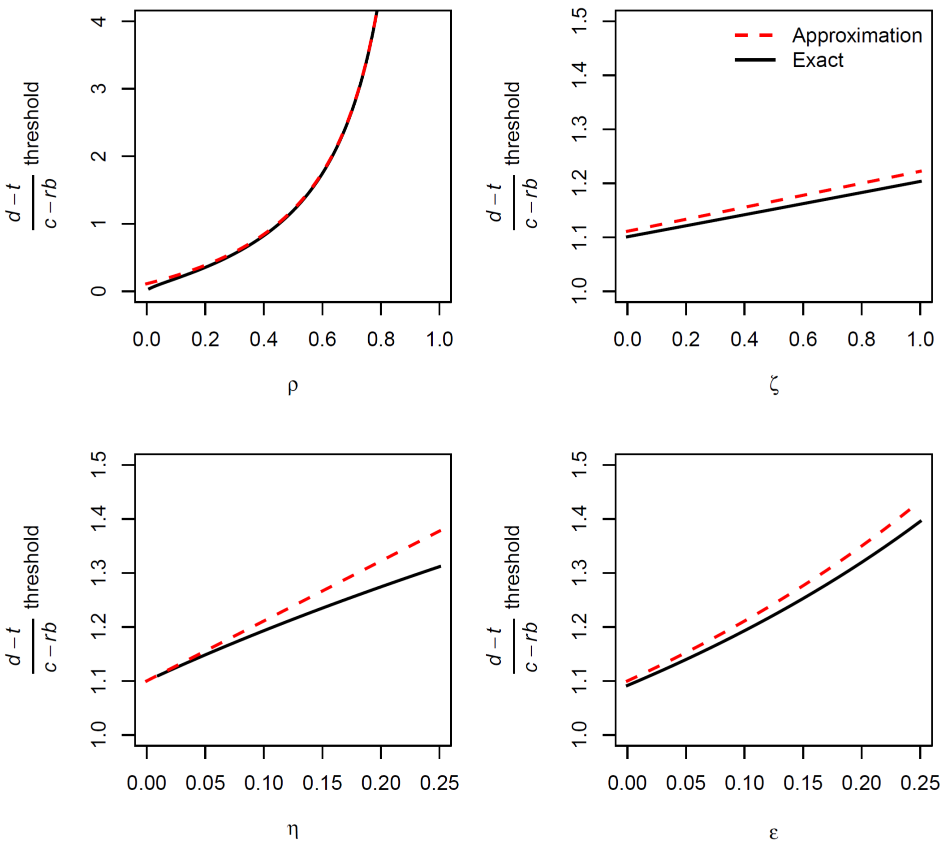

Parameters: $\rho, \zeta, \varepsilon, \eta$
Consequences: c, b, d, t
Evolving behavioural dispositions: Y, X
Social opportunity
ρ
$1 - \rho$
Contribution game
Opportunity for deliberate reputation improvement
$1 - Y$
Do nothing
$1 - \zeta$
Peers apathetic
(Nothing changes)
ζ
Peers disappointed
(Reputation worsens)
Y
Pay for reputation improvement
(Pay c-ost for public b-enefit)
Misperceived w/ probability ε
Theft game
Opportunity to exploit
Target has bad reputation
Exploit
(Inflict d-amage, earn t-akings)
Target has good reputation
$1 - X$
Do nothing
(Nothing changes)
Misperceived w/ probability η
X
Exploit
(Inflict d-amage, earn t-akings, reputation worsens)

FIGURE 2



$\frac{d-t}{c-rb}$ threshold
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
ρ
1.0
1.1
1.2
1.3
1.4
1.5
Approximation
Exact
ζ
0.00
0.05
0.10
0.15
0.20
0.25
η
ε

