

# Cooperation, Punishment, Emergence of Government, and the Tragedy of Authorities

**R. Vilela Mendes**

*CMAF, Instituto de Investigação Interdisciplinar  
Av. Gama Pinto 2, 1649-003 Lisboa, Portugal  
and  
IPFN, Instituto Superior Técnico  
Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal  
vilela@cii.fc.ul.pt*

**Carlos Aguirre**

*Escuela Politécnica Superior  
Universidad Autónoma de Madrid, Campus de Cantoblanco  
Ctra de Colmenar Km 16, 28049 Madrid, Spain  
carlos.aguirre@uam.es*

---

Under the conditions prevalent in the late Pleistocene epoch (small hunter-gatherer groups and frequent inter-group conflicts), coevolution of gene-related behavior and culturally transmitted group-level institutions provides a plausible explanation for the parochial altruistic and reciprocator traits of most modern humans. When, with the agricultural revolution, societies became larger and more complex, the collective nature of the monitoring and punishment of norm violators was no longer effective. This led to the emergence of new institutions of governance and social hierarchies. The transition from an egalitarian society and the acceptance of the new institutions may have been possible only if, in the majority of the population, the reciprocator trait had become an internalized norm. However, the new ruling class has its own dynamics, which in turn may lead to a new social crisis. Using a simple model inspired by previous work by Bowles and Gintis, these effects are studied here.

---

## 1. Introduction

---

It is a fact that humans are a highly cooperative species. Cooperative in helping each other, cooperative in achieving material and intellectual achievements unmatched by other species, but also cooperative in war and genocide. From the biological point of view, human cooperation is an evolutionary puzzle. Unlike other creatures, humans cooperate with genetically unrelated individuals, with people they will never meet again, and when reputation gains are small or absent, even engaging in altruistic punishment of defectors. These patterns of cooper-

ation cannot be explained by kin selection, signaling theory, or reciprocal altruism. The idea that group selection might explain this behavior goes back to Darwin himself who, in chapter 5 of the *Descent of Man and Selection in Relation to Sex*, states that "... an increase in the number of well-endowed men and an advancement in the standard of morality will certainly give an immense advantage of one tribe over another." However, this idea fell into disrepute because evolution does not pitch groups against groups, nor individuals against individuals, but genes against genes. A "selfish gene" analysis makes the altruistic good-of-the-group outcome virtually impossible to achieve, in particular because the late Pleistocene groups of modern man were not believed to be sufficiently different genetically to favor group selection. Therefore, human cooperation remained an evolutionary puzzle.

In recent years S. Bowles, H. Gintis, and collaborators [1–4] revived the group selection idea by showing that the particular environment and type of the hunter-gatherer groups of the late Pleistocene (which corresponds to about 95% of the evolutionary time of modern man) were such that a multilevel evolutionary dynamic involving gene-culture coevolution could account for the development of the cooperative altruistic trait they call "strong reciprocity." The cost of group beneficial behavior to an individual would be limited by the emergence of group-level social norms. On the other hand, even in the absence of these group-level norms, group selection pressures would support the evolution of the cooperative-altruistic punishment trait if intergroup conflicts were very frequent. Egalitarian practices among ancestral humans reduced the force of individual selection against altruists, while frequent warfare made altruistic cooperation among group members essential to survival. That is, parochial altruism and warfare could have coevolved. Furthermore, they developed simple mathematical models that gave quantitative support to their ideas.

We think that the analysis of Bowles and Gintis provides a convincing picture of the genesis of the cooperative nature of humans and their culture. The human capacity for social norm building and for the cultural transmission of learned behavior allowed altruistic other-regarding preferences to proliferate. But it also suggests that the other-regarding preferences that we inherited from primeval man are partly cultural, not purely genetic, and therefore liable to change at a much faster pace than if they were purely genetic. A natural question is what is happening to this human trait (that presumably developed during a period of 190 000 years) in the short time (10 000 years) since the end of the Pleistocene. Using a simple version of the Bowles–Gintis model, the evolution of the reciprocator trait has been analyzed in [5], in a situation where the size of the society and the degree of clustering precludes the collective nature of rule violator monitoring. Both an agent-based and a mean-field model were used. The main conclusion was that in this situation, the reciprocator trait would not be evolutionary stable.

