Environment for Development

Discussion Paper Series

February 2013 ■ EfD DP 13-04

On Social Sanctions and Beliefs

A Pollution Norm Example

Jorge H. Garcia and Jiegen Wei

Environment for Development

The **Environment for Development** (EfD) initiative is an environmental economics program focused on international research collaboration, policy advice, and academic training. It supports centers in Central America, China, Ethiopia, Kenya, South Africa, and Tanzania, in partnership with the Environmental Economics Unit at the University of Gothenburg in Sweden and Resources for the Future in Washington, DC. Financial support for the program is provided by the Swedish International Development Cooperation Agency (Sida). Read more about the program at www.efdinitiative.org or contact info@efdinitiative.org.

Central America
Research Program in Economics and Environment for Development in Central America
Tropical Agricultural Research and Higher Education Center (CATIE)
Email: efd@catie.ac.cr

China
Environmental Economics Program in China (EEPC)
Peking University
Email: EEPC@pku.edu.cn

Ethiopia
Environmental Economics Policy Forum for Ethiopia (EEPFE)
Ethiopian Development Research Institute (EDRI/AAU)
Email: eepfe@ethionet.et

Kenya
Environment for Development Kenya
Kenya Institute for Public Policy Research and Analysis (KIPPRA)
University of Nairobi
Email: kenya@efdinitiative.org

South Africa
Environmental Economics Policy Research Unit (EPRU)
University of Cape Town
Email: southafrica@efdinitiative.org

Tanzania
Environment for Development Tanzania
University of Dar es Salaam
Email: tanzania@efdinitiative.org

Contents

On social sanctions and beliefs: A pollution norm example

Jorge H. García and Jiegen Wei[1]

Abstract

A prevailing view in the literature is that social sanctions can support, in equilibrium, high levels of obedience to a costly norm. The reason is that social disapproval and stigmatization faced by the disobedient are highest when disobedience is the exception rather than the rule in society. In contrast, the (Bayesian) model introduced here shows that imperfect information causes the expected social sanction to be lowest precisely when obedience is more common. This, amongst other findings, draws a distinct line between social and moral sanctions, both of which may depend on others' behavior but not on action observability.

Key words: Social Interactions, Social Norms, Asymmetric Information

JEL: D82, K42, L51

[1]García (corresponding author) is at the Centre for International Climate and Environmental Research - University of Oslo (CICERO); Gaustadalléen 21, 0349 Oslo, Norway; Tel/Fax +47 22841703; Email: j.h.garcia@cicero.uio.no. Wei is at the Department of Economics, University of Gothenburg, Sweden. We have benefited from discussions with Peter Berck, Fredrik Carlsson, Shachar Kariv and Asa Lofgren, Thomas Sterner, Elias Tsakas and seminar participants at various universities. García would like to thank the hospitality of UC Berkeley where he started working on this paper.

1 Introduction

It is widely recognized that social norms are important drivers of the behaviors of individuals and organizations (Kübler 2001;Young, 2005). Actions regarded by one's social group as proper can bring rewards and have positive effects on reputation. On the other hand, breaching a social norm may lead to sanctions and losses of reputation in a society that instills feelings of shame and distress in its deviants (Elster, 1989; Kaplow and Shavell, 2007).[1] This paper studies the effects of reputation on compliance with social norms of behavior, and in particular, the role of information in mediating this relationship. It has been argued that social sanctions imposed on managers and owners of polluting firms can provide a mechanism for internalization of external costs and damages. Cropper and Oates (1992) suggest in their survey of environmental economics that public opprobrium may explain the Harrington Paradox (HP) in the US, i.e, firms' high levels of compliance with environmental regulation under low expected penalties (Harrington, 1988). Similarly, Elhauge (2005) argues extensively about the relevance of social sanctions in influencing managers' decisions to undertake environmental investments. Decision makers would rather incur costs of compliance than face stigmatization and losses in reputation in society.[2]

[1] Social norm examples studied in the economics literature include an individuals' decision to actively look for a job (Clark, 2003; Lindbecket et al. 1999), to engage in criminal activities (Patacchini and Zenou, 2012), or to ration the exploitation of a common pool resource (Sethi and Somanathan, 1996; Ostrom, 1990). Some of these examples are consistent with the view that social norms often emerge as society's reaction to compensate for market failure (Arrow 1971).

[2] In a special report on business and climate change, *The Economist* (June 2nd., 2007) explains that the current shift toward cleaner energy might be due to two factors: moral(social) pressure and

The idea that the levels of social sanctions are relatively high when disobedience is uncommon allows a high-compliance state to qualify as an equilibrium; see Akerlof (1980), Bernheim (1994), and Lindbeck et al (1999). It is argued here that the potential disgrace of violating a well-established code of behavior may be significant, and that this constitutes a strong deterrent. However, the social sanction approach does not necessarily give a unique prediction of the equilibrium. Low compliance equilibria could coexist since losses of reputation are expected to be low at high levels of disobedience. Nyborg and Telle (2004) and Lay et al. (2003) formalize this notion in the case where firms are expected to meet an environmental standard.