FIGURE 3

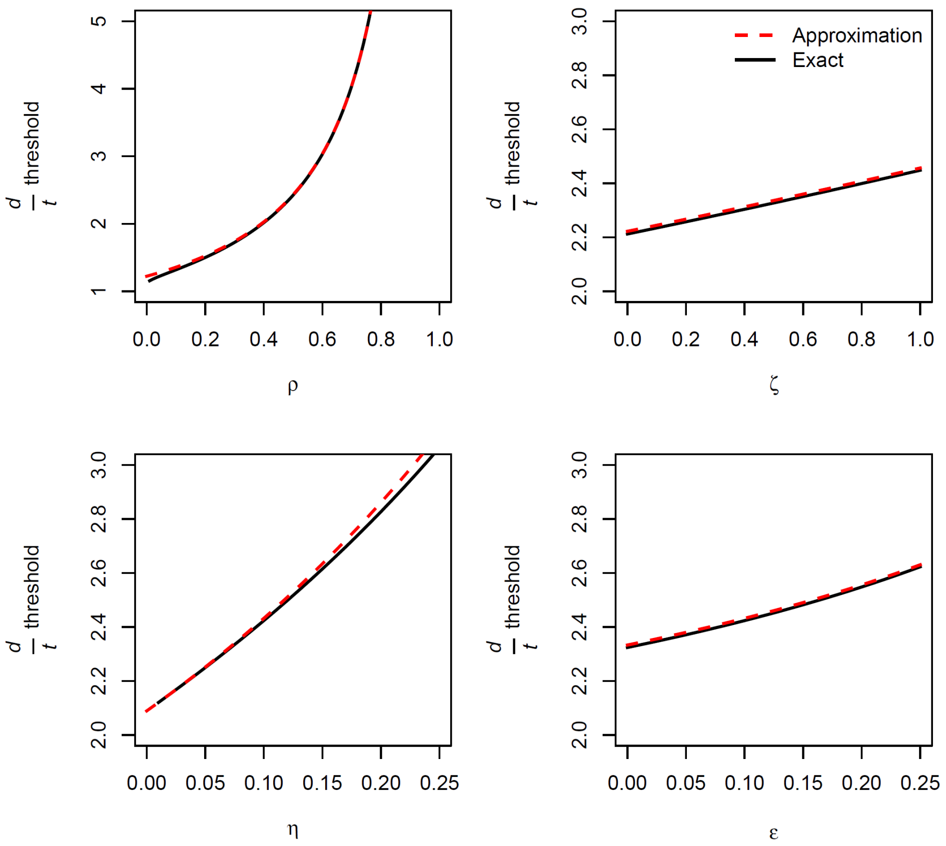
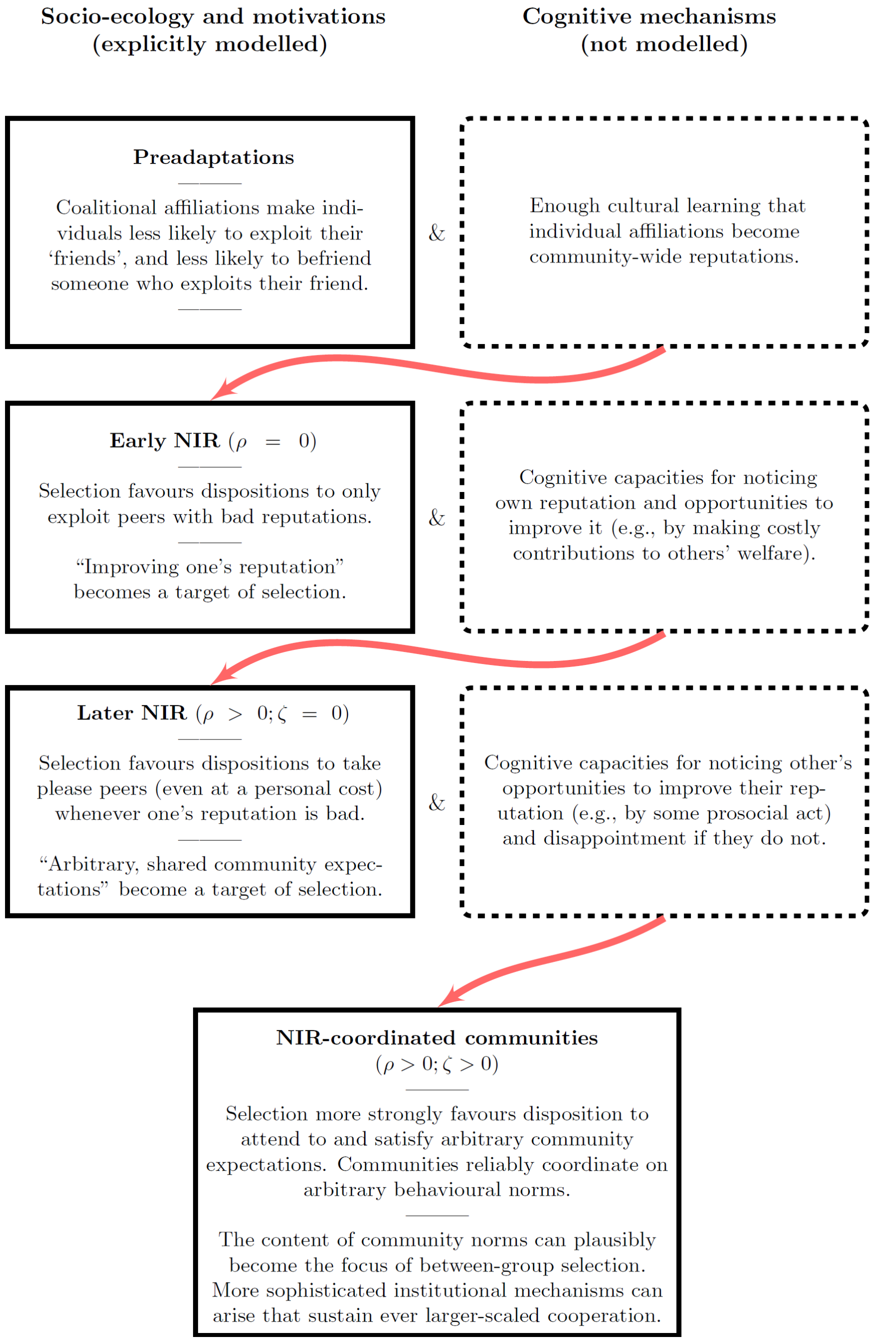

d/t threshold
5
4
3
2
1
0.0
0.2
0.4
0.6
0.8
1.0
ρ
3.0
2.8
2.6
2.4
2.2
2.0
Approximation
Exact
ζ
0.00
0.05
0.10
0.15
0.20
0.25
η
ε

FIGURE 4



Socio-ecology and motivations (explicitly modelled)

Cognitive mechanisms (not modelled)

Preadaptations

Coalitional affiliations make individuals less likely to exploit their ‘friends’, and less likely to befriend someone who exploits their friend.

&

Enough cultural learning that individual affiliations become community-wide reputations.

Early NIR ($\rho = 0$)

Selection favours dispositions to only exploit peers with bad reputations.

“Improving one’s reputation” becomes a target of selection.

&

Cognitive capacities for noticing own reputation and opportunities to improve it (e.g., by making costly contributions to others’ welfare).

Later NIR ($\rho > 0; \zeta = 0$)

Selection favours dispositions to take please peers (even at a personal cost) whenever one’s reputation is bad.

“Arbitrary, shared community expectations” become a target of selection.

&

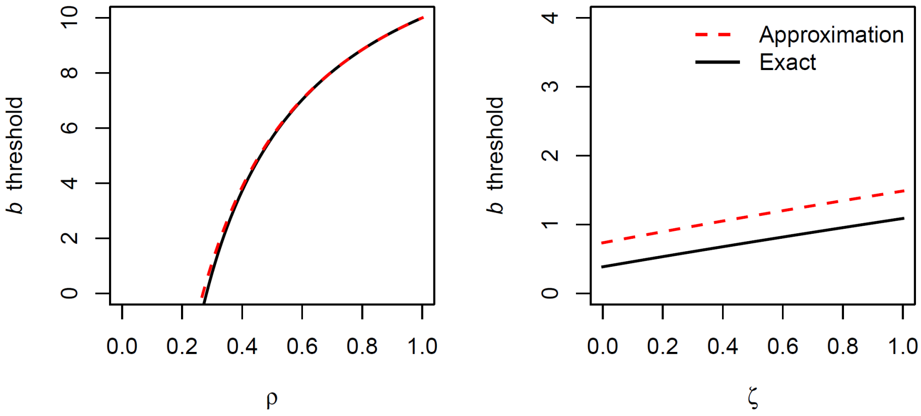
Cognitive capacities for noticing other’s opportunities to improve their reputation (e.g., by some prosocial act) and disappointment if they do not.

NIR-coordinated communities
($\rho > 0; \zeta > 0$)

Selection more strongly favours disposition to attend to and satisfy arbitrary community expectations. Communities reliably coordinate on arbitrary behavioural norms.

The content of community norms can plausibly become the focus of between-group selection. More sophisticated institutional mechanisms can arise that sustain ever larger-scaled cooperation.