Historically, it is known that such a transition from the small hunter-gatherer groups to larger sedentary population groups occurred at the time of the agricultural revolution and that the solution was “the emergence of government.” That is, new types of agents (rulers and authorities) came into play and replaced the type of egalitarian decision-making that might have existed before. It is worth noticing that the final, if difficult [6–8], acceptance of this transition of power may have to do with the internalization of the reciprocator trait, which valued the enforcing of social norms above complete freedom.

In the agricultural societies, specialization arose as well as new security needs and more intense population pressure on limited resources. This tended to produce greater organization within the community, which in turn led to social hierarchies, certain forms of chieftainship, and a whole class of people with managing roles.

In this paper, using a setting similar to the one in [5], we will study the effect of introducing in the model a new agent representing the role of the authorities. The collective monitoring and punishment of the reciprocators will be a decreasing function of the population size in the social group, which is allowed to grow with the average fitness. The need to introduce authority agents to avoid a “tragedy of the commons,” that is, a fitness crisis arising from the proliferation of self-regarding agents, is an expected effect: the emergence of government. The interesting question is that the dynamics of the authority agents may, by themselves, lead to a new fitness crisis called a “tragedy of authorities.” This crisis may or may not be related to the elite overproduction crisis that some authors [9–11] have identified. This will be discussed in Section 3.

## 2. Emergence of Government and the Tragedy of Authorities

The basic setting is similar to the one used before [2, 5] as far as the type of “public good activity” is concerned in a group of  $N$  agents, with  $N$  being in general a function of time. Here, however, three types of agents are considered. The first type (R agents) are cooperators that also have a monitoring effect on the cooperation of other agents. The second are self-regarding agents (S agents) and the third are purely monitoring agents (A agents). The labels that were chosen refer to the name reciprocators (R), self-regarding or shirkers (S), and authorities (A). The percentages of each one of the types in the population are denoted by  $f_R$ ,  $f_S$ , and  $f_A$ .

Each R or S agent can produce a maximum amount of goods  $q$  at cost  $b$  (with goods and costs in fitness units). An S agent benefits from shirking public good work by decreasing the cost of effort  $b(\sigma)$ ,  $\sigma$  being the fraction of time the agent shirks. As before, the following conditions hold:

$$b(0) = b, \quad b(1) = 0, \quad b'(\sigma) < 0, \quad b''(\sigma) > 0. \quad (1)$$

Furthermore,  $q(1 - \sigma) > b(\sigma)$  so that, at every level of effort, working helps the group more than it hurts the worker. This assumption, which restricts the class of “cost of effort” functions, also emphasizes the fact that it is desirable to explore a situation where cooperation may in the end be beneficial for everyone. Otherwise, cooperation would be pointless.

For  $b(\sigma)$ ,

$$b(\sigma) = \frac{2}{2\sigma - 1 + \sqrt{1 + 4/b}} - \frac{2}{1 + \sqrt{1 + 4/b}} \tag{2}$$

is chosen [1], which satisfies the constraints in equation (1).

R agents never shirk and punish each free rider at cost  $c\sigma$  and probability  $p(N)$ , the cost being shared by all R agents. For an S agent, the estimated cost of being punished is  $s\sigma$ , punishment being ostracism or some other fitness decreasing measure. Punishment and cost of punishment are proportional to the shirking time  $\sigma$ ,  $c$  is the reciprocator unit of punishment cost, and  $s$  is the weight given by an S agent to the possibility of being punished. It may or may not be the same as the actual fitness costs of punishment ( $\gamma, \gamma_A$ ). Each S agent chooses  $\sigma$  (the shirking time fraction) to minimize the function

$$B(\sigma) = b(\sigma) + s(f_R + f_A)\sigma - q(1 - \sigma)\frac{1}{N} \tag{3}$$

From the point of view of an S agent  $(f_R + f_A)\sigma$  is the probability of being monitored and punished. The last term is the agent’s share of his own production. The value  $\sigma_S$  that minimizes  $B(\sigma)$  is

$$\sigma_S = \max\left(\min\left(\frac{1}{2} - \sqrt{\frac{1}{4} + \frac{1}{b}} + \frac{1}{\sqrt{s(f_R + f_A) + q/N}}, 1\right), 0\right) \tag{4}$$

The contribution of each species to the population in the next time period is proportional to its fitness  $\pi_R, \pi_S$ , or  $\pi_A$ , computed from

$$\begin{aligned} \pi'_R &= q(1 - f_A - f_S \sigma_S)x - b - c p(N) f_S \frac{N \sigma_S}{N f_R} \\ \pi'_S &= q(1 - f_A - f_S \sigma_S)x - b(\sigma_S) - (\gamma p(N) f_R + \gamma_A f_A) \sigma_S \\ \pi'_A &= q(1 - f_A - f_S \sigma_S) w x - c_A f_S \frac{N \sigma_S}{N f_A} \end{aligned} \tag{5}$$

and  $\pi_{R,S,A} = \max(\pi'_{R,S,A}, 0)$  because the baseline fitness is zero. Notice that although we are using notions like fitness and evolutionary stability borrowed from genetic evolution throughout the paper, we are in fact dealing mostly with cultural processes.

