Phoenix From the Ashes: The Evolution of Mechanism Designers

David K. Levine[1]

Abstract

There is substantial empirical and experimental evidence that humans are instinctive mechanism designers. I develop an evolutionary model in which good mechanism designers have an evolutionary advantage over those who are not. In this model there are many types including mechanism designers. Some of these types are conformists who punish other types. Consequently it is difficult to leave a conformist state: it requires a large invasion of mutants to reduce punishment enough to acquire an advantage over the incumbents. However: if a shock causes a reduction in population a much smaller invasion is required, so it is much easier to leave. This favors mechanism designers because they are able to provide public goods that mitigate the chances of catastrophe. This evolutionary theory implies that social welfare should not be measured by happiness but by resilience.

Email address: david@dklevine.com (David K. Levine)
[1]Department of Economics, EUI and WUSTL
Acknowledgements: First version: September 10, 2022. I would like to thank Rohan Dutta, John Mair, Andrea Mattozzi and Salvatore Modica. I gratefully acknowledge support from the MIUR PRIN 2017 n. 2017H5KPLL_01.

> Our founding fathers, they were great mechanism designers. *Edward C. Prescott, private communication to the author.*

1. Introduction

There is substantial empirical and experimental evidence that humans are instinctive mechanism designers. I do not mean this in the sense that we actively employ the modern mathematical tools of mechanism design, but rather that we informally understand and try to create incentives for ourselves and others to reach social objectives. If this is the case then it should be that good mechanism designers have an evolutionary advantage over those who are not. My objective in this paper is to develop such a theory. In doing so I also clarify what is the social objective function that is favored by evolutionary forces.

The tool I will use to study the evolution of mechanism designers is that of modern stochastic evolutionary theory and stochastic stability as developed by Young (1993), Kandori, Mailath and Rob (1993) Ellison (2000), Levine and Modica (2016a) among others. In that setting I consider the evolution of types who vary in their ability to design mechanisms. Individual fitness in this model depends upon the private costs born by individuals in pursuing social goals, but in addition types can provide incentives by punishing other types. Conformists are those whose punishments make it strictly incentive compatible to bear the private costs associated with being that type. As is the case in practice, I consider that the population is subject to random shocks. The central hypothesis is that the public goods produced by effective mechanism designers mitigate these shocks.

The crucial idea is this. Because conformists punish other types it is difficult to leave a conformist state: it requires a large invasion of mutants to reduce punishment enough to acquire an advantage over the incumbents. Writers in the evolutionary literature such as Ellison (2000) have long been aware of this. However: if a shock causes a reduction in population then a much smaller invasion is required, so it is much easier to leave. This favors mechanism designers because they are able to provide public goods that mitigate the chances of catastrophe. Hence, like the Phoenix rising from the ashes, mechanism designers arise from the catastrophes of their predecessors. However, because they provide public goods, they are less likely to endure catastrophes of their own.

In addition to explaining the evolution of humans as mechanism designers the theory has two other implications. First, the types that are often studied by economists are selfish types and altruistic types, the former minimizing private cost, and the latter contributing to the public good. As neither of these types provides any incentives for others to do the same neither has much survival value in the presence of conformist types. The model here confirms this: we should rarely see either selfish or altruistic types. Second, the social objective function that arises from the evolutionary analysis is an index of the probability of avoiding catastrophe. That is, social welfare as favored by evolution is not about making us happy but about protecting us from catastrophes. In current terminology, social welfare should be measured by resilience.

I must emphasize that the theory does not predict that mechanism designers will always dominate. First, it can be that catastrophes are not so important - in this case I show that it is "nasty types" who minimize private cost yet demand conformity through punishment who dominate. Second, the theory is a dynamic theory in which, while mechanism designers may dominate most of the time, there are reversals and for shorter periods of time types with dysfunctional social norms may thrive.

Before turning to the theory I want to discuss the evidence that we are indeed mechanism designers. Although they do not use this terminology this is the thrust of the ideas of Coase (1960) and Ostrom (1990). Coase (1960) documents through a series of case studies how groups of people effectively overcome the free-rider problem by designing incentive schemes. Ostrom (1990) also documents the provision of public goods through the use of incentives and shows how these incentives take the form of punishments such as ostracism for those who fail to do their share. Townsend (1994) studies Thai villages and shows that an explicit mechanism design model well describes how insurance against crop failures is designed. In the context of political parties Levine and Mattozzi (2020) show how voter turnout is explained by parties designing social mechanisms to overcome the free-rider problem and Della Vigna et al (2014) discusses the punishments involved. Levine and Modica (2017) similarly discuss the role of social incentives in lobbying groups. Dutta, Levine and Modica (2021) give a series of case studies of effective mechanisms developed by groups to overcome free-riding. One important class of examples are cartels - there is an extensive empirical literature showing how cartels operate effectively to provide incentives for their members - see for example Levenstein and Suslow (2006). I should also mention that the basis of the modern contracting literature and theory of the firm (see, for example, Grossman and Hart (1986)) is based on the idea that people are good at designing mechanisms.

There is also experimental evidence that people are good at mechanism design in the laboratory. Fehr and Gachter (2000a) and Peysakhovich and Rand (2014), among others, show how given the opportunity to create incentives to provide public goods in the laboratory people do so. The experimental literature on reciprocal altruism (see for example Fehr and Gachter (2000b)) show much the same. Dutta, Levine and Modica (2022) show that many of these experimental results can be well explained as the solution of a mechanism design problem.

The theory also indicates the strong survival value of conformists who enforce their social norms by punishing other types. As indicated there are circumstances in which dysfunctional conformists can thrive for short or even long periods of time. Evidence that this is the case has been extensively developed on the literature on conformity and identity: see in particular Akerlof and Kranton (2005).

The theory also proposes a specific mechanism through which mechanism designers survive and thrive: it is through catastrophes that reduce population and make it easier for new types to enter. Human catastrophes can take place in many ways, they may be ecological as in Diamond (2005), they may involve

losses in war as in Levine and Modica (2013), or they may simply involve civil strife, arising, for example, out of inequality. In this respect the insurance schemes studied by Townsend (1994) may be seen to reduce the chances of catastrophe.

Catastrophes that reduce population are not uncommon. Homo Sapiens has been around for around 70,000 years and our history includes the severe glaciation of the Younger Dryas period that ended about 12,000 years ago (Peteet (1995)). While worldwide population declines of the level caused by the Younger Dryas have not occured since, local population declines due to catastrophes are frequent in human history: "[The population of Rome] declined from about a million persons to 30,000 between the second and sixth centuries AD" (Twine (1992)) being a particularly sharp example. These declines, it should be emphasized, are frequently through out-migration from hard hit areas, as well as from deaths. For example, the depopulation of Ireland during the Great Famine of the 1840s is thought to be split roughly equally between excess deaths and out-migration (Ross (2002)).

Finally, the idea that population reductions make it easier for new entrants to thrive is strongly supported in the biological record. A key fact (see for example Jablonski (2001)) that mass extinctions caused by events such as asteroid strikes not only reduce the population of existing species, but subsequently lead to a great increase in the diversity of species. This too is the phoenix rising from the ashes.

Related Literature

While the idea in this paper differs from the existing literature by considering variable population driven by catastrophes the idea that evolution favors functional institutions is scarcely new. This is the thrust of the work of Bowles and Choi (2013) and Choi and Bowles (2007) in their study of the emergence of institutions in the post-Dryas period. In the repeated games literature there is a long list of results on evolution leading to efficient outcomes supported by punishment strategies, particularly in the prisoners' dilemma game: Axelrod and Hamilton (1981), Binmore and Samuelson (1992), Johnson, Levine and Pesendorfer (2001), Dal Bo and Pujals (2015), and Juang and Sabourian (2021) are but a few examples.

The theory about evolution through conflict that leads to hegemony as in Levine and Modica (2012), Levine and Modica (2013), Levine and Modica (2016a) Levine and Modica (2022) , Bilancini, Boncinelli and Marcos-Prieto (2022)] has a similar flavor to the results here. In particular the evolution of the ability to withstand "outside pressure" is similar to resilience in this paper. However, in those papers are reduced form in the way in which mechanism operate and consider neither punishment nor conformism.