An underlying assumption that seems ubiquitous in the study of social sactions is that of perfect observability of agents' behavior, for example, in terms of their emissions and compliance status. We argue that, unlike other situations where social sanctions have been used to explain economic behavior, this assumption is not necessarily met in the industrial pollution case. In fact, social sanctions are generated in different environments and firms' individual actions and compliance status are unlikely to be perfectly observable in the social circles where owners and managers interact.[3] In some cases, awareness of the identity of polluting sources

economic pressure: "*Businessmen, like everyone else, want to be seen to be doing the right thing, and self-interest points in the same direction.*" This paper is concerned with the social approval explanation. The economic explanation is associated with green consumerism. For theoretical analysis of markets with environmentally aware consumers, see, for instance, Amacher et al. (2004), and Cremer and Thisse (1999). Baron (2009) introduces a theory of moral management where firms face (economic) pressure from product markets, investors and green activists.

[3]The relevance of action observability to the imposition of social sanctions has been acknowledged by several authors; see for instance, Elster (1989) and Kaplow and Shavell (2007).

may be limited to neighboring communities; even for these, it may difficult to judge whether a given emitter is in or out of compliance with the legislation. Levin and List (2007) and Fershtman et al. (2008) explain that whether a norm is activated or not depends on the characteristics of the "situation," which directly relates to the social spheres of our pollution example. While the above discussion concurs with this view, we emphasize here that, although a norm might be activated, actions could be imperfectly observable.

This paper presents a theory of social interactions with a rich informational structure. In our model, society forms (Bayesian) beliefs (or expectations) about the compliance status of individual firms based on two pieces of information: the general level of violation in the society, and signals that can convey some indication of firms' compliance status. Managers' beliefs and expected losses of reputation are in turn built on society's beliefs. It is further assumed that a single firm's action can not affect any given outcome or social equilibrium. While an environmental norm example is used to introduce our analysis, the results are applicable to a wider range of cases where similar social interactions come into play. They also highlight the role of moral (self-imposed) sanctions, which may depend on others' behavior but not on action observability.

In Section 2, the model is presented and solved for both perfect and imperfect information structures. Section 3 discusses the main results and concludes the paper. Appendix A presents some partial results omitted in the body of the text and Appendix B contains the proofs of the three theorems and the lemma introduced in Section 2.

2 A model of reputation and compliance

The social norm in our model demands that firms meet a legal pollution standard. Compliance is costly but non-compliance could lead to a loss in reputation which may also be costly. In order to recreate the HP scenario we assume that regulatory costs due to non-compliance are negligible or nonexistent. As stated earlier, the main feature of social sanctions is that agents' pay-off functions depend not only on their own action but also on other agents' actions. In a setting where the number of agents that follow a norm is relatively large, social disapproval due to deviation is high. Correspondingly, if very few agents follow the norm, costs of deviation are small.[4] Let $\alpha \in [0,1]$ represent the fraction of firms that violate the standard. The loss in reputation function is $R(\alpha)$, where $R_\alpha < 0$. By breaking the norm, violators derive pecuniary benefits represented by saved abatement expenditures a.

We will only be concerned with situations where firms adopt pure strategies, either comply or violate. Let $x \in \{c, v\}$ be a firm's strategy, where c denotes compliance and v violation. A manager's utility function is then given by:

$$U(x;\alpha) = \begin{cases} -a & \text{if} \quad x = c \\ -R(\alpha) & \text{if} \quad x = v \end{cases} \tag{1}$$

Note that society, although it has the capacity to sanction violators, is not a player as such in this model. An underlying assumption of the managers' utility

[4] In this model society rewards conformity to the prevailing patterns of behavior. While we subscribe to this behaviorist interpretation, $R(\alpha)$ may also be explained by limited capacity or decreasing returns in the punishment technology.

function in equation (1) is that of perfect observability of firms' behavior. The social sanction faced by managers is to a large extent given by society's beliefs concerning their compliance status. Under perfect information society's assessment of a given firm's compliance status matches the firm's actions. In order to make our point clear, we use the simple linear reputation function, $R(\alpha) = 1 - \alpha$. Furthermore, assume that there is a unit mass of firms with homogeneous fixed costs of compliance $a \in (0, 1)$ and that a single firm's actions do not affect the value of $R(\alpha)$.[5] This description fits that of perfect competition (or non-atomic games). In the analysis of the strategic interactions in our model, the following Nash Equilibrium (NE) concept will be used:

Definition 1. *Let $x(\alpha)$ be a firm's best response strategy to level of violation α, so that $U(x(\alpha); \alpha) \geq U(x; \alpha)$ for $x \in \{c, v\}$. A strategy profile α is a NE if all firms' strategies are best response strategies. Further, a NE is Stable if there is $\bar{\epsilon}$ such that $x(\alpha) = x(\alpha \pm \epsilon)$ holds for all $\epsilon \in (0, \bar{\epsilon})$ and for all firms.*[6]

This definition presents a natural extension of NE for N-player games to a game with a continuum of players.[7] The stability condition ensures that equilibrium strategies are also best response strategies to levels of violation that slightly differ

[5] The framework proposed here is also illustrative of situations where pro-social behavior is rewarded. When the reward function is given by $1 - \alpha$, so that compliant agents experience more satisfaction when compliance is more common, the three propositions and the lemma derived below still hold.