The first term in  $\pi'_R$ ,  $\pi'_S$ , and  $\pi'_A$  is the benefit arising from the produced public goods. The factors  $x$  and  $w x$  with

$$x = \frac{1}{w f_A + 1 - f_A}$$

account for the fact that this benefit is the same for R and S agents but might be different for A agents. The second term in  $\pi'_R$  and  $\pi'_S$  is the work effort. The third term in  $\pi'_R$  and the second term in  $\pi'_A$  represent the fitness cost of punishment for R and A agents and the third term in  $\pi'_S$  represents the cost incurred by S agents when they are punished.

The  $\gamma$  and  $\gamma_A$  coefficients code for the severity of the coercive measures affecting the fitness of S agents. The last term in  $\pi'_R$  and  $\pi'_A$  emphasizes the heavy punishing burden put on R or A agents when in small numbers. The factor  $p(N)$ , a decreasing function of  $N$ , accounts for the fact that (as studied at length in [5]), when a social group grows in size, the collective nature of monitoring of free riders becomes increasingly difficult. Essentially, the punishment probability by R agents should be a growing function of the clustering coefficient of the group. Here, for illustration purposes, a simple function of  $N$  is chosen:

$$p(N) = \sqrt{\frac{1 + \delta}{1 + \delta(N/N_0)}} ,$$

with  $N_0$  being some small initial population.

Finally, for the evolution of the population at successive generations, a replicator map is chosen:

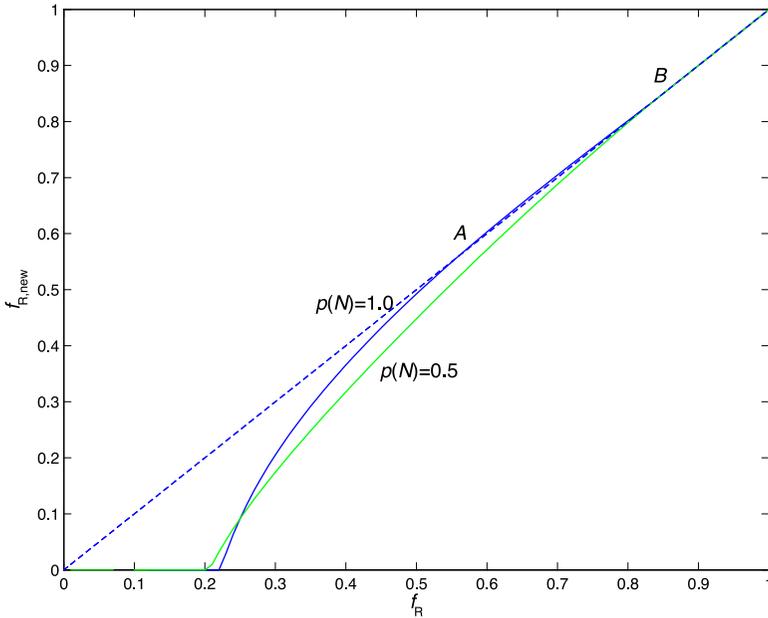
$$f_{\alpha, \text{new}} = f_{\alpha} \frac{\Pi_{\alpha}(f)}{f_R \Pi_S + f_S \Pi_S + f_A \Pi_A} \tag{6}$$

$\alpha = R, S, A$ .

A different, incremental dynamic is sometimes used for the fitness-based evolution of populations. The replicator map used here provides faster evolution but qualitatively similar results, up to a renormalization of the time scale.