FIGURE 5


b threshold
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
ρ
b threshold
0
1
2
3
4
Approximation
Exact
0.0
0.2
0.4
0.6
0.8
1.0
ζ

Supplemental Materials for:

How Exploitation Launched Human Cooperation

Rahul Bhui, Maciej Chudek, and Joseph Henrich

1. Further Approximations to Stability Conditions

The method used in the main text to yield interpretable analytical approximations to the stability conditions can be applied to yield many different expressions. Some expressions are easier to interpret but less accurate, while others are harder to interpret but more accurate. Here, we document and compare these alternative approximations to demonstrate the robustness of insights from the approximations focused on in the main text, and to explore this method further.

1a. Stability of Reputational Cooperator population against Defector/Stingy invasion

As shown in the main text, *R* is stable against *D* when

$$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right). \tag{S1}$$

(The same inequality, except without *t*, describes stability against *S*.)

Recall that the probability of an *R* type having a good reputation is given by

$$G_R = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}, \tag{S2}$$

and hence

$$\frac{1}{G_R} = \frac{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}{\rho(1-\varepsilon)} = \left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{1-\rho}{\rho}\frac{\eta}{1-\varepsilon}G_R. \tag{S3}$$

The approximations are derived by iteratively substituting the expanded expression for G_R into the stability condition, starting with substituting (S3) into (S1), and then applying the fact that G_R must be between 0 and 1. From this method we obtain four approximations: two upper bounds and two lower bounds.

The approximation in the main text results from one level of substitution:

$$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}G_R. \tag{S4}$$

Since $G_R \leq 1$, as in the main text, the RHS must be bounded above by

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}. \tag{S5}$$

Furthermore, since $G_R \geq 0$, the RHS must be bounded below by

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]. \tag{S6}$$

We refer to (S5) as Upper Bound 1 and (S6) as Lower Bound 1.

More refined approximations can be obtained by a second level of substitution, of (S2) into (S4):

$$\begin{aligned}&\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}\left(\frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}\right)\\ &=\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)G_R\eta+\big(1-\varepsilon(1-\zeta)\big)}.\end{aligned} \tag{S7}$$

The upper bound of this expression is

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon(1-\zeta)}. \tag{S8}$$

and the lower bound is

$$\frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\big(1-\varepsilon(1-\zeta)\big)}, \tag{S9}$$

We refer to (S8) as Upper Bound 2 and (S9) as Lower Bound 2. While any number of even more accurate bounds can be gained by iterating this process again and again, further expressions decline in ease of interpretation, and simulations indicate that the ones above are sufficiently faithful to the exact solution for our purposes. In fact, they are exact when $\eta = 0$.

All four bounds are collected in Table S1 and compared visually in Figure S1. As can be seen, more complex approximations tend to be more accurate; in this case, Lower Bound 2 is the most accurate. However, all approximations are generally quite good, with the partial exception of the simplest Lower Bound 1, which neglects to capture the noise parameter η at all.

1b. Stability of Reputational Cooperator population against Mafioso invasion

As shown in the main text, *R* is stable against *M* when

$$\frac{d}{t}>\frac{G_R}{G_R-G_M}, \tag{S10}$$

and

$$G_M = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R}, \tag{S11}$$

leading to

$$\frac{G_R}{G_R - G_M} = \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\left(\frac{1-\varepsilon(1-\zeta)}{G_R}\right)\right]. \tag{S12}$$

Using the same method as above, we start by substituting (S3) into (S12), and then applying the fact that G_R must be between 0 and 1.

This first level of substitution yields

$$\begin{aligned}\frac{G_R}{G_R - G_M} &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\big(1-\varepsilon(1-\zeta)\big)\left(\frac{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}{\rho(1-\varepsilon)}\right)\right] \\ &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}G_R\eta\right].\end{aligned} \tag{S13}$$

Since $G_R \leq 1$, this must be bounded above by

$$\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\right], \tag{S14}$$

and since $G_R \geq 0$, it must be bounded below by

$$\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon}\right]. \tag{S15}$$

We refer to (S5) as Upper Bound 1 and (S6) as Lower Bound 1.

Other approximations can be obtained by a second level of substitution, of (S2) into (S13):

$$\begin{aligned}&\frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\left(\frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)G_R\eta}\right)\right] \\ &= \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \frac{\eta\big(1-\varepsilon(1-\zeta)\big)}{\big(1-\varepsilon(1-\zeta)\big) + \left(\frac{1-\rho}{\rho}\right)G_R\eta}\right].\end{aligned} \tag{S16}$$

The upper bound of this expression (occurring when $G_R \to 0$) is presented in the main text:

$$\frac{d}{t} > \frac{1}{1-\eta}\left[1 + \frac{\rho}{1-\rho}\frac{\big(1-\varepsilon(1-\zeta)\big)^2}{1-\varepsilon} + \eta\right], \tag{S17}$$

and the lower bound is

$$\frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\left(1-\varepsilon(1-\zeta)\right)}\right]. \tag{S18}$$

We refer to (S17) as Upper Bound 2 and (S18) as Lower Bound 2. Again, while this process can be continually iterated, further expressions are less interpretable, and the ones above appear close to the exact solution. (Indeed, they are exact when $\eta = 0$.)

All four bounds are collected in Table S1 and compared visually in Figures S2 (high ζ) and S3 (low ζ). Again, the more complex approximations are generally more accurate; here, Lower Bound 2 has the lowest error. However, all approximations are generally good.

Table S1: Approximate stability conditions for population of reputational cooperators

	Rare invader in population of *Reputational Cooperators*	
	Defector / Stingy	*Mafioso*
Upper Bound 1	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\eta\frac{1-\varepsilon(1-\zeta)}{1-\varepsilon}\right]$
Upper Bound 2	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{1-\varepsilon(1-\zeta)}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\eta\right]$
Lower Bound 2	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\left(1-\varepsilon(1-\zeta)\right)}$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}+\frac{\eta}{\left(\frac{1-\rho}{\rho}\right)\eta+\left(1-\varepsilon(1-\zeta)\right)}\right]$
Lower Bound 1	$\frac{d-t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right]$	$\frac{d}{t} > \frac{1}{1-\eta}\left[1+\frac{\rho}{1-\rho}\frac{\left(1-\varepsilon(1-\zeta)\right)^2}{1-\varepsilon}\right]$

Note: Stability conditions assume $d > t$ and $c > rb$. Expressions for *Stingy* are same as for *Defector* except with $t = 0$. In the main text, Upper Bound 1 is presented for the *Defector* invasion and Upper Bound 2 for the *Mafioso* invasion.