[6] Naturally, the stability condition is one sided for the extreme cases, $\alpha = 0, 1$. The best responses must, respectively, meet $x(0) = x(0 + \epsilon)$ and $x(1) = x(1 - \epsilon)$ for all $\epsilon \in (0, \bar{\epsilon})$ and for all firms.

[7] Schmeidler (1973) first proved existence of pure strategy equilibrium in games with a continuum

from equilibrium so that small masses of firms do not have incentives to deviate. Also, if a small mass of firms makes a mistake in equilibrium, the remaining set of firms will not change their original strategies.

Proposition 1 (Perfect Information Equilibria). *Under perfect information concerning firms' compliance status, two Stable NE coexist: the full compliance equilibrium, $x(0) = c$ for all firms, and the full violation equilibrium, $x(1) = v$ for all firms. A third Non-Stable NE with partial compliance, $\alpha = 1-a$, is also present.*[8]

Figure 1 illustrates the insight provided by this proposition by showing the (dis)utilities of compliance and violation for different levels of violation. Proposition 1 presents two Stable NE, namely states k and m in the figure, where all firms behave identically. The social sanction at high levels of compliance is high enough to keep this society in full compliance, state k. Nevertheless, the compliance incentives are undermined at low levels of compliance in such a way that a violation equilibrium could persist, state m. State l emerges as a possible NE but it does not meet the stability requirement.

[Figure 1 about here]

Society's attitude toward pollution in the above analysis contrasts with the traditional view used to study the industrial pollution control problem. The existence

of players. For a comprehensive account of this class of games, see Khan and Sun (2002).

[8] This proposition is the equivalent of Proposition 1 of Nyborg and Telle (2004)

of increasing marginal damages of pollution implies that the optimal pressure imposed by society on polluting firms ought to be increasing in pollution. While we do not attempt to develop a normative theory of pollution here, it is interesting to see that, under a behavioristic lens, society might be more tolerant of pollution at higher levels of environmental degradation.[9] In our model, higher levels of violation are naturally associated to higher levels of pollution.

We now turn to study the imperfect information case. We assume that society has fragmentary information based on which it forms expectations about the compliance status of firms. Because beliefs are now formed with partial information, losses in reputation could be imputed to both compliant firms and violators. We assume that society knows the actual level of violation in the economy α. This in fact constitutes society's (prior) belief about any given firm being in violation.[10] Further, although society does not observe the compliance status of firms, it does receive a signal from each manager that conveys information about their actions. A signal could be denoted as either a violation signal or a compliance signal. Signals are mutually exclusive and the occurrence of a compliance signal is equivalent to the non-occurrence of a violation signal. Let $\theta \in (0,1)$ be the probability that society

[9] When marginal environmental damage is given by $D(\alpha)$ with $D(0) > a$ and $D_\alpha > 0$, it is clear that full compliance generates the largest social surplus. Note, however, that all levels of violation are Pareto efficient.

[10] Assume compliant firms emit 0 and violating firms emit z units of pollution. Since the number of firms is normalized to unity, if they were all noncompliant, total pollution would be "z." If total pollution can be observed and is measured as W, then the statistic used by society to calculate the share of polluting firms is given by $\tilde{\alpha} = \frac{W}{z}$. Alternatively, when the frequency of violation signals, i.e. $f = \pi\alpha + \theta(1-\alpha)$, is observable, this statistic can take the following form $\tilde{\alpha} = \frac{f-\theta}{\pi-\theta}$.

receives a violation signal from a compliant firm and π be the probability that such signal comes from a violator with $\pi \in [\theta, 1)$, that is society cannot be less (more) likely to receive a violation (compliance) signal from a violator than from a compliant firm. Consequently, $1-\pi$ and $1-\theta$ are the probabilities that a compliance signal is received from a violator and a compliant firm respectively. Note that these primitive probabilities are exogenous and firms cannot influence them.[11] Once signals are realized, society's beliefs on the compliance status of firms are calculated using Bayes' rule. Specifically, society's beliefs about an individual firm being in violation when a violation signal is received take the following form:

$$A(\alpha, \pi) = \frac{\pi}{\pi\alpha + \theta(1-\alpha)} \ \alpha \tag{2}$$

Without loss of insight, θ is assumed invariant through most of the analysis and was omitted in $A(\alpha, \pi)$. In fact, increases (decreases) in π can always be interpreted as decreases (increases) in θ in this type of models. Society's prior belief α is updated via the ratio factor given by the first part the expression. When signals are uninformative, that is when $\pi = \theta$, the updating factor equals 1 for all values of $\alpha \in [0, 1]$. With informative signals, that is when $\pi > \theta$, this factor is higher than 1 for $\alpha \in [0, 1)$ and equal to 1 for $\alpha = 1$. Equation (2) provides society with an estimate of the probability that a received violation signal comes from a violator after correcting for the fact that violation signals could also come from non-violators.