First the dynamics of R and S agents are studied alone, keeping  $f_A = 0$ . In this case, using equations (4) and (5), the evolution of  $f_R = 1 - f_S$  corresponds to a one-dimensional map that is illustrated in Figure 1 for two values of  $p(N)$  (1.0 and 0.5). For  $p(N) = 1$ , the map has an unstable fixed point at A ( $f_R(A) \approx 0.57$ ), a left-stable fixed point at B ( $f_R(B) \approx 0.85$ ), and a continuum of neutral fixed points after that. For  $p(N) = 0.5$ , only the neutral fixed points remain. The neutral fixed points correspond to the situation where S

agents do not shirk for fear of being punished. For initial conditions smaller than  $f_R(A)$  in the first case or  $f_R(B)$  in the second, the population of R agents is always invaded by S agents. However, the neutrality of the fixed points means that the population of S agents is not completely invaded by the R agents.



**Figure 1.** One-dimensional map of the evolution of R agents corresponding to  $f_A = 0$ ,  $q = 2$ ,  $b = 1$ ,  $c = 0.1$ ,  $\gamma = 4$ ,  $s = 3$ , and  $N = 20$ .

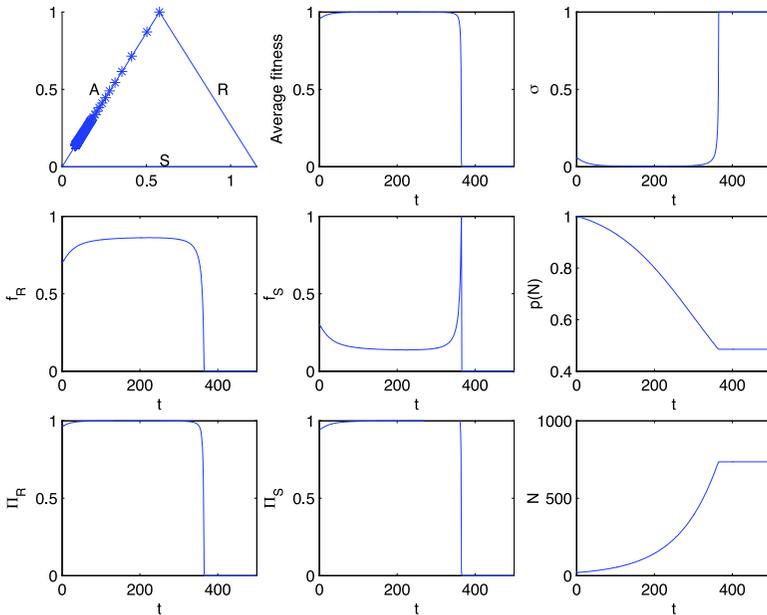
Next, still keeping  $f_A = 0$ , the evolution of the population of R and S agents is studied when the population increases in time according to a global fitness-dependent law, chosen as

$$N(t + 1) = N(t) e^{\beta\pi},$$

where  $\pi = \sum_{\alpha} f_{\alpha} \pi_{\alpha}$ .

Figure 2 displays the results for a time evolution starting from  $N_0 = 20$ ,  $f_R = 0.7$ , and  $f_S = 0.3$ . In the upper-left plot, the percentages  $f_R$ ,  $f_S$ , and  $f_A$  ( $f_A = 0$  in this case) of each agent type are displayed as the distances to the three sides of a triangle. As long as the population ( $N$ ) remains small, the monitoring effects of R agents controls shirking ( $\sigma$ ) by the S agents and, as a result, their percentage ( $f_R$ ) and fitness ( $\Pi_R$ ) increases, as well as the average fitness of the group. However, with further population growth the punishment probability ( $p(N)$ ) of shirkers decreases, leading for a while to a higher degree of

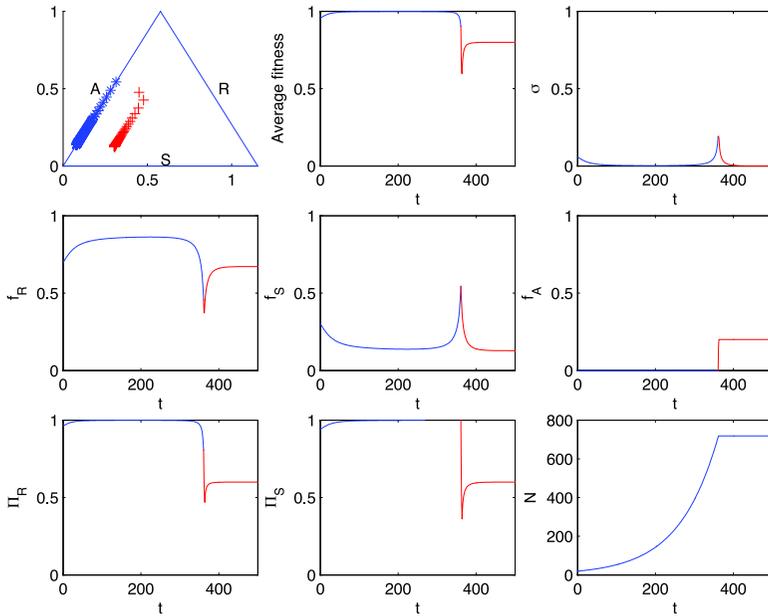
shirking ( $\sigma$ ) and higher fitness ( $\Pi_S$ ) and percentage ( $f_S$ ) of S agents. But because S agents with high  $\sigma$  produce much less goods, the fitness of all agents finally decreases and the group collapses. This is the well-known tragedy of the commons, here induced by the fact that monitoring of the public good behavior of the agents cannot be a fully collective activity in a large society.