In the context of evolution favoring efficiency I should mention the early work of Winter (1971) who showed how the survival of more profitable firms leads to efficient competitive equilibrium. In turn, there is a literature that studying the evolution of altruism, without, however, being supported by punishment and conformity. Positive assortative matching as in Alger and Weibull (2013) and

voluntary migration as in Ely (2003), for example, give reason for the survival of altruism. I should indicate while in the presence of conformist types pure altruism has no survival value the evidence it exists in practice is overwhelming, and while that same evidence suggests it is quantitatively relatively modest, it serves in an important way as "grease on the wheels of mechanism design" as made explicit in Dutta, Levine and Modica (2022).

2. The Model

In each of $t = 1, 2, \ldots$ periods there is a finite group whose members belong to a finite collection of types $\mathcal{T}$. Each type $\tau \in \mathcal{T}$ incurs a private cost $c(\tau)[\varphi] \geq 0$ where $\varphi \in [0, 1]$ is the population fraction of that type, provides resilience $w(\tau) \geq 0$ which is a public good, and imposes a utility punishment $P(\tau) \geq 0$ on other types.

Private cost should satisfy several assumptions, which I examine below in the context of an example. Cost $c(\tau)[\varphi]$ is assumed to be differentiable in φ and to satisfy $c'(\tau)[\varphi] \leq P(\tau)$. In addition if $c(\tau)[\varphi] > 0$ then $c(\tau)[0] > 0$ and $c(\tau)[1] \neq P(\tau)$. In case $c(\tau)[\varphi]$ is independent of φ for all τ I say that *cost is flat*.

This formulation allows for a wide variety of types. A *selfish type* or *homo economicus* has zero cost, provides no resilience and does not punish: $c(\tau)[\varphi] = 0, w(\tau) = 0, P(\tau) = 0$. A *nasty type* also has zero cost, provides no resilience, but does punish: $c(\tau)[\varphi] = 0, w(\tau) = 0, P(\tau) > 0$. An *altruistic type* provides resilience, but does not punish: $c(\tau)[\varphi] =?, w(\tau) > 0, P(\tau) = 0$. More broadly, I say that a type τ is *conformist* if for any other type σ we have $c(\tau)[1] < c(\sigma)[0] + P(\tau)$, which will mean that the punishment is sufficient to induce conformity. A nasty type, at least is conformist, and I assume that there is at least one type that is nasty.

In period t there are n_t^τ members of type τ and the vector n_t constitutes the *state*. A number $m^\tau > 0$ of each type are *residual* and these populations do not change over time, while the remaining *variable* populations of $n_t^\tau - m_t^\tau$ evolve according to an evolutionary process that is Markov on the state space and will be described. I denote by $M = \sum_\tau m^\tau$ the residual population and by $N_t = \sum_\tau n_t^\tau$ the total population. It is convenient as well to record the population fractions $\phi_t^\tau = n_t^\tau / N_t$. The total private cost including punishment by other types incurred by type τ at t can than be computed as

$$C^\tau(\phi_t) = c(\tau)[\phi_t^\tau] + \sum_{\sigma \neq \tau} \phi_t^\sigma P(\sigma).$$

Two forces determine the evolution of the variable population: catastrophes and reproduction. Catastrophes are governed by $v > 0$ which is a measure of the importance of the public good in preventing catastrophes, by $\underline{N} \geq M$ which is the population after a catastrope, and by $\overline{N} > \underline{N}$ which is the largest sustainable population. I study the case where catastrophes and reproduction of the unfit are rare events and $\epsilon > 0$ is a measure of how rare these events

are. Specifically, it will be convenient to specify probabilities in terms of their resistances: events at time t will have the form

$$h(n_t) = H(\phi_t, N_t/\overline{N}, v)\epsilon^{r(\phi_t, N_t/\overline{N}, v)}$$

where $0 < H(\phi_t, N_t/\overline{N}, v)$ is uniformly bounded away from zero, and the resistance $r(\phi_t, N_t/\overline{N}, v) \geq 0$.

I assume that catastrophes are unlikely and the chance of catastrophe is mitigated by resilience which per capita is $W_t = \sum_\tau \phi_t^\tau w(\tau)$. Resistance to a catastrophe is given by $r_\kappa(vW_t, N_t/\overline{N})$ which is strictly positive, strictly increasing and has slope bounded above zero in the first argument, that is, for fixed N_t, increasing resilience increases resistance to catastrophe. I am agnostic as to the dependence on $N_t/\overline{N}$: if catastrophes are due to invasion by neighbors then a large population is likely to reduce their probability, while if they are due to ecological collapse a large population is likely to increase their probability. When a catastrophe takes place $N_t - \underline{N}$ members are randomly removed from the population so that it falls to $\underline{N}$.

If there is no catastrophe then evolution takes place according to a Darwinian dynamic. As birth rates should be proportional to the population, for some $0 < \beta < 1$ with probability βN_t a type is chosen randomly to reproduce. The chances that this type is able to reproduce successfully depends upon the private cost that they bear relative to other types: that is, private cost measures individual lack of fitness. Letting z be a $\mathcal{T}-1$ vector of non-negative pairs ϕ^τ, C^τ resistance to successful reproduction for types τ is determined by a common function $r_p(C^\tau, z)$ with $r_t^\tau(\phi_t, N_t/\overline{N}, v) = r_p\left(C^\tau(\phi_t), (\phi_t^\sigma, C^\sigma(\phi_t))_{\sigma\neq\tau}\right)$. I assume that r_p is weakly increasing in C^τ and anonymous with respect to z in the sense that it is invariant to permuting the pairs. I also assume that the resistance to reproduction $r_p(C^\tau, z)$ is zero if $C^\tau \leq \min C^\sigma$, that is, for the least cost type. Denote the average cost of other types by $\overline{z} = \sum_{\sigma\neq\tau} \phi^\sigma C^\sigma / \sum_{\sigma\neq\tau} \phi^\sigma$. For above average cost types, who I refer to as *unfit*, that is, $C^\tau > \overline{z}$, resistance $r_p(C^\tau, z)$ is at least one but no greater than $\overline{r} \geq 1$. In particular least cost types reproduce much more rapidly than unfit types.

Finally, if reproduction leads to a population that is greater than $\overline{N}$ then one member is randomly chosen to be removed from the variable population so that the population never exceeds $\overline{N}$.

Stochastic Stability

I call the state in which $n_t^\tau = \overline{N} - M$, that is, the population is at a maximum and the variable population consists entirely of type τ, the *τ-state*. If τ is conformist the τ state is a *conformist state*. I say that a type τ is a *mechanism designer* if the type is conformist and for any other conformist σ we have $W(\tau) \geq W(\sigma)$. That is, a mechanism designer maximizes welfare as measured by resilience. If τ is a mechanism designer then the τ-state is a *mechanism design state*.

For $\epsilon > 0$ there is a positive probability path from any state to any other state, and because the resistance to no reproduction is zero ($\beta < 1$) there are

no deterministic cycles. From Young (1993) this implies that the system has a unique ergodic distribution with a unique limit as $\epsilon \to 0$. This unique limit is an ergodic distribution of the system with $\epsilon = 0$. Following the usual definition I say that a state is *stochastically stable* if it has positive probability in the limiting ergodic distribution. The implication of stochastic stability is that when ϵ is small "most of the time" we will observe one of these stochastically stable states. I will illustrate this later by Monte Carlo simulation.

My goal is to characterize stochastically stable states. To minimize notation and maximize readability I will abbreviate "there exists an N such that for $\overline{N} > N$" as "if $\overline{N}$ is sufficiently large."

3. Stochastic Stability of Conformists and Mechanism Designers

The main result of the paper is this:

Theorem 3.1. *For given $\underline{N}$*

(i) if $\overline{N}$ is sufficiently large only conformist states are stochastically stable

(ii) there exists a $\hat{v}$ and for any $v > \hat{v}$ if $\overline{N}$ is sufficiently large only mechanism design states are stochastically stable.

Proof Outline

For $\epsilon = 0$ since there is a positive probability of reproduction and no deaths when $N_t < \overline{N}$ it will be no surprise that states with $N_t < \overline{N}$ are transient. When $N_t = \overline{N}$ if $\overline{N}$ is sufficiently large the residual population matters little and reproductive success dominates so conformist states, by discouraging deviations, are absorbing, while all other states are transient. Hence ergodic distributions place weight only on conformist states so only these can be stochastically stable.