[11] Lyon and Maxwell (2011) introduce a (market interaction) model where firms make strategic use of environmental information but may be punished by activists for lying. Society's knowledge about polluters in our model model resembles that of the regulator's in a non-point source pollution problem.

Society's beliefs about the violation strategy when a compliance signal is received take the following form:

$$B(\alpha,\pi) = \frac{(1-\pi)}{(1-\pi)\alpha + (1-\theta)(1-\alpha)}\ \alpha \tag{3}$$

In this case, the updating factor with informative signals is lower than 1 for $\alpha \in [0,1)$ and equal to 1 for $\alpha = 1$. It thus follows that $A(\alpha,\pi) > \alpha > B(\alpha,\pi)$ for $\alpha \in (0,1)$ when signals are informative. The probability that a firm is in violation is higher when it emits a violation signal than when it emits a compliance signal. When there is either total violation, $\alpha = 1$, or total compliance, $\alpha = 0$, signals become irrelevant and society is fully certain about all firms strategies: $A(0,\pi) = B(0,\pi) = 0$ and $A(1,\pi) = A(1,\pi) = 1$. When signals are uninformative firms are completely anonymous and the level of violation, α, is the most sensible estimate of the chances that any given firm is in violation: $A(\alpha,\pi) = B(\alpha) = \alpha$. Note that the realized sanction when a violation signal is emitted is $A(\alpha,\pi)R(\alpha)$ and $B(\alpha,\pi)R(\alpha)$.

Firms make their compliance decisions taking into account their own expectations of being identified as violators. Unlike society, managers know their own actions. Firms' unconditional expectations of being identified as violators when in compliance and in violation are given by the following expressions:

$$f^v(\alpha,\pi) = \pi A(\alpha,\pi) + (1-\pi)B(\alpha,\pi) \tag{4}$$

$$f^c(\alpha,\pi) = \theta A(\alpha,\pi) + (1-\theta)B(\alpha,\pi) \tag{5}$$

Figure 2 shows the form these beliefs take under perfect and imperfect information. The solid curves represent firms' unconditional beliefs whereas the dashed curves represent society's beliefs. With uninformative signals, $f^c(\alpha,\pi) = f^v(\alpha,\pi) =$

α (see Figure 2a). With informative signals, $f_v(\alpha,\pi) > \alpha > f_c(\alpha,\pi)$ for $\alpha \in (0,1)$ (see Figure 2b). That is, signals allow compliant managers to decrease the chances of being identified as violators, whereas violating managers see these chances increase. In fact, Appendix A indicates that $f^c_\pi(\alpha,\pi) < 0$ and $f^v_\pi(\alpha,\pi) > 0$ for $\alpha \in (0,1)$. Because signals are irrelevant in the extreme cases $f^c(0,\pi) = f^v(0,\pi) = 0$ and $f^c(1,\pi) = f^v(1,\pi) = 1$.[12] In the perfect information case society's beliefs always match firms' actual behavior in such a way that only violators face losses in reputation (see Figure 2c).

[Figure 2 about here]

We started by looking at certain losses in reputation with perfect information and then turned to probabilities of violation detection with imperfect information. We are now in a position to synthesize and look at expected losses in reputation. These are now given by $f^v(\alpha,\pi)\,R(\alpha)$ for the violation strategy and $f^c(\alpha,\pi)\,R(\alpha)$ for the compliance strategy. Following the notation used in equation (1), managers' expected utility is:

$$U^E(x;\alpha,\pi) = \begin{cases} -f^c(\alpha,\pi)\,R(\alpha) - a & \text{if} \quad x = c \\ -f^v(\alpha,\pi)\,R(\alpha) & \text{if} \quad x = v \end{cases} \tag{6}$$

Ultimately, managers make decisions based on the difference in expected losses in reputation and how this relates to abatement costs. Let us denote the difference

[12] Firms in violation can be unveiled with a probability $f^c < 1$ but firms in compliance may be wrongly perceived or accused of violating with probability $f^c > 0$. This is sometimes referred to as errors of type I and II.

in expected losses in reputation between the violation and the compliance strategies by the following function:

$$F(\alpha, \pi) = \Big[f^v(\alpha, \pi) - f^c(\alpha, \pi) \Big] R(\alpha) = f(\alpha, \pi) R(\alpha) \tag{7}$$

When $F(\alpha, \pi) > a$, the compliance strategy dominates the violation strategy. From the properties of $f^v(\alpha, \pi)$ and $f^c(\alpha, \pi)$, it directly follows that $F_\pi > 0$ for $\alpha \in (0, 1)$. That is, an increase in the accuracy of signals makes the compliance strategy more attractive. Further, $F(0, \pi) = F(1, \pi) = 0$. Lemma 1 presents other important properties of the difference in expected utilities.

