A theory of norm compliance: Punishment and reputation

Shinji Teraji*

Department of Economics, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8514, Japan

ARTICLE INFO

ABSTRACT

Why do not social norms simply collapse from the violation? This paper analyzes two distinct mechanisms on norm compliance: punishment and reputation. The model considers two groups of agents in a society with one norm. Agents in one group choose whether to comply with the norm, while agents in the other group potentially punish opponents who violate the norm. The paper investigates two scenarios to account for the long-run stability of the norm. In one scenario, the norm is enforced due to a higher level of punishment of the violation. In another scenario, everyone is motivated due to reputation formation, despite a lower level of punishment by others. The interaction of two mechanisms provides a convenient way to norm compliance.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Societies have social norms or, for short, norms; members of the society are required to follow standards of behavior. Social norms are informal rules, as opposed to formal, legal rules promulgated by a court or a legislature. Social norms often direct individuals to undertake actions that are inconsistent with selfish actions. For example, in the dictator game, 50–50 division is generally viewed as norm-compliant (Andreoni and Bernheim, 2009).[1] People may deviate from such norms.[2] In the case of legal compliance, individual incentives most often refer to deterrence (Becker, 1968). That is, individuals are deterred from criminal activities by a higher fine and by a higher probability of conviction. Unlike legal rules, social norms are not supported by formal sanctions. Why do not social norms simply collapse from the violation? This paper studies two distinct mechanisms on norm compliance. The incentive to comply with norms derives not only from the enforcement of costly punishment by others, but also from reputation building for oneself.

The importance of decentralized punishments (i.e., punishments carried out by individuals without the intervention of a central authority) is documented in experimental studies. Ostrom et al. (1992) show the existence of such punishment opportunities in a common-pool resource use game. The fear of punishment has a positive effect on cooperation. In public goods experiments, subjects begin by contributing on average about half of their endowments to the public account. However, the level of contribution decays over the course of multiple rounds (Andreoni, 1995). When costly punishment is permitted, cooperation does not deteriorate. Fehr and Gächter (2000, 2002) indicate that many individuals are willing to punish unfair behavior at a personal cost in public goods games. Costly punishment is administered by "third parties" (Fehr and Fischbacher, 2004). Potential punishers are not themselves the victims but have merely witnessed unfair behavior. This is called "altruistic" punishment as individuals sacrifice for no direct benefits (Gintis et al., 2003). It suggests that cooperation has evolved through the sacrifice of altruistic punishers who are ready to incur some costs to prevent unfair behavior. People are fair because they have a psychological motivation to restore fairness.[3] Then, punishment can be seen as a consequence of a sense of fairness.

---

* Corresponding author. Tel.: +81 83 933 5554.
  E-mail address: steraji@yamaguchi-u.ac.jp

[1] The dictator game in theory gives rise to very inequitable distributions of resources. However, when the game is played for real, fair allocations figure prominently. Many game experiments offer abundant evidence that contradicts the hypothesis that all players are motivated only by their own material interest (see Camerer, 2003).

[2] In the Law and Economics literature (Polinsky and Shavell, 2000), agents obey the law only when their expected compliance utility (i.e., the payoffs expected to be obtained when they obey the law) is greater than their expected violation utility (i.e., the expected payoffs when they violate the law). In recent years, Law and Economics scholars have shown interest in the relations between social norms and the law (Posner, 2000).

---

[3] For example, Rabin (1993) examines concerns for fairness.

Certain groups of individuals can maintain a strong reputation over time. Akerlof (1980) develops an economic model to show that social norms that involve pecuniary disadvantage to individuals may persist without erosion.[4] Disobedience to the norm may involve a loss of reputation. People want to achieve the reputation of being fair. People are fair because they care about their reputation. They may not be genuinely fair. They have to be rewarded for good reputation, and they have to be willing to comply with the norm. Individuals are influenced in their convictions by what they think others will do.[5] Conformity to the norm is conditional on expectations about other people's behavior. Norms are constituted by expectations shared by members in a population and are jointly recognized among them.[6]

Thus, we have a set of solutions to the problem of norm compliance. The first solution is the punishment-based account. Following this account, people comply with norms because the threat of punishment makes it in their interest to do so. Altruistic punishment seems to have a solid foundation in human interaction. However, such costly punishment leads to a large increase in losses for altruistic punishers. The second solution is, on the other hand, the reputation-based account. Social norms can be sustained if the pecuniary advantage from breaking norms is not sufficient to offset the forgone reputation effect. This is related to indirect reciprocity. According to Alexander (1987), indirect reciprocity is arranged in the form of chain; a person is eventually helped by someone else who may not have been directly helped by him.[7] Altruistic actions can be sustained if people who support others receive support in return. To achieve such indirect reciprocity, building up a positive reputation is needed.[8]