**Figure 2.** Time evolution of R and S agents with  $f_A = 0$ ,  $q = 2$ ,  $b = 1$ ,  $c = 0.1$ ,  $\gamma = 4$ ,  $s = 3$ , and  $N_0 = 20$ .

It is then natural that a population group whose success is based on cooperation and control of selfish behavior would recognize the need, beyond a certain population level, to assign the control and punishing role to specialized agents with extra power and authority. This is called the emergence of government. The model now starts from the same initial conditions, but when  $f_R$  reaches a value below 0.5, the dynamics of A agents are unfrozen, imposing for the moment the constraint that  $f_A$  should not exceed 0.2 and, to isolate the effect of the A agents, the population is assumed to be constant after that moment. The result is shown in Figure 3.

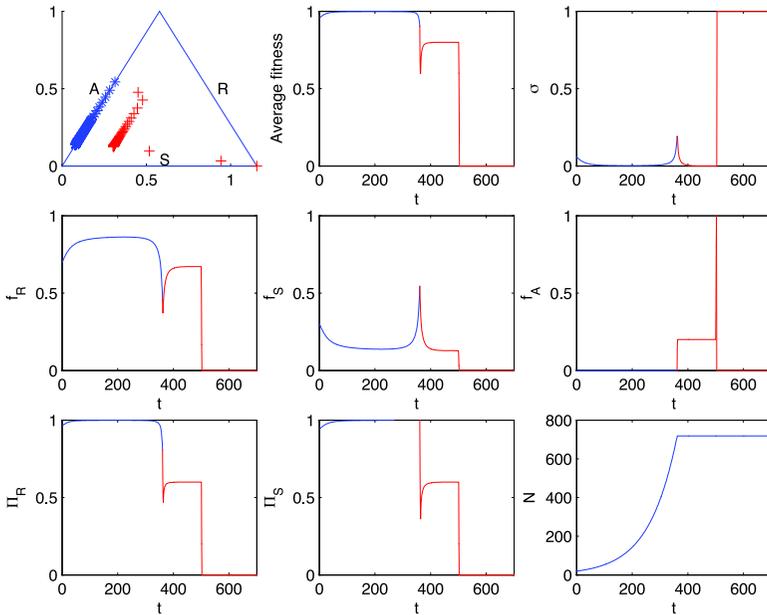
The outcome is rather satisfactory. After the unfreezing of the  $f_A$  dynamics, the percentage of R agents still decreases for a while, but then it starts to grow and the group stabilizes at a high level of average fitness.



**Figure 3.** Time evolution with the three types of agents but  $f_A \leq 0.2$  ( $q = 2$ ,  $b = 1$ ,  $c = 0.1$ ,  $s = 3$ ,  $c_A = 0.45$ ,  $\gamma = 4$ , and  $\gamma_A = 11$ ).

Notice that the growth of the number of A agents is rather fast. The reason is that as soon as they start controlling the behavior of the S agents, both  $\sigma$  and  $f_S$  decrease. This greatly increases the fitness of the A agents, because they benefit from the goods produced without incurring the cost of control because there is almost nothing to control anymore. If the 0.2 bound is now removed on  $f_A$  (Figure 4), the A agents' population continues to grow but, because they produce no goods, the average fitness finally decreases to zero as the group collapses. This is a crisis of a different type called the tragedy of authorities. What this means for actual societies will be discussed in Section 3.