Lemma 1. *When signals are informative, that is $\pi > \theta$, there exists $\hat{\alpha} \in (0, \frac{1}{2})$ such that $\hat{\alpha} = \operatorname{argmax} F(\alpha, \pi)$. Further $F_\alpha > 0$ for all $\alpha \in (0, \hat{\alpha})$, $F_\alpha = 0$ for $\alpha = \hat{\alpha}$, and $F_\alpha < 0$ for all $\alpha \in (\hat{\alpha}, 1)$.*

Starting at full compliance, as the proportion of violators α increases, signals become less coarse, thus increasing the difference in expected losses in reputation $F(\alpha, \pi)$ and managers' incentives to adopt a compliance strategy. At the same time, however, a decreasing loss in reputation, $R(\alpha)$, would have the opposite effect. This effect is reinforced and dominates at much higher levels of compliance when signals become coarse again. The social equilibria that may emerge under imperfect information are described in Proposition 2.

Proposition 2 (Imperfect Information Equilibria). *Under imperfect information about firms' compliance status we have that:*

- *The full violation state is a Stable NE, that is,* $x(1) = v$ *for all firms, whereas the full compliance state does not qualify as a NE.*

- *Two NE with partial compliance exist if and only if* $F(\hat{\alpha}(\pi), \pi) > a$ *with* $\frac{dF}{d\pi} > 0$. *The higher compliance equilibrium* α^k *is Stable, while the lower compliance equilibrium* α^l *is Non-Stable. Further,* $\alpha^k_\pi < 0$, $\alpha^k_a > 0$, $\alpha^l_\pi > 0$ *and,* $\alpha^l_a < 0$.

The first part of the proposition follows from the Bayesian belief formation. Because beliefs are completely accurate when there is full violation, the pay-offs in the perfect and imperfect information cases are exactly the same. The full violation state is thus preserved as a stable equilibrium under imperfect information. On the other hand, an important consequence of the existence of imperfect information is the ruling out of full compliance as a possible equilibrium. Note that the expected losses in reputation due to violation are zero at full compliance under imperfect information. In a society where most people conform, people find it hard to conceive that anyone would be in disobedience.

Figures 3a, 3b and 3c help illustrate the possible emergence of partial compliance equilibria. Appendix A presents the second order condition that ensures that losses in reputation for the violation action are concave with respect to α. This starts at zero, because the risk of being unveiled is zero when no one violates. The function will rise as detection risk rises until a maximum when the effect of a decreasing $R(\alpha)$ sets in. The expected costs of compliance function is also concave (See Appendix A) and follows a similar pattern but naturally it does not fall below the costs of compliance, a. When signals are uninformative (Figure 3a), the losses in reputation faced

by compliant and violating managers are the same. Since obedient firms also incur a compliance cost, disobedience is the only best strategy for the manager at all levels of violation. As signals become informative (Figures 3b and 3c) the expected costs of violation typically increase, while the expected costs of compliance decrease. Note that the partial compliance equilibrium emerges only when the maximum possible difference between expected losses in reputation are actually higher than abatement costs a. From the discussion above on belief formation, it is clear that, in both, the full compliance and full violation states, expected utilities are not sensitive to signals: $U^E(v;0,\pi) = U^E(v;1,\pi) = 0$ and $U^E(c;0,\pi) = U^E(c;1,\pi) = -a$ because $f^c(0,\pi) = f^v(0,\pi) = R(1) = 0$.

[Figures 3a,b,c about here]

Obtaining an analytical solution for the condition $F(\hat{\alpha}(\pi),\pi) > a$, introduced in Proposition 2, is virtually impossible. On the other hand, by fixing $\theta = \frac{1}{2}$, we were able to establish an intuitive sufficient condition for the emergence of interior equilibria (the derivation is algebraically involved and is omitted here for brevity but is available from the authors). In particular, if $\pi > \frac{1}{2} + \frac{\sqrt{7a}}{2}$, two interior equilibria exist.[13] This expression has some interesting characteristics. Note that π is higher than $\theta = \frac{1}{2}$ and is increasing in abatement costs, a. Because $\pi < 1$, it can also easily be concluded that, for $a > \frac{1}{7}$, no interior equilibrium can emerge.

[13] We also established that $\pi > \frac{1}{2} + \frac{\sqrt{5a}}{2}$ is a necessary condition for the emergence of interior equilibria. The necessary and sufficient condition thus has the following form: $\pi > \frac{1}{2} + \frac{\sqrt{Na}}{2}$ with $N \in (5,7)$.

The last part of Proposition 2 states that, as the violation signal from non-compliant managers π becomes more precise, the high compliance equilibrium k, moves toward full compliance, while the low compliance equilibrium l moves toward the full violation state. A similar pattern occurs if the abatement costs a are reduced. Figure 3c shows that the equilibrium state k has moved, in relation to the perfect information case, to the interior of $\alpha \in [0, 1]$. Note also that although equilibrium l has been preserved in its original form (Non-Stable), it now occurs at higher levels of violation. While a high compliance equilibrium may be attainable under imperfect information, it requires a relatively low compliance costs and a relatively high level of accuracy of signals. The following proposition presents how the different equilibrium points behave as signals become extremely informative.