If v is sufficiently large then the public good is important for preventing castrophes. If $\overline{N}$ is sufficiently large then, in addition, it is very difficult to move between conformist states when $N_t = \overline{N}$ because it requires a great deal of reproduction of the unfit. The key idea is that once the population has fallen only few reproductions by the unfit are enough to tilt the system to a new state. Since it is relatively easy to move between conformist states once the population has fallen stochastic stability requires a high level of resilience.

Before turning to the details of the proof let I want to underline the assumptions that lead to the result.

Private Cost

I imagine that producing resilience and issuing punishments are costly and that there is an underlying common cost function $\xi(w, P)[\varphi]$ and that only *feasible types* with $c(\tau)[\varphi] \geq \xi(w(\tau), P(\tau))[\varphi]$ are possible. I now want to illustrate the assumptions by developing a simple model of such a common cost function.

There are two sources of cost: the first is the cost of producing resilience. As the units of resilience are arbitrary, we may assume that they are measured so that resilience has a constant marginal cost of $\alpha > 0$. It is also costly to

issue punishments. As punishment may be costly to the punisher as well as the punished cost of type τ will in general depend upon the population fraction of the type φ. These costs may arise from two sources: identification of types may be imperfect so that there can be a chance of "accidentally" punishing the own type as in Levine and Modica (2016b) or Levine and Mattozzi (2020). In addition issuing punishments may be costly to the punisher as well as the punished. I have also assumed the existence of a nasty type that punishes without cost. This is consistent with increasing marginal cost of punishment, but is stronger.

To develop a model, I will first tell a story. A signal noisy signal is received by type τ about whether members of the population are of type τ or not. For simplicity we may assume that other types give the "bad" signal of being different with probability one, while same type gives the bad signal with probability π. Punishments are issued against members who have bad signals and may be one of two kinds. One possibility is to costless issue an insult against other types "if you are not type τ you might as well be a frog" or something of that sort. This is offensive to other types but not to members who are type τ. On the other hand, there is a limit to how offensive these types of remarks are, so the utility loss to other types is at most $\underline{P} > 0$. To punish more than $\underline{P}$ more direct action must be taken, and this punishment hurts all members regardless of whether they are type τ and in addition is costly to the punisher.

Define the excess punishment, then, as $Q(P) = \max\{0, P - \underline{P}\}$. Letting $\psi > 0$, the direct cost of issuing such a punishment is ψQ and the group issues $(1-\varphi) + \varphi\pi$ of these punishments. In addition group τ members receive an additional $\varphi\pi Q$ punishments "by accident" due to inaccurate signals. Hence the common cost function is given by

$$\xi(w, P)[\varphi] = \alpha w + \varphi\pi(1+\psi)Q(P) + (1-\varphi)\psi Q(P). \tag{3.1}$$

I refer to this as the *linear case*.

Feasible costs are then given by $c(\tau)[\varphi] \geq \xi(w(\tau), P(\tau))[\varphi]$. This allows the possibility of incurring additional costs by engaging in activities that neither produce the public good nor help with punishment: for example, building monuments to the gods that to not reduce the chances of catastrophe. Clearly efficient costs for which $c(\tau)[\varphi] = \xi(w(\tau), P(\tau))[\varphi]$ are of particular interest. When do these satisfy the assumptions of private cost? The answer is: almost always.

Theorem 3.2. *$\xi(w, P)[\varphi]$ is differentiable in φ and satisfies $\xi'(w, P)[\varphi] \leq P$. If $\xi(w, P)[\varphi] > 0$ for some φ then $\xi(w, P)[0] > 0$ and for given w, ψ and generic P, π we have $\xi(w, P)[1] \neq P$.*

Proof. Clearly $\xi(w, P)[\varphi]$ is differentiable in φ, it is, in fact linear. The assumption that $\xi'(w, P)[\varphi] \leq P$ says that $(\pi(1+\psi) - \psi)\max\{0, 1 - \underline{P}/P\} \leq 1$. This holds for all P if and only if $\pi(1+\psi) - \psi \leq 1$, which, since $\pi \leq 1$ means it holds always.

If $\xi(w, P)[\varphi] > 0$ then it must be that either $w > 0$ or $Q(P) > 0$. Since

$\xi(w,P)[0] = \alpha w(\tau) + \psi Q(P)$ this also is strictly positive. If $P < \underline{P}$ then $\xi(w,P)[1] = P$ only in the non-generic case in which $P = \alpha w$. If $P \geq \underline{P}$ then $\xi(w,P)[1] = P$ only in the non-generic case in which $\alpha w + \pi(1+\psi)(P - \underline{P}) = P$. □

This simple model allows a wide variety of types. A nasty type τ type that has no cost, produces no public good, but does punish with $0 < P(\tau) \leq \underline{P}$ is feasible as required: this is the purpose of introducing $\underline{P} > 0$. Selfish types $c(\tau)[\varphi] = w(\tau) = P(\tau) = 0$ are feasible as are altruistic types $c(\tau)[\varphi] \geq w(\tau) > 0, P(\tau) = 0$. Finally, for $0 < \mu < 1$ we have $c(\tau)[\varphi] = \mu\underline{P}, w(\tau) = \mu\underline{P}/\alpha, P(\tau) = \underline{P}$ is a conformist that produces positive resilience.

A useful special case is $\pi = \psi/(1+\psi)$ in which case the cost function is $\xi(w,P)[\varphi] = \alpha w + \psi Q$ independent of φ. This is consistent with flat cost.

Efficient Production

Mechanism design types are not all equal. In particular, it seems as though by choosing very large punishments, with associated large costs, since the cost to opponents is greater than the cost to the group, these large punishments might be "more stable" than smaller punishments. Is it true that evolution favors large punishments? Not if resources for public good production and punishment are limited: the resources needed for large punishments would leave little for the production of resilience. To show this specifically, consider the linear case with $\pi = \psi/(1+\psi)$ and flat cost, so

$$c(\tau)[\varphi] \geq \alpha w(\tau) + \psi Q(\tau).$$

Assume, moreover, that resources available are limited so that $c(\tau) \leq \overline{c}$ where $\overline{c} > 0$.

A key assumption is that types do not have punishments that lead to indifference. In practice quantities such as punishment are not infinitely divisible and neither information nor calculation so precise that it would be possible to calibrate a mechanism to give exact indifference. As a simple means of ruling this out, take $0 < \mu < 1$ and we assume that feasible punishments satisfy either $P(\tau) \leq \mu c(\tau)$ or $P(\tau) \geq (1/\mu)c(\tau)$, that is, cannot be exactly equal to cost. We refer to the latter condition as μ-strict incentive compatibility.

It is natural to inquire into the best feasible type: a type that maximizes output of the public good subject to incentive compatibility. If such a type is present in the population then these type are mechanism designers, and under the conditions of Theorem 3.1 only these types can be stochastically stable.

It is natural to think of $\underline{P}$ to be small and μ as close to 1. Moreover, to be useful, punishment should be more costly for those punished than those punishing, that is, $\psi < 1$. If this is the case then $\overline{c} > \mu\underline{P}$ and $\psi/\mu < 1$. Suppose this is true. As it is possible to produce more than μP it must be that for the best feasible type τ the resource constraint binds, $c(\tau) = \overline{c}$, that $P(\tau) > \underline{P}$ and that production is efficient in the sense that there is no excess cost $c(\tau)[\varphi] = \xi(w(\tau), P(\tau))[\varphi] = w(\tau) + \psi(P(\tau) - \underline{P})$ and no excess punishment

$P(\tau) = (1/\mu)c(\tau)[\varphi]$ as either would waste resources that could be used to produce more resilience. This easily enables us to find the best feasible type from

$$\overline{c} = w(\tau) + \alpha(\overline{c}/\mu - \underline{P})$$

as $c(\tau)[\varphi] = \overline{c}, w(\tau) = (1 - \alpha/\mu)\overline{c} + \alpha\underline{P}, P(\tau) = \overline{c}/\mu$. In particular punishment far from being as large as possible should be the smallest consistent with μ-strict incentive compatibility.

4. Proof Details

In detail, the proof of the main theorem 3.1 proceeds in a series of steps which I present as Lemmas and Corollaries.