Figure S1. Minimum threshold values of *d-t/c-rb* required for reputational cooperation to be stable against rare defectors. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

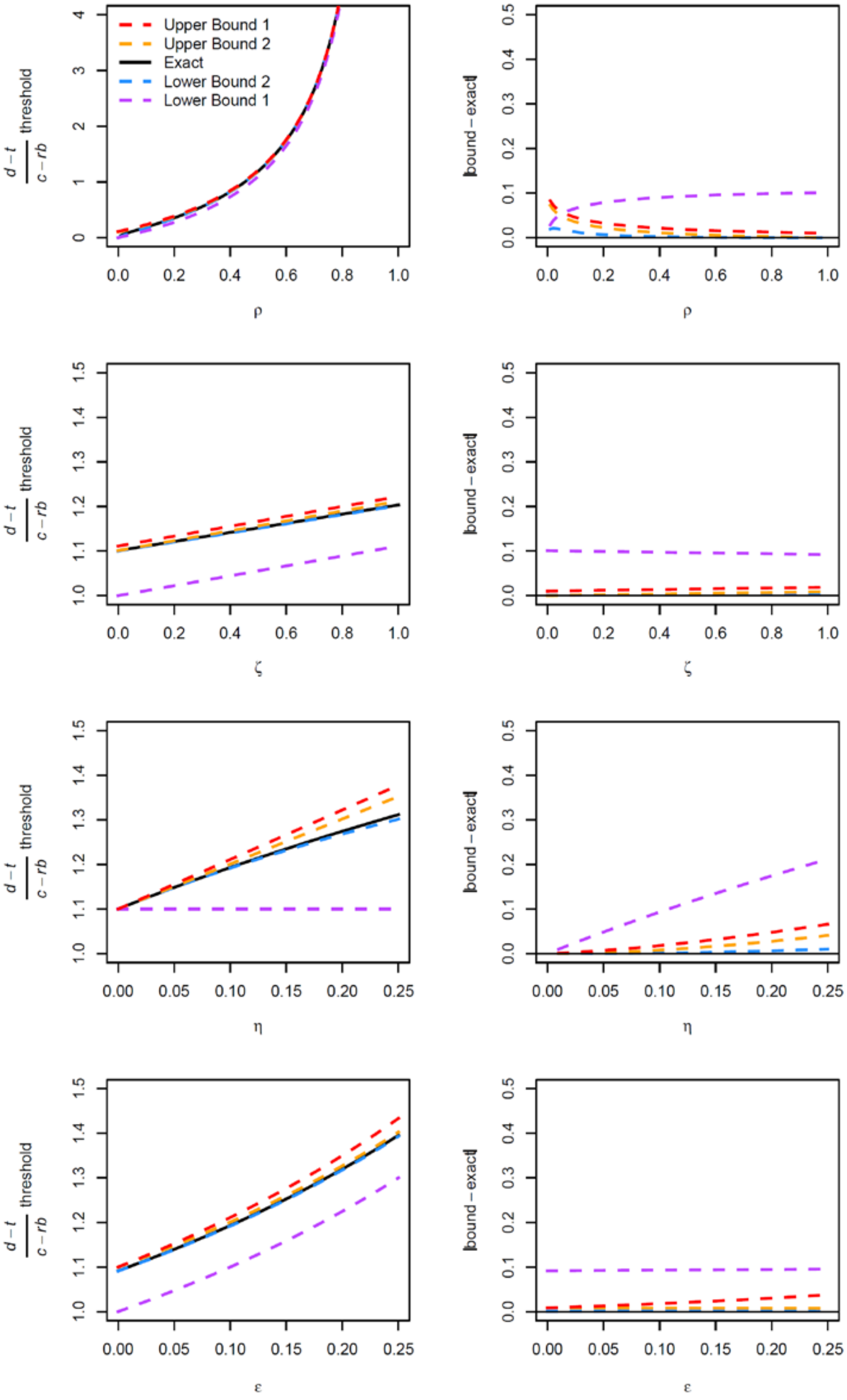


Figure S2. Minimum threshold values of *d/t* required for reputational cooperation to be stable against rare Mafiosos. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{9}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.