Proposition 3 (Almost Perfect Information Equilibria). *When information is almost perfect, and independent of costs of compliance, partial compliance equilibria* α^k *and* α^l *(Proposition 2) emerge. Further, as* $\pi \to 1$ *and* $\theta \to 0$*, we have that* $\alpha^k \to 0$ *and* $\alpha^l \to 1 - a$*. In this sense, social equilibria under perfect information are limiting situations of social equilibria under imperfect information.*

An increase in the preciseness of signals drive both interior equilibria to different limit points. With society almost certainly receiving a violation signal from a violator and a compliance signal from a compliant manager, $\pi \to 1$ and $\theta \to 0$, the stable high compliance equilibrium α^k will get infinitely close to the stable full compliance equilibrium under perfect information, while the non-stable low compliance equilibrium α^k moves infinitely close to the unstable equilibrium $1 - a$ under per-

fect information. As shown in graphs 3a, 3b and 3c, as signals become informative expected utilities tend to resemble perfect information utilities for $\alpha \in (0, 1]$.

2.1 Social opprobrium based on the frequency of violation signals

We have assumed that the loss of reputation function depends solely on the aggregate level of violation. This implies a degree of separability in expected payoffs that has been instrumental in our derivations. If the social sanction depends on the accuracy of signals, $R(\alpha, \pi)$, the conclusion from the analysis above might not be valid any more. In order to extend the analysis, we opt to study the sensible case in which the loss of reputation depends on the (observed) frequency of violation signals $\pi\alpha + (1-\alpha)\theta$. The reputation function can thus be represented as $R(\alpha, \pi) = 1 - [\pi\alpha + (1-\alpha)\theta]$. Note here that, as signals become very informative, losses in reputation tend to losses in reputation under the perfect-information benchmark scenario, that is $R \to 1 - \alpha$. Under the new assumption, and even though some of the results under imperfect information differ from our previous analysis, our most important conclusions still hold. We are in fact still able to find $\hat{\alpha}$ similar to Lemma 1, so that $F_\alpha > 0$ for all $\alpha \in (0, \hat{\alpha})$, and $F_\alpha < 0$ for all $\alpha \in (\hat{\alpha}, 1)$, but are not precise about its location, which largely depends on π and θ. For the same reasons that are behind Proposition 2, we also find under the new assumption that full violation is an equilibrium while full compliance is not. However, (interior) social equilibria may respond differently to changes in the accuracy of signals. We explained earlier that, when the loss in reputation function is independent of signals,

the benefit of complying the norm increases with the improvement of signal quality, i.e. $f_\pi(\alpha,\pi)R(\alpha) > 0$. When the reputation loss function depends on signals, as the quality of signals improves, the compliance strategy is more appealing due to the increased chances of being unveiled as a violator $f_\pi(\alpha,\pi)R(\pi\alpha+(1-\alpha)\theta) > 0$ but, at the same time, it is less interesting due to a laxer reputation requirement in society, $f(\alpha,\pi)R_\pi(\pi\alpha+(1-\alpha)\theta) < 0$. If the additional gains do not compensate the costs, the high compliance equilibrium moves away from full compliance, i.e. $\alpha_\pi^k > 0$. However, if signals are precise enough, and high partial compliance equilibrium is close to full compliance equilibrium, we find $\alpha_\pi^k < 0$. Similarly to Proposition 3, this will bring the high partial compliance equilibrium toward full compliance. The same argument applies to the lower compliance equilibrium

2.2 Alternatives to Bayesian updating

Using experimental data, El-Gamal and Grethner (1995) report that, while Bayes rule was the most important rule used by individuals, a representativeness rule (overweighting the signals) and conservativism (overweighting the prior) were also implemented, but to a lesser extent. In this section we explore the equilibrium configurations that may emerge in societies that use these (non-statistical) belief formation rules. We maintain the basic elements of the model so that Proposition 1 holds and can be used as a benchmark. In the characterization of imperfect information equilibria, and for ease of exposition, we assume that compliant managers do not emit violation signals ($\theta = 0$).

A society that uses the representative rule takes signals at face value. Regard-

less of the prior and the likelihood of signals, society presumes who managers emit violation (compliant) signals are in violation (compliance). These managers are not given the benefit of the doubt and bear the full weight of social punishment $R(\alpha)$. The manager's expected payoff is thus given by $\{-a, -\pi R(\alpha)\}$ for the compliance and the violation decisions. It is easy to see that, when the probability of emitting a violation signal is not high enough in relation to the costs of compliance, that is, when $\pi < a$, full violation is the only equilibrium. On the other hand, when $\pi \geq a$, full compliance and full violation equilibria coexist (a non-stable compliance equilibrium also occurs at $\alpha = \frac{\pi - a}{\pi}$). Notably, what some may interpret as "limited rationality" has the potential to preserve the perfect information equilibria configuration of Proposition 1. Similar qualitative results hold when social punishment depends on the frequency of violation signals so that $R(\alpha, \pi) = 1 - \pi\alpha$. Managers' expected utility would then be given by $\{-a, -\pi\, R(\pi\alpha)\}$ for the compliance and the violation decisions. It is easy to see that, when $\pi - \pi^2 < a < \pi$, full compliance and full violation are stable equilibria (non-stable compliance will occur at $\alpha = \frac{a}{\pi - \pi^2}$). However, when $\pi - \pi^2 \geq a$, full compliance is the only equilibrium in the game.