A very similar effect is obtained if, while keeping  $f_A$  bounded,  $w$  is allowed to grow with the fitness of A agents, that is, allowing the share of goods allotted to A agents to grow. It may also happen that with the introduction of the authority agents they may, for example, demand an increase in public works for their benefit, essentially equivalent to a change of the  $b(\sigma)$  cost of effort function. However, the qualitative effects would not be changed with perhaps an even earlier tragedy of authorities outcome.



**Figure 4.** Time evolution with the three types of agents and  $f_A$  allowed to grow above 0.2 after time 500 ( $q = 2$ ,  $b = 1$ ,  $c = 0.1$ ,  $s = 3$ ,  $c_A = 0.45$ ,  $\gamma = 4$ , and  $\gamma_A = 11$ ).

A conclusion is that a stable society with government is only possible if the tragedy of authorities is avoided. To illustrate this effect, in our model we have imposed an artificial cut off on the number of A agents. In real life, a similar limitation may be obtained by internal competition of the A agents themselves.

Here we would only like to emphasize the delicate nature of the balance between the several agents in a viable society and the emergence of what seem to be universal features in the human social evolution. Cooperation is at the root of success in human groups. However, a natural, perhaps biological, tendency of humans to minimize effort and to maximize benefits requires that a certain amount of control of shirking is required. This led some humans to internalize the idea that shirkers should be controlled. Apparently, it is the societies where more humans adopted this norm that were the most successful. When, after the agricultural revolution, the human groups became larger, collective control became more difficult. Then, the evolved acceptance of social norms led naturally to the acceptance of government as a specialized body. However, the dynamic of the authority agents may, by itself, lead to a new fitness crisis.

### 3. Remarks and Conclusions

---

Stylized mathematical models, both in natural and human sciences, are not intended to take care of all the details that each particular system possesses. Rather, they are intended to extract general features or universal mechanisms, if any, that rule the dynamics of the system. Then, of course, the detailed characteristics that each physical system or society has will determine the time scales and intensity of the universal features.

The important point to retain is that the general behavior extracted from the simple equations (5) and (6) is qualitatively the same for a large range of parameter values. Hence, the general features that may be extracted from this and previous works are as follows.

1) Under the conditions prevalent in the late Pleistocene (i.e., small population groups, frequent inter-group conflicts, and a species with the capacity for norm enforcing and cultural transmission of learned behavior), the reciprocator trait may become dominant although, in general, not completely invasive of the self-regarding type. In fact, recent results in experimental games seem to indicate the existence of diverse subpopulations even now (defectors that always defect and cooperators that always cooperate, as well as consistently tit-for-tat individuals).

2) In a large population, monitoring of public good behavior cannot be a fully collective activity, rather being the chore of those in close contact with the free riders. Because punishment of free riders requires a local consensus among reciprocators, the clustering nature of the society would play an important role in the maintenance and evolution of the reciprocator trait. Although large human societies tend to be “small worlds” in the sense of short path lengths, they do not necessarily maintain a high degree of clustering. Therefore, norm monitoring and enforcing requires new special institutions of governance. However, the new institutions bring with them social hierarchies, which imply inequalities. Therefore, acceptance of the new institutions may have been possible only if, in the majority of the population, the reciprocator trait had become an internalized norm.

3) The evolutionary dynamics of the agents associated to governance, that is, the ruling class, may by its proliferation or by assigning to itself a higher share of the production (a high  $w$  factor in the Section 2 model) provoke a decrease of the average fitness, a crisis, or even a collapse of the society. This is what has been called here the tragedy of authorities. Some authors [9–11] have studied the historical effects of “elite overproduction” as generating crisis and revolutions. However, not all cases of elite overproduction that they characterize can be identified with the phenomena of the tragedy of authorities. If elite overproduction is, for example, the proliferation of an aristocratic class that, under the protection of the ruler, lives from the society production without contributing to it, then it has all the marks of a tragedy of authorities. But if, instead, elite overproduction

is associated with a higher access of the youth to higher education, this is not a tragedy of authorities. The eventual crisis that may occur in this case results from the fact that the new educated agents are not incorporated either in the productive sector or as beneficiaries of the society production. Hence, it is not a tragedy of authorities. In fact, they are only reacting against an authority structure that wants to preserve their privileges; therefore, to associate these two distinct situations under the same elite overproduction label may be quite misleading.