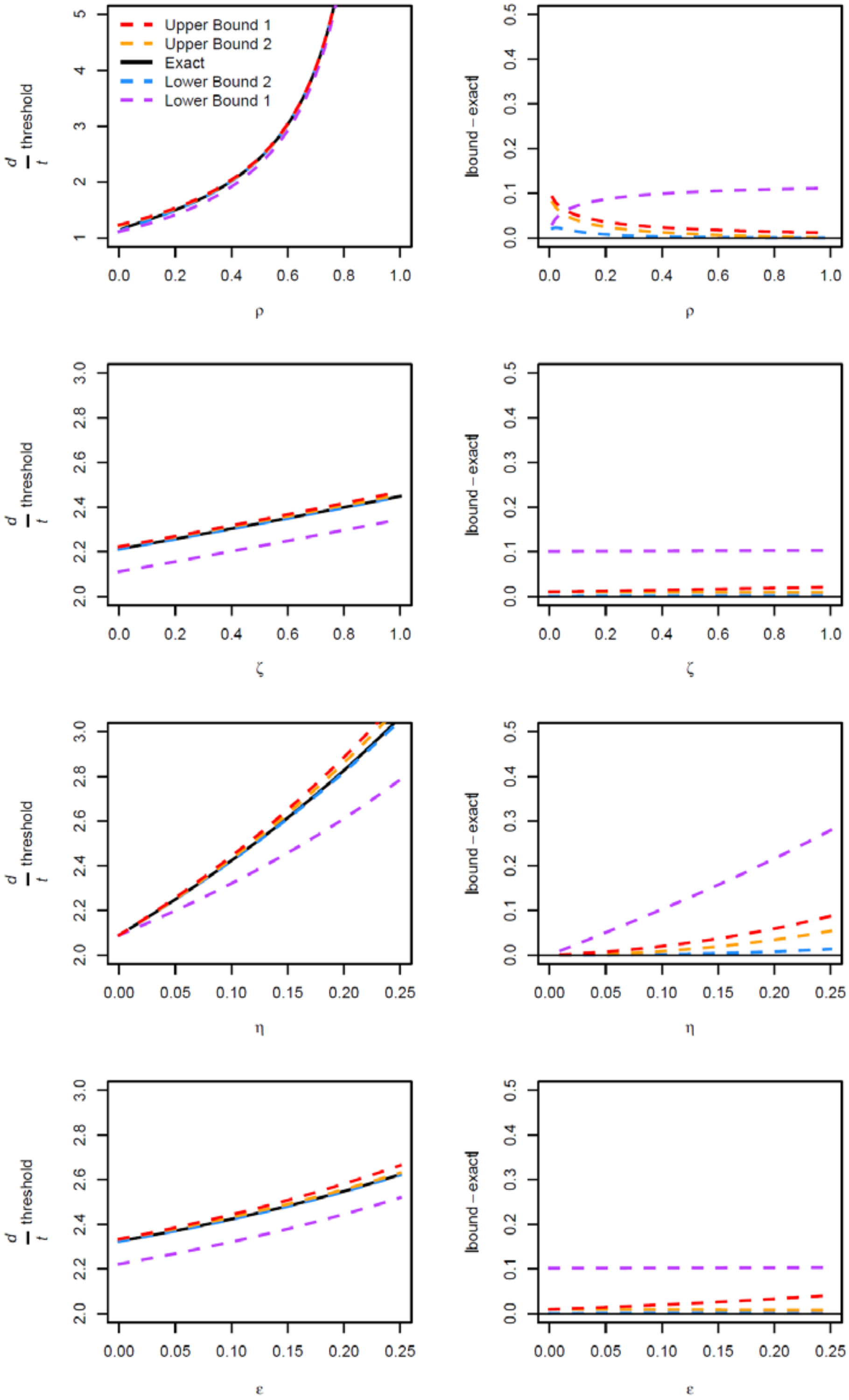
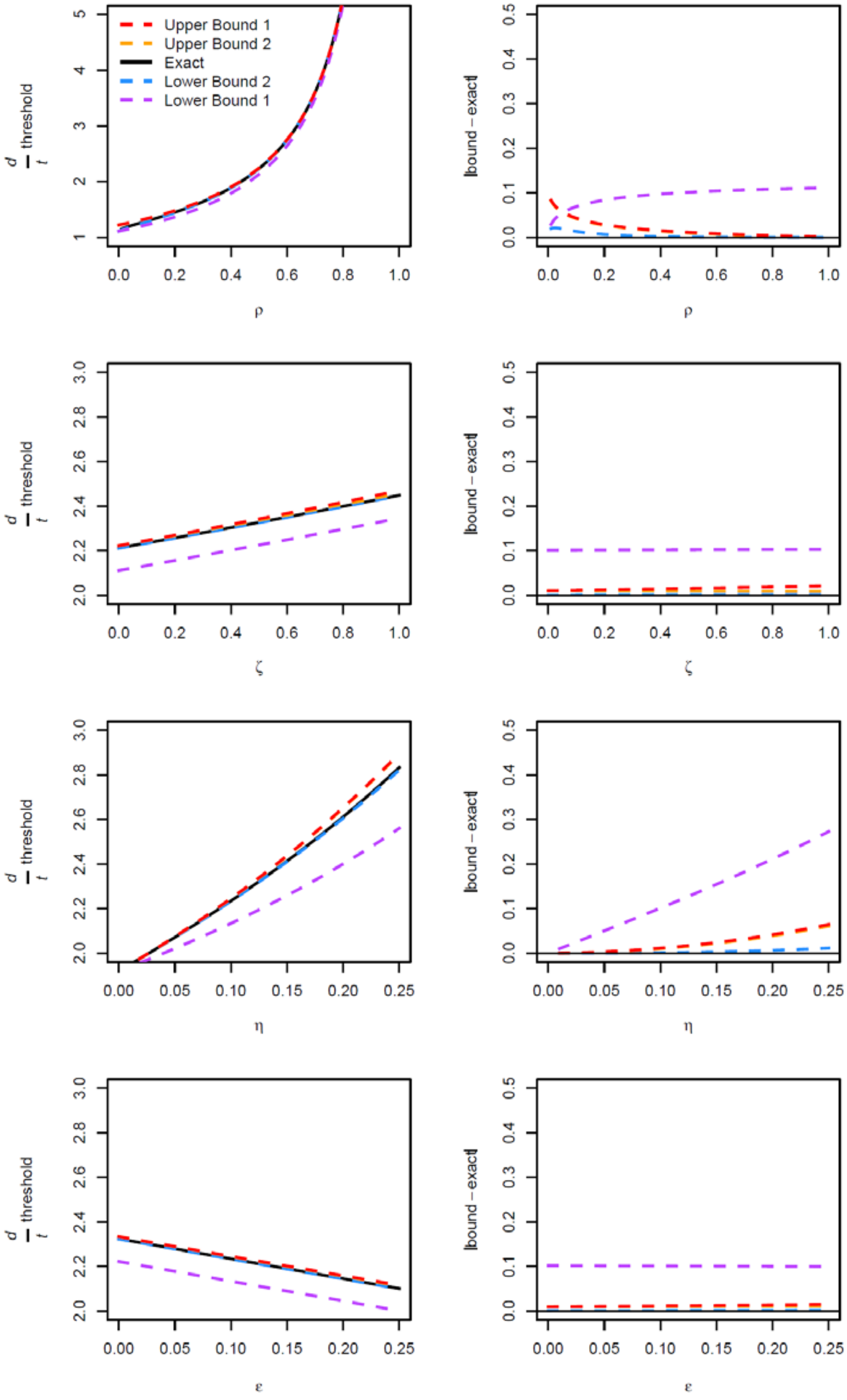


Figure S3. Minimum threshold values of *d/t* required for reputational cooperation to be stable against rare Mafiosos, with low ζ. Non-varied parameters are set at $\rho = \frac{1}{2}$, $\zeta = \frac{1}{10}$, and $\eta = \varepsilon = \frac{1}{10}$.



2. Basins of Attraction for Reputational Cooperation

Here we suppose that R types are not necessarily predominant, but constitute proportion p_R of the population. Then we determine the minimum initial value of p_R such that the population will converge to entirely R (i.e., the basin of attraction). In this section, we assume $\eta = 0$ for simplicity.

2a. Basin of Attraction against Defectors

In a population with fraction p_R Rs and $p_D = 1 - p_R$ Ds, and recognizing that $G_D = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\{t(p_R(1-G_R) + p_D) - d(p_R(1-G_R) + p_D)\}$$

$$= \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)(t-d)(1 - p_R G_R),$$

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t - d(p_R(1-G_D) + p_D)\}$$

$$= \rho\{b(1-r)p_R\} + (1-\rho)(t-d).$$

Then $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)(d-t)p_R G_R > 0$$

$$p_R > \frac{c - rb}{d - t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right).$$

Here, $G_R = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, so $\frac{1}{G_R} = 1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)$ and the condition becomes

$$p_R > \frac{c - rb}{d - t}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right).$$

Accordingly, p_R must meet some minimum threshold.

2b. Basin of Attraction against Stingy

In a population with fraction p_R Rs and $p_S = 1 - p_R$ Ss, and recognizing that $G_S = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\{t\big(p_R(1-G_R) + p_S(1-G_S)\big) - d(1-G_R)\}$$

$$= \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\big(t(1 - p_R G_R) - d(1-G_R)\big),$$

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)\{t\big(p_R(1-G_R) + p_S(1-G_S)\big) - d(1-G_S)\}$$

$$= \rho\{b(1-r)p_R\} + (1-\rho)(t(1 - p_R G_R) - d).$$

Then $w_R > w_S$ when

$$\rho(rb-c) + (1-\rho)dG_R > 0$$

$$\frac{d}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_R}\right).$$

Since $\frac{1}{G_R} = 1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)$, we need

$$1 > \frac{c-rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right).$$

which does not depend on p_R.