We interpret the case where society uses the conservative rule as having the same structure as the Bayesian belief formation model with completely uninformative signals. In fact, in both situations, signals are irrelevant and the only piece of information used by society is the prior. Specifically, manager's expected payoffs are $\{-a - \alpha\, R(\alpha), -\alpha\, R(\alpha)\}$ and it is clear that the only possible equilibrium is this game is full violation. Note that it makes no sense to consider conservatism in the case where social opprobrium depends on the frequency of violation signals.

In a more general setting, one may envision society as consisting of three types of individuals: Bayesian individuals, those using the representativeness rule, and conservatives. Firms' expected losses in reputation will then be the weighted sum of losses in reputation induced by each type. A difficulty with this approach is that a complete characterization of equilibrium is not granted. In fact, the difference in expected losses in reputation between the violation and the compliance strategies will fail to meet Lemma 1 for a number of societal and probability of signals configurations.

2.3 Heterogeneous managers

Assuming homogenous managers has allowed us to convey the basic message on the divergence between the social sanction and the beliefs functions and its consequences on compliance. This section briefly illustrates the equilibrium configurations in a world with managers (firms) with heterogeneous costs of compliance. Assume the same punishment function $R(\alpha)$ and a unit mass of firms with costs of compliance uniformly distributed in the range $[\underline{a}, \overline{a}]$, with $0 < \underline{a} < \overline{a} < 1$ and *cdf* function $G(a)$. It is easy to see that, under perfect information, full compliance and full violation equilibria coexist, because $R(0) > \overline{a}$ and $R(1) < \underline{a}$. A partial (unstable) compliance equilibrium also occurs when the proportion of firms whose costs of compliance are higher than the social penalty equals the proportion of non-compliant firms that generate that penalty, that is, when α such that $R(\alpha) = G^{-1}(1-\alpha)$.[14] Note

[14]Full compliance might be reached through an unraveling mechanism in which relatively low-cost managers complying with the environmental standard push up the social pressure face by disobedient managers. As firms shift to the compliance strategy, managers of the highest cost of

that $G^{-1}(1-0) = \overline{a}$ and $G^{-1}(1-1) = \underline{a}$. Regarding the imperfect information case, it should be first noted that the difference in expected losses in reputation between the violation and the compliance strategies $F(\alpha, \pi)$ is independent of costs. Similarly to Proposition 2, the full compliance state can no longer be an equilibrium, because $F(0, \pi) < \overline{a}$, while the full violation state is a stable equilibrium, because $F(1, \pi) > \underline{a}$. Interior equilibrium points, on the other hand, will occur whenever $F(\alpha, \pi) = G^{-1}(1-\alpha)$. Noting that $G^{-1}(1-\alpha) = (\overline{a} - \underline{a})(1-\alpha) + \underline{a}$, we obtain a similar result to that presented in Proposition 2, namely that a stable and an unstable interior equilibrium emerge with relatively high accuracy of signals.

3 Conclusions and discussion

The existing literature on social norms explains that, in a norm-abiding world, the loss in reputation from being caught "cheating" could be devastating and that this constitutes a strong force explaining why high levels of compliance might be preserved. In contrast, and using an industrial pollution example, we show here how imperfect information makes the expected loss of reputation due to violation the lowest precisely when compliance is relatively high. The linchpin of our argument is that the likelihood of being unveiled is a very different function from the loss off reputation function. In a society where most agents conform, it is hard to conceive or believe that anyone would be in disobedience, in particular when actions are not

compliance firms will find it less and less profitable to sustain a violation strategy.

Consistent with Herrendorf et al. (2000), under large cost heterogeneity, i.e. $\underline{a} < 0$ and $1 < \overline{a}$, a unique stable compliance equilibrium emerges at α such that $R(\alpha) = G^{-1}(1-\alpha)$.

fully observable. Consequently, the veil of anonymity drawn over violators becomes thicker as the proportion of firms that meet the standard increases. Imperfect information can also lead to mistakes in judgment so that compliant individuals could wrongly be stigmatized. These results suggest that a society where social pressure is somewhat unimportant could exhibit higher obedience than a society where social disapproval does play a more important role. This is so if the latter suffers more acute information asymmetries than the former. Due to the way beliefs are formed in our model, the loss in reputation functions due to violation in the perfect and imperfect information worlds are diametrically different at high levels of compliance. One may refer to this as a "belief curse."

To a certain extent, the "classical" environmental regulator can be viewed as an agent that solves an information asymmetry between polluters and the judiciary (Garvie and Keeler, 1994). In fact, its budget is spent in two different activities, namely monitoring and enforcement, or the actual process of prosecuting firms. If provision of information to the general public is relatively cheap, as seems to be the case with today's information technologies, the regulator could publicly disclose polluters' environmental indicators (e.g. Afsah et al. 2013) and make use of social sanctions (rewards) as a substitute for conventional enforcement. A drawback of this approach is that provision of information may not be enough to move society from low to high compliance equilibria. Additional incentives may be required so that a critical mass of law-abiding managers have the ability to trigger overall compliance. This, however, can only be achieved if the additional incentives, e.g. a green tax, do not erode underlying social motivations.