Lemma 4.1. *If state n_t is not a τ-state and type τ has least cost C^τ then there is a zero resistance transition to a state in which type τ has reproduced and still has least cost. In particular there is a zero resistance path to the τ-state.*

Proof. The least cost type has zero resistance to reproducing. The cost difference between τ and $\sigma \neq \tau$ is

$$C^\tau(\phi_t) - C^\sigma(\phi_t) = c(\tau)[\phi_t^\tau] - c(\sigma)[\phi_t^\sigma] + \sum_{\omega \neq \tau} \phi_t^\omega P(\omega) - \sum_{\omega \neq \sigma} \phi_t^\omega P(\omega)$$

$$= c(\tau)[\phi_t^\tau] - c(\sigma)[\phi_t^\sigma] + \phi_t^\sigma P(\sigma) - \phi_t^\tau P(\tau).$$

A reproduction by τ weakly increases ϕ^τ and weakly decreases ϕ^σ regardless of whether or not some type $\gamma \neq \tau$ is removed from the population and whether or not $\gamma = \sigma$. It follows from $c'(\gamma)[\varphi] \leq P(\gamma)$ for $\gamma = \tau, \sigma$ that for all $\sigma \neq \tau$ the cost difference $C^\tau(\phi_t) - C^\sigma(\phi_t)$ is weakly decreased, so τ remains least cost. □

Lemma 4.2. *For given M if $\overline{N}$ is sufficiently large and if τ is not conformist then in the τ-state there is zero resistance to reaching a conformist state.*

Proof. By Lemma 4.1 it suffices to show that if τ is not conformist if $\overline{N}$ is sufficiently large there is a conformist γ with least cost in the τ-state.

For type τ cost is $C^\tau(\phi_t) = c(\tau)[\phi_t^\tau] - \sum_{\sigma \neq \tau}(m^\sigma/N_t)P(\sigma)$. As τ is not conformist there is a γ with $c(\tau)[1] \geq c(\gamma)[0] - P(\tau)$. Hence the same inequality holds if $c(\gamma)[\varphi] \equiv 0$ and there is at least one such γ, a nasty type, that is conformist by assumption. Let ϕ_0 denote the distribution of types in the τ-state: for $\overline{N}$ sufficiently large by continuity of c it must be that $c(\tau)[\phi_0^\tau] \geq -P(\tau)$. Let γ be chosen as a minimizer of $\sum_{\sigma \neq \gamma} m^\sigma P(\sigma)$ subject to $c(\gamma)[\varphi] \equiv 0$ and notice that it must be that $P(\gamma) > 0$ so that γ is conformist.

For any $\omega \neq \tau$ we have

$$C^\omega(\phi_0) = c(\omega)[\phi_0^\omega] - (1/\overline{N}) \sum_{\sigma \neq \omega} m^\sigma P(\sigma) - (1 - M/\overline{N})P(\tau)$$

$$= c(\omega)[\phi_0^\omega] - P(\tau) - (1/\overline{N})\left(\sum_{\sigma\neq\omega} m^\sigma P(\sigma) - MP(\tau)\right).$$

If $c(\omega)[\varphi] \equiv 0$ then by construction $C^\gamma(\phi_0) \leq C^\omega(\phi_0)$. Otherwise, by assumption $c(\omega)[0] > 0$. Hence for $\overline{N}$ sufficiently large by continuity of c we have $c(\omega)[\phi_0^\omega] > 0$ and since $c(\gamma)[\phi_0^\gamma] = 0$ we have $c(\gamma)[\phi_0^\gamma] - P(\tau) < c(\omega)[\phi_0^\omega] - P(\tau)$. Hence for $\overline{N}$ is sufficiently large $C^\gamma(\phi_t) < C^\omega(\phi_t)$.

Finally I must compare $C^\gamma(\phi_t)$ with $C^\tau(\phi_t)$. If $c(\tau)[\phi_0^\tau] = 0$ then $c(\tau)[\varphi] \equiv 0$ so since τ is not conformist it must be that $P(\tau) = 0$ and by construction $C^\gamma(\phi_t) \leq C^\omega(\phi_t)$. If $c(\tau)[\phi_0^\tau] > 0$ then $P(\tau) \neq c(\tau)[1]$, hence for $\overline{N}$ sufficiently large $P(\tau) \neq c(\tau)[\varphi_0^\tau]$ and since $c(\tau)[\phi_0^\tau] \geq -P(\tau)$ in fact $c(\tau)[\phi_0^\tau] > -P(\tau)$. Hence if $\overline{N}$ is sufficiently large $C^\gamma(\phi_t) \leq C^\omega(\phi_t)$. □

Lemma 4.3. *There exists $\lambda > 0$ and $\theta < 1$ such that if τ is a conformist type for $\phi_t^\tau > \theta$ and $\sigma \neq \tau$ we have $C^\tau(\phi_t) + \lambda < C^\sigma(\phi_t)$.*

Proof. Since there are finitely many conformist types is suffices to find a λ and θ for each conformist τ.

Type σ has cost

$$C^\sigma(\phi_t) \geq c(\sigma)[\phi_t^\sigma] + \phi_t^\tau P(\tau) \geq \phi_t^\tau P(\tau)$$

and letting $\overline{P} = \max_\omega P(\omega)$ type τ has cost

$$C^\tau(\phi_t) \leq c(\tau)[\phi_t^\tau] + (1 - \phi_t^\tau)\overline{P}.$$

Hence it is sufficient for $C^\tau(\phi_t) + \lambda < C^\sigma(\phi_t)$ that

$$c(\tau)[\phi_t^\tau] + (1 - \phi_t^\tau)\overline{P} + \lambda < \phi_t^\tau P(\tau)$$

or

$$\frac{c(\tau)[\phi_t^\tau] + \lambda + \overline{P}}{P(\tau) + \overline{P}} < \phi_t^\tau.$$

Since τ is conformist we have $c(\tau)[1] < P(\tau)$. Consequently there is a $\theta_1 < 1$ such that $\bar{c}(\tau) \equiv \max_{\varphi\geq\theta_1} c(\tau)[\varphi] < P(\gamma)$. Hence I choose $0 < \lambda < P(\tau) - \bar{c}(\tau)$ implying

$$\frac{c(\tau)[\phi_t^\tau] + \lambda + \overline{P}}{P(\tau) + \overline{P}} < \theta_2 < 1.$$

Choosing $\theta = \max\{\theta_1, \theta_1\}$ yields the desired bound. □

Corollary 4.4. *For given $\underline{N}$ if $\overline{N}$ is sufficiently large there is $\alpha > 0$ so that in a conformist τ-state to reach another conformist state without a catastrophe has a resistance at least $\alpha\overline{N}$. This means that for $\epsilon = 0$ conformist states are absorbing.*

By Lemmas 4.1 and 4.2 all other states have zero resistance paths to one of these absorbing states so they are transient. Hence with $\epsilon = 0$ only conformist

states have positive weight in any ergodic distribution, so only they can be stochastically stable. This is part (i) of the Theorem.

Proof. Let τ be conformist and choose λ and θ by Lemma 4.3 so that for $\phi_t^\tau > \theta$ and $\sigma \neq \tau$ we have $C^\tau(\phi_t) + \lambda < C^\sigma(\phi_t)$. Let $\overline{C}^\tau = \max_{\varphi \geq \theta} C^\tau(\varphi)$ and $\overline{C} = \max c(\sigma) + \max P(\sigma)$. Then for $\phi_t^\tau > \theta$ the average cost $\overline{z}$ is at most $(1-\phi_t^\tau)\overline{C} + \phi_t \overline{C}^\tau$ while

$$C^\sigma(\phi_t) > \overline{C}^\tau + \lambda \geq \frac{\overline{z} - (1-\phi_t^\tau)\overline{C}}{\phi_t^\tau} + \lambda$$

$$= \overline{z} - \frac{1-\phi_t^\tau}{\phi_t^\tau}\left(\overline{C} - \overline{z}\right) + \lambda.$$

Hence for

$$\phi_t^\tau \geq \frac{\overline{C} - \overline{z}}{\overline{C} - \overline{z} + \lambda/2} \equiv \overline{\theta}$$

$C^\sigma(\phi_t) > \overline{z} + \lambda/2$, that is $\sigma \neq \tau$ has above average cost. Taking $\underline{\theta} = \sum_\sigma m_t/\overline{N}$ we may choose $\alpha = 1 - \min\{\theta - \underline{\theta}, \overline{\theta} - \underline{\theta}\}$ so that at least $\alpha\overline{N}$ reproductions of cost at least one are needed to escape from a τ-state. □

Corollary 4.5. *After a catastrophe the resistance to reach a conformist τ-state is at most $\hat{R} \equiv \overline{r}(1 + \underline{N}/(1-\theta))$.*

Proof. To see this add $K \geq \underline{N}/(1-\theta)$ of conformist type τ to the post-catastrophe population of $\underline{N}$ and recall that each addition has resistance no greater than $\overline{r}$. By Lemma 4.3 τ has cost at least λ less than any other type. Hence the resistance to reach the τ-state after a catastrophe is at most $\hat{R}$.