2c. Basin of Attraction against Mafiosos

In a population with fraction p_R *R*s and $p_M = 1 - p_R$ *M*s, we have

$$w_R = \rho\{b-c\} + (1-\rho)\{t(p_R(1-G_R) + p_M(1-G_M)) - d(p_R(1-G_R) + p_M)\}$$

$$= \rho\{b-c\} + (1-\rho)\{t((1-G_M) - p_R(G_R - G_M)) - d(1-p_RG_R)\},$$

$$w_M = \rho\{b-c\} + (1-\rho)\{t - d(p_R(1-G_M) + p_M)\}$$

$$= \rho\{b-c\} + (1-\rho)(t - d(1-p_RG_M)).$$

Then $w_R > w_M$ when

$$dp_R(G_R - G_M) > t(G_M + p_R(G_R - G_M))$$

$$p_R > \frac{t}{d-t}\left(\frac{G_M}{G_R - G_M}\right).$$

Here,

$$G_M = \frac{\rho(1-\varepsilon)}{\rho(1-\varepsilon(1-\zeta)) + (1-\rho)(p_RG_R + p_MG_M)},$$

and observing that $\frac{G_M}{G_R - G_M} = \frac{G_R}{G_R - G_M} - 1 = \frac{\rho}{1-\rho}\left(\frac{1-\varepsilon(1-\zeta)}{p_RG_R + p_MG_M}\right)$, we have

$$p_R > \frac{t}{d-t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1-\varepsilon(1-\zeta)}{p_RG_R + p_MG_M}\right).$$

However, since G_M is the solution to $(1-\rho)p_MG_M^2 + [\rho(1-\varepsilon(1-\zeta)) + (1-\rho)p_RG_R]G_M - \rho(1-\varepsilon) = 0$, this does not permit an easily interpretable analytical solution and must be numerically computed.

2d. Basin of Attraction against Defectors and Stingy

In a population with fraction p_R *R*s, p_S *S*s, and $p_D = 1 - p_R - p_S$ *D*s, and recognizing that $G_S = G_D = 0$, we have

$$w_R = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\big(t(1 - p_R G_R) - d(1 - (1 - p_D)G_R)\big)$$

$$w_S = \rho\{b(1-r)p_R\} + (1-\rho)(t(1 - p_R G_R) - d)$$

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)(t - d).$$

Then $w_R > w_S$ when

$$\rho(rb - c) + (1-\rho)(1-p_D)dG_R > 0$$

$$1 - p_D > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right)$$

$$1 - p_D > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right),$$

and $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)(d(1-p_D)G_R - tp_R G_R) > 0$$

$$(1 - p_D)d - tp_R > (c - rb)\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_R}\right)$$

$$1 - p_D - \left(\frac{t}{d}\right)p_R > \frac{c - rb}{d}\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right),$$

so p_D cannot be too large. Figure S4 shows the barycentric plot considering all three of these strategies.

2e. Basin of Attraction against Defectors and Mafiosos

In a population with fraction p_R *R*s, p_M *M*s, and $p_D = 1 - p_R - p_M$ *D*s, and recognizing that $G_D = 0$, we have

$$w_R = \rho\big\{b\big(r + (1-r)(1-p_D)\big) - c\big\} + (1-\rho)\big(t(1 - p_R G_R - p_M G_M) - d(1 - p_R G_R)\big)$$

$$w_M = \rho\big\{b\big(r + (1-r)(1-p_D)\big) - c\big\} + (1-\rho)\big(t - d(1 - p_R G_M)\big)$$

$$w_D = \rho\{b(1-r)(1-p_D)\} + (1-\rho)(t - d).$$

Then $w_R > w_M$ when

$$dp_R(G_R - G_M) - t(p_R G_R + p_M G_M) > 0$$

$$p_R[G_R(d-t) - G_M d] > p_M G_M t$$

$$p_R\left(\frac{G_R(d-t) - G_M d}{G_M t}\right) > p_M$$

$$p_R\left(\left(\frac{d-t}{t}\right)\left(\frac{G_R - G_M}{G_M}\right) - 1\right) > p_M.$$

This must be numerically computed; observe that

$$G_R = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$$

$$G_M = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)(p_R G_R + p_M G_M)}$$

$$\frac{G_R}{G_M} = \frac{\rho\big(1-\varepsilon(1-\zeta)\big) + (1-\rho)(p_R G_R + p_M G_M)}{\rho\big(1-\varepsilon(1-\zeta)\big)} = 1 + \left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right),$$

so we have

$$p_R\left(\left(\frac{d-t}{t}\right)\left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right) - 1\right) > p_M.$$

Also $w_R > w_D$ when

$$\rho(rb - c) + (1-\rho)\big(dp_R G_R - t(p_R G_R + p_M G_M)\big) > 0$$

$$p_R G_R(d-t) - p_M G_M t > (c - rb)\left(\frac{\rho}{1-\rho}\right)$$

$$p_R\left(\frac{G_R}{G_M}\right) > \frac{c-rb}{d-t}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{G_M}\right) + p_M\left(\frac{t}{d-t}\right).$$

$$p_R\left[1 + \left(\frac{1-\rho}{\rho}\right)\left(\frac{p_R G_R + p_M G_M}{1-\varepsilon(1-\zeta)}\right)\right]$$

$$> \frac{c-rb}{d-t}\left[\left(\frac{\rho}{1-\rho}\right)\left(1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right) + \frac{p_R G_R + p_M G_M}{1-\varepsilon}\right] + p_M\left(\frac{t}{d-t}\right).$$

Again, numerical solutions must be obtained. Figure S5 shows the barycentric plot considering all three of these strategies.

Figure S4. Barycentric plot reflecting evolution of *R*, *D*, and *S* strategies. Parameters are set at $\rho = \frac{1}{3}$, $b = 2$, $c = 1$, $r = \frac{1}{10}$, $d = 2$, $t = 1$, $\zeta = \frac{1}{2}$, and $\varepsilon = \frac{1}{10}$. The entire *S*-*D* axis consists of equilibria.