This paper analyzes the interaction between the potential for costly punishment and building personal reputation. The model considers two groups of agents in a society with one norm. Agents in one group (group i) choose whether to comply with the norm by incurring some cost. They acquire utility from the reputation derived from complying with the norm. This utility depends positively on the proportion of motivated compliers. Individuals may differ in their motivation to comply with the norm. Punishment will be imposed on individuals who deviate from the norm. In the other group (group j), there are agents who value compliance and potentially punish non-compliance (i.e., the sanctioning individuals). The paper investigates individual punishment decisions. Agents choose to punish violators at some cost (decentralized punishment).

This paper asks how individual values evolve endogenously over time and analyzes the long-run dynamics of norm formation. The present framework systematically investigates the different forces to account for the long-run stability of the norm. There are two scenarios as follows. In one scenario, there is some possibility that the erosion of the reputation effect induces individuals to break the norm. Then, the norm is enforced due to a higher level of punishment of non-compliance. Punishment would be used to enforce the norm if a substantial fraction of people has little reputation-derived utility by obeying the norm. This paper, however, suggests that altruistic punishment may play a limited role in sustaining the norm. In another scenario, everyone is motivated due to reputation formation, despite a lower level of punishment by others. For a lower level of punishment, effective reputation building provides a way to sustain the norm.

The rest of the paper proceeds as follows. In Section 2, we present a simple model to consider the problem of norm compliance. Section 3 investigates two scenarios to account for the long-run stability of the norm. Section 4 concludes.

2. The model

Consider a society populated by a continuum of agents at each period of time t. The population size is constant over time and normalized to 1. We assume that the population is composed of two groups, i and j, which differ in characteristics. The first group i amounts to a share 1/2 of the whole population. Accordingly, the half part of the population belongs to group j. Matching between two groups take place randomly. A member of group i randomly matches with the opponent of group j at each period. This may suggest a large society in which one-shot encounters with unrelated strangers are common and information is rarely transparent.

There is one norm in the society. For simplicity, we assume that it is only possible to either follow the norm or not. An agent of group i ('he') chooses an action $x \in \{0, 1\}$ at each period. His action is represented by a discrete variable, one or zero. That is, $x = 1$ if the agent of group i complies with the norm, and $x = 0$ if he violates it. Inertia is introduced with the assumption that every agent of group i cannot switch actions at each point in time. He must make a commitment to a particular action in the short run. Opportunities to switch actions arrive randomly; some fraction α, $0 < \alpha < 1$, of individuals is drawn randomly from group i and makes a new choice of either $x = 1$ or $x = 0$. Thus, we may interpret a norm as a prescription indicating how a person ought to behave at any situation at which he may be called to move.

An agent in group i may deviate from the norm, but this deviation is costly. Punishment (such as ostracism) will be imposed on group i members who deviate from the norm. If deviation from the norm is observed ($x = 0$), the opponent of group j ('she') decides whether to punish the deviator, i.e., chooses p on the closed interval [0,1]. Here, with probability p the agent of group j punishes non-compliance, and with probability $1 - p$ she does not. The agent of group j is only an outside party who happens to know that norm violation has occurred. Altruistic punishment is motivated to restore norm compliance even though they are not expected to interact again in the future. It is costly for agents of group j to punish norm violators. Thus, in the model, agents in group i choose whether to comply with the norm, while agents in group j value compliance and potentially punish violators. Punishment is then confined to interactions with others that share the same norm in the population.

Individuals in group i are assumed to be heterogeneous with respect to their social concerns. We denote an agent's type of group i as g. An agent of type $g = 0$ is not concerned with the social meaning of a certain action. A higher g implies higher social concerns. The distribution of g is assumed to be exogenous and uniform in the model. Let the uniform distribution of g be $F(g)$, with $g \in [0, G]$. Furthermore, let the density of the uniform distribution be $f = 1/G$.

In group i, the short-run utility function of an agent of type g is given by

$$U = x\{R(g, \mu) - D\} - (1 - x)pC. \tag{1}$$

Each individual in group i is assumed to maximize the utility function (1), which constitutes the short-run equilibrium. The first term

---