Let $\overline{W}$ denote the maximum of $W(\tau)$ over conformist τ. □

Lemma 4.6. *There is a $\hat{v}$ large enough that for $v > \hat{v}$ the resistance of a mechanism designer state to a catastrophe is at least $\hat{R}$ greater than a non-mechanism designer state.*

Proof. Since $r_\kappa(vW_t, N_t/\overline{N})$ is strictly increasing and has slope bounded above zero in the first argument it follows that there exists a $\overline{v}$ large enough that for $v > \overline{v}$ and all $W(\tau) < \overline{W}$ we have $r_\kappa(vW, N_t/\overline{N}) - r_\kappa(vW(\tau), N_t/\overline{N}) > \hat{R}$. □

Lemma 4.7. *For fixed v there is an $\overline{N}$ large enough the least resistance route from one conformist τ-state to another is by having a catastrophe.*

Proof. The greatest resistance if there is an immediate catastrophe is $r_\kappa(v\overline{W}, 1) + \hat{R}$, while if there is no catastrophe at all it is at least $\alpha\overline{N}$. Hence for $\overline{N} > r_\kappa(v\overline{W}, 1) + \hat{R}$ the least resistance paths between conforming states must have a catastrophe.

Finally, I show that if $\overline{N}$ is large enough it is not possible to reduce resistance below $r_\kappa(vW(\tau), N_t/\overline{N})$ by increasing a type with low resilience then having a catastrophe. As I just showed that we cannot leave the basin of the conformist

state τ without greater resistance than an immediate catastrophe it suffices to show that an above average cost reproduction with cost at least one reduces the resistance to a castrophe by less than one. Consider then a given W and an above average cost of reproduction leading to $\hat{W}$. Let $\overline{w} = \max_\sigma w(\sigma)$. As resilience is an average, $|W - \hat{W}| \leq \overline{w}/\overline{N}$. Recall that r_κ has slope bounded away from zero in the first argument, say by $b > 0$. Hence

$$|r_\kappa(vW, 1) - r_\kappa(v\hat{W}, 1)| \leq vb\overline{w}/\overline{N}$$

and we see that if $\overline{N}$ is sufficiently large this is less than one. □

Lemma 4.8. *For given $\underline{N}$ there exists a $\hat{v}$ and for any $v > \hat{v}$ if $\overline{N}$ is sufficiently large only mechanism design states are stochastically stable.*

This is part (ii) of the Theorem.

Proof. Fix $\underline{N}$ and choose $\hat{v}$ from Lemma4.6. Choose $v > \hat{v}$ and define sufficiently large $\overline{N}$ by Lemma 4.7.

From Young (1993) stochastically stable states are the roots of least resistance trees on the conformist states. Suppose τ is at the root of a least resistance tree and is not a mechanism design state. Find some mechanism design state σ in the tree and cut it out from the state γ to which is was connected. By Lemma 4.7 this saves a resistance of at least $r_\kappa(v\overline{W}, 1)$. Attach the previous root τ to σ making σ the root. Also by Lemma 4.7 and Corollary 4.5 this adds a resistance of at most $r_\kappa(vW(\tau), 1) + \hat{R}$. Hence resistance is decreased by

$$r_\kappa(v\overline{W}, 1) - r_\kappa(vW(\tau), 1) - \hat{R}$$

and by Lemma 4.6 this is strictly positive. Consequently no state that is not a mechanism design state can be at the root of a least cost tree, so is not stochastically stable. □

5. Simulations

There is a tendency to think that small ϵ means "negligible" and that $\overline{N}$ large relative to $\underline{N}$ means $\overline{N}/\underline{N}$ is "neglible." In fact the evolutionary processes are numerically quite robust: this is supported by a recent theoretical literature including Kreindler and Young (2014) and Ellison, Fudenberg and Imhof (2016). I illustrate this through a Monte Carlo simulation that also highlights the main findings.

For simplicity, in the simulations, I take the linear case with flat cost, choosing types so that $c(\tau)[\varphi] = \xi(w(\tau), P(\tau))[\varphi] = \alpha w(\tau) + \psi \max\{0, P(\tau) - \underline{P}\}$. I take

α	$\underline{P}$	ψ
0.10	1.00	0.33

I will work with four benchmark types: selfish types, nasty types, altruistic types and mechanism designer types given by

τ	$c(\tau)[\varphi]$	$w(\tau)$	$P(\tau)$
selfish	0	0	0
nasty	0	0	$1 = \underline{P}$
altruistic	$1 = \alpha w(\tau)$	10	0
mechanism designer	$2 = \alpha w(\tau) + \psi(P(\tau) - \underline{P})$	10	4

Note that if the maximum cost $\bar{c} = 2$ and the strictness coefficient $\mu = 3/2$ then the mechanism designer is the best feasible type. The Monte Carlo was conducted in R for $40,000$ periods.

The residual population has one of each type, the maximum population is $\overline{N} = 40$, the minimum population is $\underline{N} = 6$ and $\epsilon = 0.5$. Notice that the population fall in the case of a catastrophe is 85% which is large but not extreme, and $\epsilon = 0.5$ which is hardly negligible. The numbers bear a reasonable relationship to human history. From Bowles and Choi (2013) we know that most human evolution took place in relatively small groups, on the order of $\overline{N} = 40$. With typical ages on the same order, the replacement rate is about one per year, so that the relevant length of a period is one year, so $40,000$ periods is roughly half the history of homo sapiens.

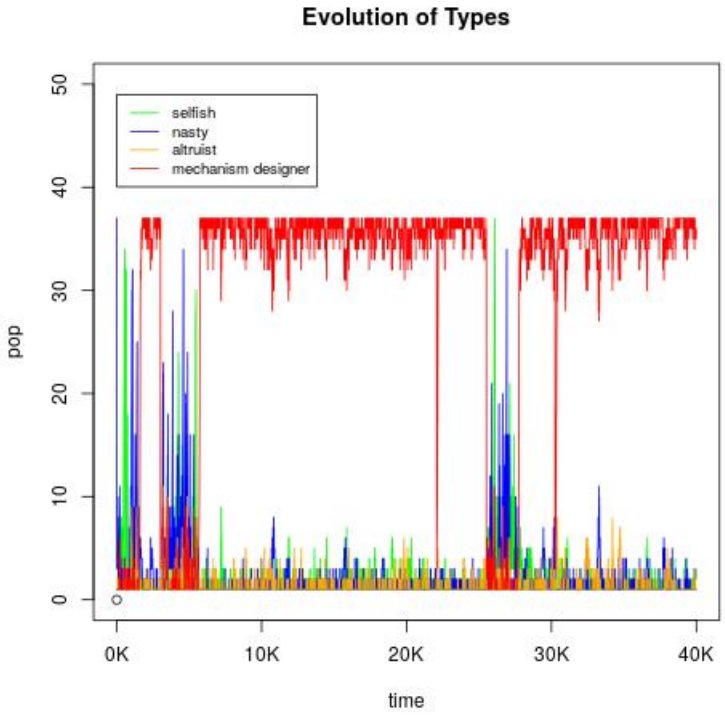
The probability of catastrophe is given by

$$h_\kappa(vW, N_t/\overline{N}) = \epsilon^{3.2+W}$$

and in particular $h_\kappa(0, N_t/N) \approx 1/10$ and $h_\kappa(10, N_t/N) \approx 1/1000$, which is to say for the types that produce no resilience catastrophes are roughly once every ten years, while for the types that do produce resilience it is about once every thousand years.