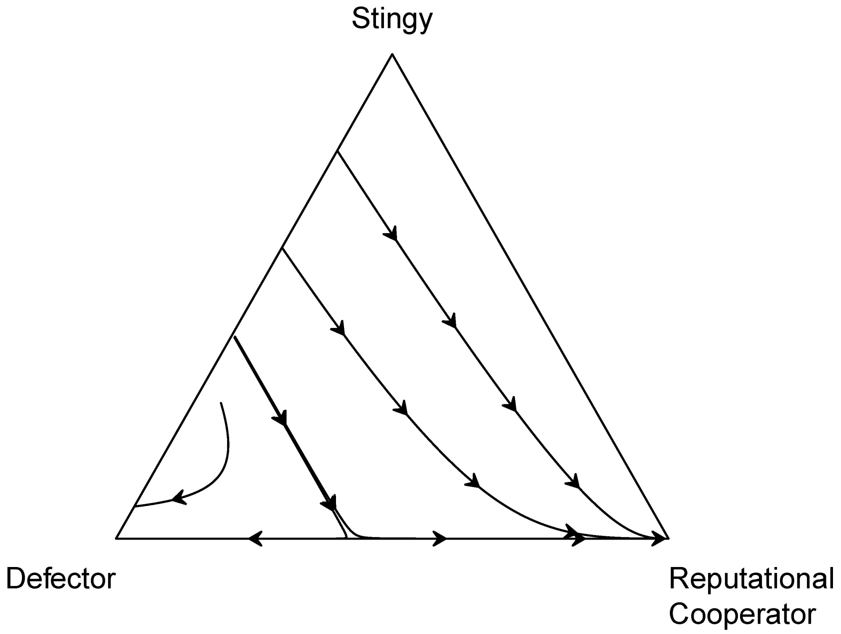
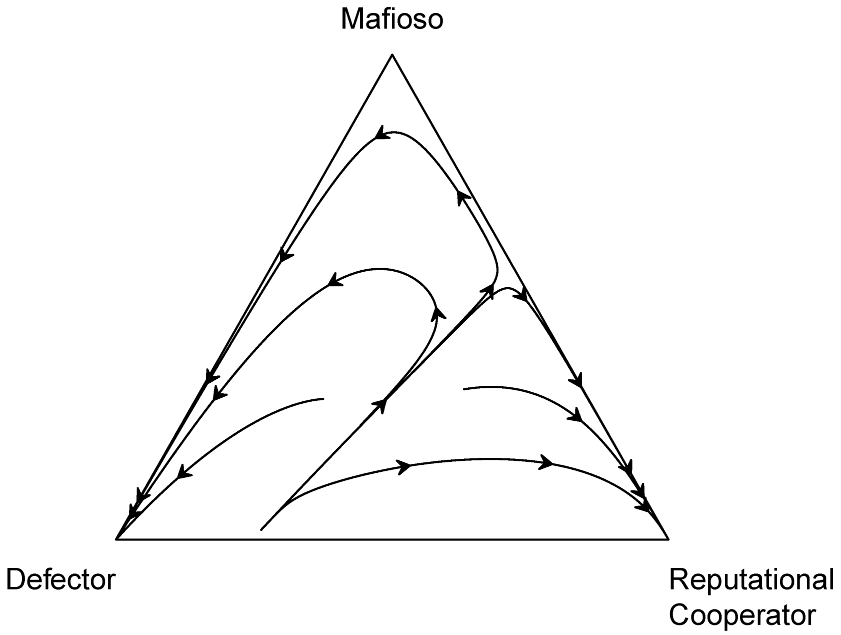


Figure S5. Barycentric plot reflecting evolution of *R*, *D*, and *M* strategies. Parameters are set at $\rho = \frac{1}{2}$, $b = 3$, $c = 1$, $r = \frac{1}{10}$, $d = 4$, $t = 1$, $\zeta = \frac{1}{2}$, and $\varepsilon = \frac{1}{10}$.



3. Exact Solution to Frequency of *R* in Good Standing

As laid out previously, in a population of R, $G_R = \frac{\rho(1-\varepsilon)}{\rho\big(1-\varepsilon(1-\zeta)\big)+(1-\rho)G_R\eta}$ is the solution to the quadratic equation $(1-\rho)\eta G_R^2 + \rho\big(1-\varepsilon(1-\zeta)\big)G_R - \rho(1-\varepsilon) = 0$, which is

$$G_R = \frac{-\rho\big(1-\varepsilon(1-\zeta)\big) + \sqrt{\rho^2\big(1-\varepsilon(1-\zeta)\big)^2 + 4\rho(1-\rho)(1-\varepsilon)\eta}}{2(1-\rho)\eta}.$$

Because the solution must be non-negative, only the positive result is accepted.

4. Obligate (Non-Reputational) Cooperation is Dominated by Reputational Cooperation

In a population of type j, the payoff of an individual who always cooperates (never exploits regardless of partner's standing, and always contribute) is

$$w_{OC} = \rho\big\{b\big(r + (1-r)p_{OC} + (1-r)p_jY_j\big) - c\big\} + (1-\rho)\{-d(1-G_{OC})\}.$$

The payoff of an individual who cooperates based on reputation (exploits only when partner is in bad standing, and always contribute) is

$$w_R = \rho\big\{b\big(r + (1-r)p_R + (1-r)p_jY_j\big) - c\big\} + (1-\rho)\{t(1-G_R) - d(1-G_R)\}.$$

Intuitively, the former behaves exactly like the latter except they relinquish the free takings from exploiting those in bad standing. Since this exploitation has no reputational consequences, both types have good standing identical fractions of the time (i.e., $G_{OC} = G_R$), and thus in any given population, reputational cooperators always perform at least as well as obligate cooperators.

5. Stability Conditions for Populations of Other Strategies

In the main text, stability conditions were presented for populations of reputational cooperators R resisting invasion by rare defectors D, stingy types S, and Mafiosos M. Here, we present stability conditions for populations comprising each one of the other strategies to flesh out the full space of combinations.

5a.i. Stability of Defector population against Reputational Cooperator invasion

D will earn the public benefit *b* only when meeting a rare *R* in the contribution game, which happens with probability $(1-r)p_R$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game due to virtually always meeting defectors. Thus,

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t-d\}.$$

R will earn *b* when meeting another *R* in the contribution game, which happens with probability $r+(1-r)p_R$, and they will always pay the cost *c*. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1-G_D$, and will always suffer damage *d*. So,

$$w_R = \rho\{b(r+(1-r)p_R)-c\} + (1-\rho)\{t(1-G_D)-d\}.$$

Here, $G_i = \frac{\rho Y_i(1-\varepsilon)}{\rho[Y_i(1-\varepsilon)(1-\zeta)+\zeta]+(1-\rho)G_D[X_i+(1-X_i)\eta]},.$ This entails $G_D = 0$, so *D* is stable when

$$w_D > w_R \Leftrightarrow rb < c.$$

5a.ii. Stability of Defector population against Mafioso invasion

D will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = \rho\{b(1-r)p_R\} + (1-\rho)\{t-d\}.$$

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r+(1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r+(1-r)p_R)-c\} + (1-\rho)\{t-d\}.$$

Since $G_D = 0$, *D* is stable when

$$w_D > w_M \Leftrightarrow rb < c.$$

5a.iii. Stability of Defector population against Stingy invasion

D will never earn the public benefit *b* in the contribution game, nor will they ever pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = (1 - \rho)\{t - d\}.$$

S will also never earn *b* or pay the cost *c* in the contribution game. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1 - G_D$, and will always suffer damage *d*.

$$w_S = (1 - \rho)\{t(1 - G_D) - d\}$$

Since $G_D = 0$, *S* has equal fitness to *D* here.