As shown in Section 2, the existence of authority agents is beneficial to society as long as their number and their share of the goods remains limited. The problem is the old question of “who guards the guardians” [12]. Democracy is in principle a way to implement limitations and accountability of the rulers. But even then, nothing is guaranteed. Economic power easily escapes constraints of democratic control. And even more subtle effects may occur. For example, through exploration of the co-evolved parochial feelings of the population, it is easy to erect as a goal the proliferation of local or regional government structures, coordinating committees, and layers and layers of control when there is nothing else to control.

Another example may be found in [13], where it is shown how well-organized groups can use seemingly irrational government policies to exploit poorly organized groups. Given rational predatory behavior between these groups, protection or any other redistributive policy that improves the chances of election of a party increases political efficiency. This can create an economic black hole, conditions under which an entire economy can disappear into lobbying.

4) Even subtler effects of emergent tragedies of authorities are found everywhere. The solidary form of collective government of the hunter-gatherer groups was probably the most successful invention of modern man, leading to his dominance over other species and even over other hominids. It was also the most extensively tested of all, lasting for 95% of the evolutionary history of modern man. Centralized, professional forms of government, by comparison, are a very recent development and not always very successful. Hence, it could be rationally expected that, whenever applicable, “community government” would be used. In fact, except in very rare cases, this is not so. Instead, centralized forms of government tend to migrate to all local levels carrying with them the kind of political party-oriented issues that are not necessarily the most relevant at the local community level.

5) Evolutionary stability of the reciprocator trait is very much dependent on social norms and transmission of culture; it is a trait that depends as much on genetics as on culture. Some direct evidence of this comes from the fact that experimental games played by adults and young children have different results. Culturally inherited traits may have a much faster dynamic than gene-based ones. If the reciprocator trait has a high cultural component, it is critical to understand

how modern society might be acting on or modifying it. A considerable loss of cooperative behavior might change society in many unexpected ways. Could less altruism come along with less hostility to strangers? If contemporary man were becoming more like *Homo economicus*, maybe it would not be necessary to rewrite the classical economy books.

## References

- [1] S. Bowles and H. Gintis, "The Evolution of Reciprocal Preferences," *Santa Fe Institute Working Papers* 00-12-072, 2000. <http://www.santafe.edu/media/workingpapers/00-12-072.pdf>.
- [2] S. Bowles and H. Gintis, "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations," *Theoretical Population Biology*, 65(1), 2004 pp. 17–28.
- [3] S. Bowles, J.-K. Choi, and A. Hopfensitz, "The Co-evolution of Individual Behaviors and Social Institutions," *Journal of Theoretical Biology*, 223(2), 2003 pp. 135–147.
- [4] S. Bowles and H. Gintis, *A Cooperative Species: Human Reciprocity and Its Evolution*, Princeton, NJ: Princeton University Press, 2011.
- [5] R. Vilela Mendes, "Network Dependence of Strong Reciprocity," *Advances in Complex Systems*, 7(3–4), 2004 pp. 357–368.
- [6] R. L. Carneiro, "A Theory of the Origin of the State," *Studies in Social Theory*, 3, 1977 pp. 3–21.
- [7] R. L. Carneiro, "From Autonomous Villages to the State: An Irresistible Trend in the Grand Sweep of Human History," *General Semantics Bulletin*, 72, 2005 pp. 15–19. <http://www.generalsemantics.org/wp-content/uploads/2011/04/gsb-72-carneiro.pdf>.
- [8] L. E. Grinin, "The Early State and Its Analogues," *Social Evolution History*, 2, 2003 pp. 131–176.
- [9] P. Turchin, "Arise Cliodynamics," *Nature*, 454, 2008 pp. 34–35. doi:10.1038/454034a.
- [10] P. Turchin, "Modeling Periodic Waves of Integration in the Afro-Eurasian World System," in *Globalization as Evolutionary Process: Modeling Global Change* (G. Modelski, T. C. Devezas, and W. R. Thompson, eds.), New York: Routledge, 2008 pp. 161–189.
- [11] A. Korotayev and D. Khaltourina, *Introduction to Social Macrodynamics: Secular Cycles and Millennial Trends in Africa*, Moscow: URSS, 2006.
- [12] Plato, *The Republic*, translated by A. D. Lindsay, Rutland, VT: C. E. Tuttle, 1992.
- [13] S. P. Magee, W. A. Brock, and L. Young, *Black Hole Tariffs and Endogenous Policy Theory: Political Economy in General Equilibrium*, Cambridge: Cambridge University Press, 1989.