The probability that reproduction takes place in the absence of a catastrophe is $0.9N_t/\overline{N}$. The Darwinian dynamic is given by a reproduction probability of $\epsilon = 0.5$ for unfit types, and 1.0 for those with weakly below average cost. The initial population $N_1 = 40$ and the inital variable population consists entirely of nasty types.

The results of the Monte Carlo are shown in the graph below.



As can be seen the types that are not conformist, the selfish and altruistic type, play little role. Despite the fact that the simulations are slanted in favor of the nasty types by starting in the nasty state, never-the-less the mechanism designers predominate. In 82% of the periods they constitute more than 75% of the population. To verify this, I took the seed used in the figure the Monte Carlo and repeated the Monte Carlo 100 times incrementing the seed by one each time. The average number of periods during which the mechanism designers constituted more than 75% of the population over these 100 simulations was slightly above 50%. However, this understates the importance of mechanism designers. When there are many mechanism designers the population tends to be large as they have few catastrophes: this can be seen in the figure. A better measure of evolutionary success is the fraction of mechanism designers in the cumulative population. This is considerably higher: 69%.

6. Survival of the Nastiest

What happens when catastrophes are not so important? We know that only conformist types can be stochastically stable, but which ones? In this section I give an indication of what can happen.

To begin with, I modified the first Monte Carlo example reducing the importance of catastrophes by using

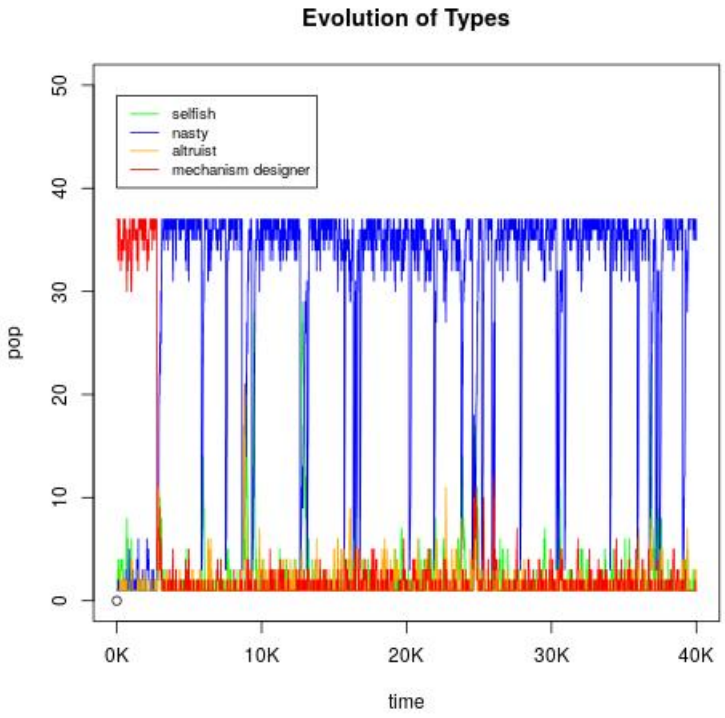
$$h_\kappa(vW, N_t/\overline{N}) = \epsilon^{8.1+W}$$

in place of

$$h_\kappa(vW, N_t/\overline{N}) = \epsilon^{2.7+W}.$$

I also switched to an initial population of mechanism designers.

The result is plotted in the graph below, in which we see that now the nasty type dominates.



What can be said more generally about diminished importance of catastropes? There is an uninteresting theorem which never-the-less shows some of the difficulties involved. Suppose that there is a super-conformist type punishes vastly more than any other type. Once a decent fraction of these enter the population only this super-conformist type is above average, and if catastrophes do not play an important role only the super-conformist state is stochastically stable. I say this theorem is uninteresting because it is hard to understand why there should be a super-conformist type.

If we rule out super-conformist types, again assuming that catastrophes do not play an important role, a natural candidate for stochastic stability is a type

that minimizes costs but still demands conformity. Let us assume that this is a type τ with zero costs, and continue to assume flat cost, that $c(\sigma)[\varphi] = c_0(\sigma)$. Suppose that while the ability of types to punish can rise with cost, it does not do so too fast in the sense that for every type $\sigma \neq \tau$ we have $c_0(\sigma) \geq (P(\sigma) - P(\tau))/2$. This rules out superconformist types. In this case we refer to type τ as the *nastiest type*. Notice that this is satisfied in the example for the nasty type since for the mechanism designer $c_0(\sigma) = 2$ while the punishment difference is $P(\sigma) - P(\tau) = 3$.

Simply reducing the chances of a catastrophe, however, is not enough to lead to dominance by the nastiest type. Without catastrophes we need to deal carefully with the size of the basins of the absorbing states and this is complicated by the possibility of a mixture of different types entering. This is a problem that has bedeviled the literature on evolution in repeated games, as can be seen from the difficult analysis of Johnson, Levine and Pesendorfer (2001) and is often avoided by limiting the number of types, or introducing special assumptions such as assuming that evolution largely proceeds through imitation as in Levine and Pesendorer (2007).

The problem is this: suppose some fairly high cost type enters in a decent fraction. This raises the average cost quite a bit and may mean that some third type now has below average cost so might not have resistance to reproduction. However: this cannot occur in the *best-response dynamics* in which only least cost types have zero resistance to reproduction and all other types have resistance ϵ. With this assumption we have the following theorem:

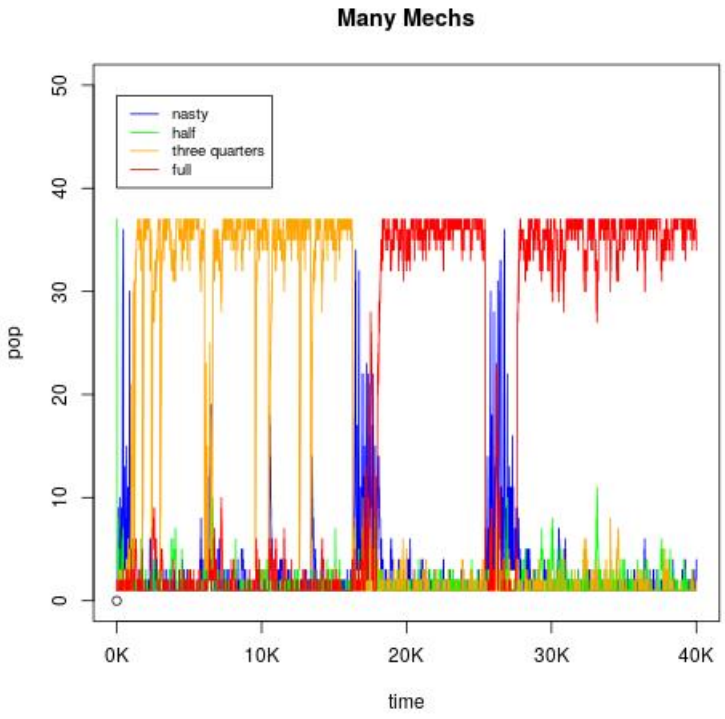
Theorem 6.1. *With flat cost, in the best response dynamic for given M and sufficiently large $\overline{N}$ there is a $\overline{R} > 0$ such that for $\inf_\varphi r_\kappa(0, \varphi) > \overline{R}$ the nastiest state is the unique stochastically stable state.*

Proof. The cost condition guarantees that if the nastiest type is 50% or more of the population then it does strictly better than than any other type. So the same remains true for some $\theta < 1/2$. By making $\overline{N}$ large enough we insure this is true accounting for the residual types and that the actual number of variable nastiest types needed is strictly less than half the population. If we then take $\overline{R}$ large enough that $\epsilon^{\overline{N}} < \overline{R}$ then least resistance paths cannot include a catastrophe so we are down to the standard case of analyzing least resistance paths for a fixed population of $\overline{N}$. However, the fact that it takes more than half the population to reproduce with resistance ϵ to escape the nastiest state, while it takes the nastiest type to reproduce with resistance ϵ strictly fewer times to go from any other τ-state to the nastiest state means leads to the standard result, for example, Morris, Rob and Shin (1995), that only the nastiest state is stochastically stable. This is easily proven by taking any tree in which the nastiest state is not the root, cutting it and attaching the root to it and observing that this strictly reduces the resistance. □

7. The Role of the Residual Population

After a collapse the residual population plays a key role in regenerating the group. Theorem 3.1 says that for v and $\overline{N}$ big enough the composition of the residual population does not matter, but in practice it does. Monte Carlo simulations enable us to explore this more carefully.