5b.i. Stability of Stingy population against Defector invasion

S will never earn the public benefit *b* in the contribution game, nor will they ever pay the cost *c*. They will take t when meeting someone in bad standing, which happens with probability $1 - G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1 - G_S$. Thus,

$$w_S = (1 - \rho)\{t(1 - G_S) - d(1 - G_S)\}.$$

D will also never earn *b* or pay the cost *c* in the contribution game. In the theft game, they will always take *t*, and suffer damage *d* when they are in bad standing.

$$w_S = (1 - \rho)\{t - d(1 - G_D)\}$$

Here, $G_i = \frac{\rho Y_i(1-\varepsilon)}{\rho[Y_i(1-\varepsilon)(1-\zeta)+\zeta]+(1-\rho)G_S[X_i+(1-X_i)\eta]}$, This entails $G_S = G_D = 0$, and so *D* has equal fitness to *S* here.

5b.ii. Stability of Stingy population against Reputational Cooperator invasion

S will earn the public benefit *b* only when meeting a rare *R* in the contribution game, which happens with probability $(1 - r)p_R$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens with probability $1 - G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1 - G_S$. Thus,

$$w_S = \rho\{b(1 - r)p_R\} + (1 - \rho)\{t(1 - G_S) - d(1 - G_S)\}.$$

R will earn *b* when meeting another *R* in the contribution game, which happens with probability $r + (1 - r)p_R$, and they will always pay the cost *c*. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1 - G_S$, and will suffer damage *d* when they are themselves in bad standing, occurring with probability $1 - G_R$. So,

$$w_R = \rho\{b(r + (1 - r)p_R) - c\} + (1 - \rho)\{t(1 - G_S) - d(1 - G_R)\}.$$

Note that $G_S = 0$ and $G_R = \frac{\rho(1-\varepsilon)}{\rho\left(1-\varepsilon(1-\zeta)\right)+(1-\rho)G_S\eta} = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, hence S is stable when

$$w_S > w_R \Leftrightarrow \rho(c - rb) > (1-\rho)d\left[\frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}\right]$$

$$\Leftrightarrow \frac{d}{c - rb} < \frac{\rho}{1-\rho}\left[1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right].$$

5b.iii. Stability of Stingy population against Mafioso invasion

S will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens with probability $1 - G_S$, and will suffer damage d in the theft game when they are themselves in bad standing, occurring with probability $1 - G_S$. Thus,

$$w_S = \rho\{b(1-r)p_M\} + (1-\rho)\{t(1-G_S) - d(1-G_S)\}.$$

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r + (1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t*, and will suffer damage *d* when they are themselves in bad standing, occurring with probability $1 - G_M$. So,

$$w_M = \rho\{b(r + (1-r)p_R) - c\} + (1-\rho)\{t - d(1-G_M)\}.$$

Note that $G_S = 0$ and $G_M = \frac{\rho(1-\varepsilon)}{\rho\left(1-\varepsilon(1-\zeta)\right)+(1-\rho)G_S} = \frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}$, hence *S* is stable when

$$w_S > w_M \Leftrightarrow \rho(c - rb) > (1-\rho)d\left[\frac{1-\varepsilon}{1-\varepsilon(1-\zeta)}\right]$$

$$\Leftrightarrow \frac{d}{c - rb} < \frac{\rho}{1-\rho}\left[1 + \zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right].$$

5c.i. Stability of Mafioso population against Defector invasion

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r + (1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r + (1-r)p_M) - c\} + (1-\rho)\{t - d\}.$$

D will earn the public benefit *b* when meeting an *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will always take t and suffer damage d in the theft game. Thus,

$$w_D = \rho\{b(1-r)p_M\} + (1-\rho)\{t-d\}.$$

Thus *M* is stable when

$$w_M > w_D \Leftrightarrow rb > c.$$

5c.ii. Stability of Mafioso population against Reputational Cooperator invasion

M will always earn *b* and pay *c* in the contribution game. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b-c\} + (1-\rho)\{t-d\}.$$

R will also always earn *b* and pay *c* in the contribution game. In the theft game, they will take *t* when the other player is in bad standing, which happens here with probability $1-G_M$, and will always suffer damage *d*. So,

$$w_R = \rho\{b-c\} + (1-\rho)\{t(1-G_M)-d\}.$$

Since $G_M > 0$, a population full of *M* is always stable against *R*.

5c.iii. Stability of Mafioso population against Stingy invasion

M will earn *b* when meeting another *M* in the contribution game, which happens with probability $r+(1-r)p_M$, and they will always pay the cost *c*. In the theft game, they will always take *t* and suffer damage *d*. So,

$$w_M = \rho\{b(r+(1-r)p_M)-c\} + (1-\rho)\{t-d\}.$$

S will earn the public benefit *b* only when meeting a rare *M* in the contribution game, which happens with probability $(1-r)p_M$, and will never pay the cost *c*. They will take t only when meeting someone in bad standing, which happens here with probability $1-G_M$, and will always suffer damage d in the theft game. Thus,

$$w_S = \rho\{b(1-r)p_M\} + (1-\rho)\{t(1-G_M)-d\}.$$

Recognizing that $\frac{1}{G_M} = \frac{\rho\left(1-\varepsilon(1-\zeta)\right)+(1-\rho)G_M}{\rho(1-\varepsilon)} = \left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{1-\rho}{\rho}\frac{G_M}{1-\varepsilon}$, *M* is stable when

$$w_M > w_S \Leftrightarrow \frac{t}{c-rb} > \frac{\rho}{1-\rho}\left(\frac{1}{G_M}\right)$$

$$\Leftrightarrow \frac{t}{c-rb} > \frac{\rho}{1-\rho}\left[1+\zeta\left(\frac{\varepsilon}{1-\varepsilon}\right)\right] + \frac{G_M}{1-\varepsilon}.$$