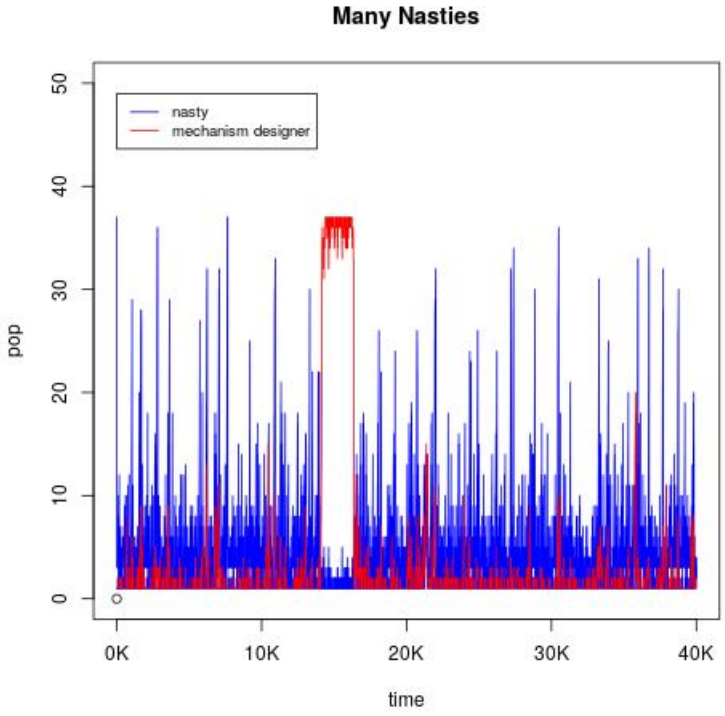
We think of good ideas as scarce in a sea of bad ideas. The residual population consisting of one of each type contains two types that are dysfunctional in the sense that they are not conformist so that the real contest is between the nasty type and the mechanism designer. To better capture the idea that good ideas are scarce I replaced the two dysfunctional types with "imperfect" mechanism design types. I refer to these as the half and three quarters types, and to the original mechanism design type as the full type. All incur the same cost $c(1/2) = c(3/4) = c(1) = 2$ and issue the same punishment $P(1/2) = P(3/4) = P(1) = 4$. They differ, however, in how much of the public good they produce: the full mechanism designer produces $w(1) = 10$ as before, but the half type produces half as much $w(1/2) = 5$ and the three quarters type three quarters as much $w(3/4) = 7.5$. In this case the graph below shows that the three quarters type provides serious competition for the full type.



This particular sample is somewhat atypical, however, in that in a hundred samples it is still the case that the fraction of the time that the full type consti-

tutes 75% of the population remains slightly above 50% and while the fraction of full types in the total population drops from 69% to 57%, it is still relatively high.

An alternative form of competition is to replace the dysfunctional types with nasty types. With three nasty types and $\epsilon = 0.5$ can the mechanism designers emerge? The answer, shown in the graph below, is no.

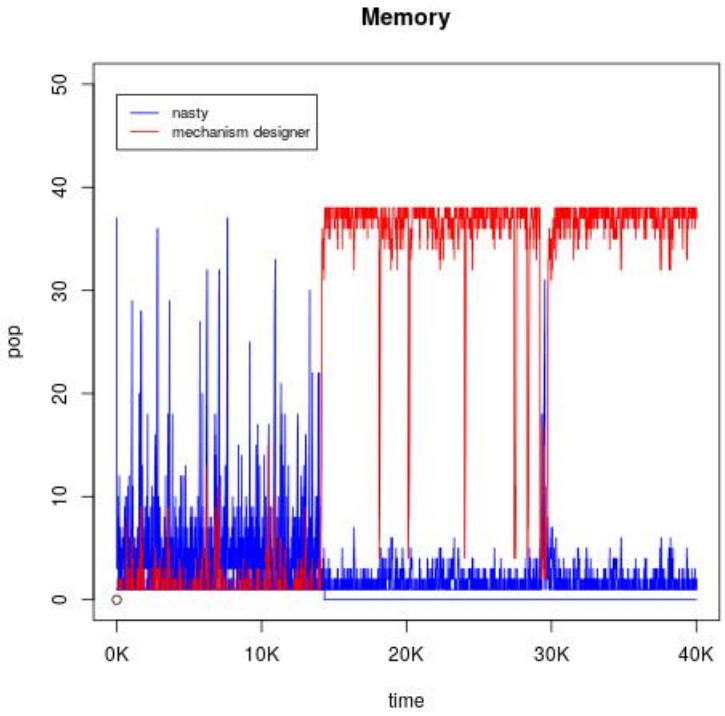


Indeed in a hundred samples only 27% of the time do mechanism designers constitute more than 75% of the population, although due to low populations during the nasty episodes their overall fraction of the population is still 59%.

This brings me to the final point I would like to make. So far I have loaded the dice against mechanism designers. In particular while I have been agnostic about whether evolution is biological or social, the model is largely biological in the sense that DNA has no memory and neither does the evolutionary process here. Look again at the graph above. There are vast vistas of time dominated by nasty types with repeated catastrophes, and a period of mechanism design about a thousand years long with a stable population and no catastrophes. Is it so hard to imagine that after the end of the mechanism design episode people talk to each other and say: "gosh, that was a lot better than those nasty periods, maybe we ought to consider keeping the mechanism design institutions?" In practice memory of past institutions is quite strong: even today Greece and

North Macedonia argue over who is the rightful heir of Alexander the Great, who died more than 2300 years ago.

To reflect fond memories of past episodes of mechanism design I modified the model slightly so that after the first successful mechanism design episode the residual population shifts and one of the original three nasty types instead becomes the mechanism design type. In other words after the first successful mechanism design episode the residual population shifts from three nasty to one mechanism designer to equal proportion of both. The consequence shown in the graph below is dramatic.



Once the mechanism designers arrive they are very hard to get rid of. In the hundred iterations 66% of the time mechanism designers constitute more than 75% of the population, and their overall fraction of the population is 83%.

8. Identification of Types

Crucial to the theory is the assumption that types are observable, although possibly with error. This has been a fraught subject in the evolutionary literature. Robson (1990)'s discussion of the secret handshake, Levine and Pesendorer (2007)'s assumption that lying is difficult, Levine and Szentes (2006) discussion of the feasibility of identifying those using the same rules, and in the repeated

game setting the difficulty in observing off path play, are all examples of the obstacles.

In this setting, if costs, public good production, and the act of punishment are observable, the issue is not so problematic. Individuals can be required to identify themselves as a type by making a public announcement, that is stating a τ. It is the duty of each individual of a particular type to incur the prescribed cost and produce the prescribed amount of public good, and to punish anyone who announces a different type or fails to carry out the duties of their own type. If costs, public good production, the act of punishment, and the announcements are perfectly observed, then so are types, and indeed, the probability of accidentally punishing the own type, $\pi = 0$.

There is a recursive nature this type of scheme: failure to punish another type must be punished, and failure to punish an individual who failed to punish another type must be punished and so forth. This can be incorporated in a repeated game of auditing and punishment that ends rather quickly as in Levine and Modica (2016b). More to the point, this is the way real social norms work: as Skarbek (2014) documents even prison gangs have figured this out, so I take it that we as people know how to do this. Another way to say that is that being a type that punishes means figuring out how to act recursively and as we have seen only these types can be stochastically stable.

9. Robustness

I want to emphasize how the limits work in Theorem 3.1: the theorem is not about what happens "at the limit" but about what happens "as the limit is approached." In particular, it says that for any given $\zeta < 1$ we can find an $\epsilon > 0$ so that the fraction of the time f_ϵ system spends in stochastically stable states is strictly greater than ζ. Fixing that ϵ the system is a finite state ergodic Markov process: the ergodic distribution is therefore a locally continuous function of parameters that continuously change the transition matrix. That means that for sufficiently small perturbations of these parameters the system will still spend at least the fraction ζ of the time in the stochastically stable states. In particular: if you are bothered by the fact that it is possible to punish a small amount without any cost at all, introducing a slight cost of punishment for the nasty types (strictly increasing cost of punishment) will not change the result that the system spends most of its time in mechanism designer states.

10. Conclusion

We are mechanism designers. We understand that we and other people have motives, have an idea what those motives are, and organize incentives to direct those motives to serve social purposes. The writers of the US Constitutions were aware of human fallability and designed a system of democracy guarded by checks and balances to both elicit preferences through voting and see that those preferences are reflected in public policy. Their design was based on explicit

consideration of incentives and on millenia of thought and writing about human motivation and incentives: two thousand years earlier, even, Plato considered five different mechanisms and analyzed how they might serve the common good. While Plato is perhaps the first to explicitly analyze alternative mechanisms, mechanisms explicitly designed to enforce social rule in the forms of codes of law are considerably more ancient. The written code of law that has survived to the current day is the Code of Hammurabi dating from about 1750 BCE, but we know of similar earlier codes of law such as the Code of Uruk dating to about 2400 BCE. Prior to the advent of writing we know little about social organization and the types of incentives used, but we do know that public goods, and indeed public goods providing resilience, were produced. Town and village walls are a clear example, being both a public good, and protecting from the catastrophe of invasion by neighbors. These date back to at least shortly after the end of the Younger Dryas, as the Walls of Jericho were built around 8000 BCE. From the perspective of the theory, this suggests that we had already evolved to be mechanism designers by the end of the Younger Dryas and before the advent of agriculture.

Here I have explored how it might have come about that we are mechanism designers and what is the objective function we maximize. The key insight of the paper is that a shock that causes a reduction in population leads to a state that is relatively easy to leave. This favors mechanism designers who provide public goods that mitigate the chances of catastrophe. That we design mechanisms for resilience and not for efficiency sheds light on some seeming failures of mechanism design. In some cases, of course, efficiency leads to greater resilience: for example, a more efficient private firm is better able to survive. However, it has long been noted that government bureaucracies such as those approving drugs are not efficient. Indeed, drug approval bureaucracies are designed to err on the side of disapproval, costing countless lives by failing to approve or delaying the approval of live-saving drugs, while saving few by keeping defective drugs off the market. It has equally been observed that the reason for this is that failing to approve drugs has little consequence for the organization, while approving a defective drug is disastrous: seen through the lens of resilience, the structure of these bureaucracies makes perfect sense.

References

Akerlof, G. and R. E. Kranton (2005): "Identity and the Economics of Organizations," *Journal of Economic Perspectives*.

Alger, I., and Weibull, J. W. (2013): "Homo moralis—preference evolution under incomplete information and assortative matching," *Econometrica* 81: 2269-2302.

Axelrod, R. and W. D. Hamilton (1981): "The evolution of cooperation," *Science*.

Binmore, Kenneth G., and Larry Samuelson (1992): "Evolutionary stability in repeated games played by finite automata," *Journal of Economic Theory* 57: 278-305.

Bowles, Samuel and Jung-Kyoo Choi (2013): "Coevolution of farming and private property during the early Holocene" *Proceedings of the National Academy of Science*, doi:10.1073/pnas.1212149110

Dal Bó and Pujals (2015): "The Evolutionary Robustness of Forgiveness and Cooperation," mimeo.

Bilancini, Ennio, Leonardo Boncinelli, and Pablo Marcos-Prieto (2022): "Conflict Initiation Function Shapes the Evolution of Persistent Outcomes in Group Conflict" mimeo IMT Lucca

Coase, R. H. (1960): "The Problem of Social Cost," *Journal of Law and Economics* 3: 1-44.

Choi, Jung-Kyoo and Samuel Bowles (2007): "The coevolution of parochial altruism and war," *Science* 318: 636-40.

Della Vigna S., J. A. List, U. Malmendier and G. Rao (2014): "Voting to Tell Others," NBER Working Paper.

Diamond, Jared (2005): *Collapse: How Societies Choose to Fail or Succeed*, Viking Press.

Dutta, R., D. K. Levine and S. Modica (2022): "The Whip and the Bible: Punishment Versus Internalization," *Journal of Public Economic Theory*, 23: 858-894

Dutta, R., D. K. Levine and S. Modica (2021): "Interventions with Sticky Social Norms: A Critique," *Journal of the European Economic Association*, forthcoming.

Ellison, Glenn (2000): "Basins of Attraction, Long Run Stochastic Stability and the Speed of Step-by-step Evolution," *Review of Economic Studies* 67: 17-45.

Ellison, G., D. Fudenberg and L. A. Imhof (2016): "Fast convergence in evolutionary models: A Lyapunov approach," *Journal of Economic Theory* 161: 1-36.

Ely, Jeffrey C. (2002): "Local conventions," *Advances in Theoretical Economics* 2.

Fehr, E. and S. Gächter (2000a): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review* 90: 980-994.

Fehr, E., and S. Gächter (2000b): "Fairness and Retaliation: The Economics of Reciprocity," *The Journal of Economic Perspectives* 14: 159–181.

Grossman, S. J., and O. D. Hart (1986): "The costs and benefits of ownership: A theory of vertical and lateral integration," *Journal of Political Economy* 94: 691-719.

Juang, W-T. and Sabourian, H. (2021): "Rules and Mutation - A Theory of How Efficiency and Rawlsian Egalitarianism/Symmetry May Emerge," mimeo Cambridge.

Jablonski, David (2001): "Lessons from the past: Evolutionary impacts of mass extinctions," *Proceedings of the National Academy of Sciences* 98: 5393-5398.

Johnson, P., D. K. Levine and W. Pesendorfer (2001): "Evolution and Information in a Gift Giving Game," *Journal of Economic Theory* 100: 1-22.

Kandori, Michihiro, George Mailath, and Rafael Rob (1993): "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica* 61: 29-56.

Kreindler, G. E. and H. P. Young (2014): "Rapid innovation diffusion in social networks," *Proceedings of the National Academy of Sciences* 111: 10881-10888.

Twine, K. (1992): "The city in decline: Rome in late antiquity," *Middle States Geographer* 25: 134-138.

Levenstein, M. C. and Suslow, V. Y. (2006): "What determines cartel success?" *Journal of Economic Literature* 44: 43-95.

Levine, David K. and Andrea Mattozzi (2020): "Voter Turnout with Peer Punishment," *American Economic Review* 110: 3298–3314.

Levine, David and Salvatore Modica (2012): "Conflict and the Evolution of Societies," working paper, www.dklevine.com

Levine, David and Salvatore Modica (2013): "Anti-Malthus: Conflict and the Evolution of Societies," *Research in Economics* 67: 289-306

Levine, David and Salvatore Modica (2016a): "Dynamics in stochastic evolutionary models", *Theoretical Economics* 11: 89-131

Levine, David and Salvatore Modica (2016b): "Peer Discipline and Incentives within Groups," *Journal of Economic Behavior and Organization* 123: 19-30

Levine, D. K. and S. Modica (2017): "Size, Fungibility, and the Strength of Organizations," *European Journal of Political Economy* 49: 71-83

Levine, D. K. and S. Modica (2022): "Survival of the Weakest: Why the West Rules," *Journal of Behavior and Economic Organization*, forthcoming.

Levine, David K., and Wolfgang Pesendorfer (2007): "The evolution of cooperation through imitation," *Games and Economic Behavior* 58: 293-315.

Levine, D. K. and B. Szentes (2006): "Can A Turing Player Identify Itself?," *Economics Bulletin* 1: 1-6

Morris, S., R. Rob and S. Shin (1995): "p-Dominance and belief potential," *Econometrica*: 145-157.

Ostrom, Elinor (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.

Peteet, D. (1995): "Global younger dryas?" *Quaternary International* 28: 93-104.

Peysakhovich, A. and D. Rand (2014): "Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory," mimeo.

Robson, A. J. (1990): "Efficiency in evolutionary games: Darwin, Nash and the secret handshake.," *Journal of theoretical Biology*, 144: 379-396.

Ross, David (2002): *Ireland: History of a Nation*, New Lanark: Geddes & Grosset.

Skarbek, D. (2014): "The social order of the underworld: How prison gangs govern the American penal system," O*xford University Press.*

Townsend, R. M. (1994): "Risk and insurance in village India," *Econometrica*, 539-591.

Winter, S. G. (1971): "Satisficing, selection, and the innovating remnant," *Quarterly Journal of Economics* 85: 237-261.

Young, P. (1993): "The Evolution of Conventions," *Econometrica* 61: 57-83