Although the discussion has focused on an industrial pollution example, the basic framework lends itself to study other situations where similar social interactions and information asymmetries are present. Direct examples may be found in the exploitation of (other) common property resources and the contribution to a public good. The "belief curse" of our model could also help us understand, for instance, the persistent presence of corruption in some societies. As Bardhan (1997) puts it "*...the tenacity with which it [corruption] tends to persist in some cases easily leads to despair and resignation on the part of those who are concerned about it.*" In this context, the social norm demands that public officials (and possibly private managers) not engage in corruption, whereas the costs of compliance with the norm are represented by the forgone bribery benefits. Since corruption activities are carried out behind closed doors, the most likely equilibrium in light of our model, is one in which most officials are corrupt and *society knows it* with certainty, but it *does not care*, i.e. the social sanction is very low. On the other hand, individuals may have internal motives to follow a certain norm (Smith, 1979; Kaplow and Shavell, 2007). It may also be the case that, although the individual's incentives to follow the norm depend on her peers' behavior, these incentives do not depend on observability. In some societies, it may suffice for an individual to know that most of her peers conform to deter her from breaking a social code. This is, in fact, the case of moral norms and this paper illustrates how valuable such norms may be.

Appendix A

Derivations are omitted but available from the authors.

$$\begin{aligned}
\frac{\partial f^v(\alpha,\pi)}{\partial \pi} &= \left(\frac{1-\pi}{1-\theta}\frac{\alpha}{1-\alpha}+1\right)^{-2} - \left(\frac{\pi}{\theta}\frac{\alpha}{1-\alpha}+1\right)^{-2} > 0 \quad \text{and} \\
\frac{\partial f^c(\alpha,\pi)}{\partial \pi} &= \frac{(1-\alpha)}{\alpha}\left[\left(\frac{1-\alpha}{\alpha}+\frac{\pi}{\theta}\right)^{-2} - \left(\frac{1-\alpha}{\alpha}+\frac{1-\pi}{1-\theta}\right)^{-2}\right] < 0 \quad \text{for } \alpha \in (0,1)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 f^v(\alpha,\pi)R(\alpha)}{\partial^2 \alpha} &= -\left[\pi\frac{2\theta\pi^2}{1-(1-\alpha)\theta-\alpha\pi} + (1-\pi)\frac{2(1-\theta)(\pi-1)^2}{(1-\alpha)\theta+\alpha\pi}\right] < 0 \quad \text{and} \\
\frac{\partial^2 f^c(\alpha,\pi)R(\alpha)}{\partial^2 \alpha} &= -\left[\theta\frac{2\theta\pi^2}{1-(1-\alpha)\theta-\alpha\pi} + (1-\theta)\frac{2(1-\theta)(\pi-1)^2}{(1-\alpha)\theta+\alpha\pi}\right] < 0 \quad \text{for } \alpha \in [0,1]
\end{aligned}$$

Appendix B

Proof Proposition 1. The proposition consists of three statements that are proven separately:

- $x(0) = c$ for all firms is a NE since $U(c;0) > U(v;0)$, which holds given the assumption $-a > -1$. The equilibrium is Stable since there always exists small enough ϵ such that $U(c;\epsilon) > U(v;\epsilon)$, that is $-a > -(1-\epsilon)$.

- $x(1) = v$ for all firms is a NE since $U(v;1) > U(c;1)$, which holds given the assumption $0 > -a$. The equilibrium is Stable since there always exists small enough ϵ such that $U(v;1-\epsilon) > U(c;1-\epsilon)$, that is $-(1-\epsilon) > -a$.

- $x(1-a) = v$ for a fraction $\alpha = 1-a$ of firms and $x(1-a) = c$ for the remaining population of firms is a NE since $U(v;1-a) \geq U(c;1-a)$ and $U(c;1-a) \geq U(v,1-a)$ hold simultaneously so that $U(c;1-a) = U(v;1-a) = a$. Suppose that a small mass of compliant firms ϵ deviate so that the new level of violation is $1-(a+\epsilon)$. Since $U(c;1-(a+\epsilon)) = -a < -[1-(\alpha+\epsilon)] = U(v;1-(a+\epsilon))$, the deviants' new best response is violation. Since this differs from their equilibrium response, that is compliance, the equilibrium is Non-Stable. Q.E.D.

Proof Lemma 1. Replacing equations (5) and (6) and $R(\alpha) = 1-\alpha$ into Equation (7), we obtain:

$$F(\alpha,\pi) = (\pi-\theta)(1-\alpha)\left[\frac{1}{1+\frac{\theta}{\pi}\frac{1-\alpha}{\alpha}} - \frac{1}{1+\frac{1-\theta}{1-\pi}\frac{1-\alpha}{\alpha}}\right] \tag{8}$$

Let $m = \frac{1-\alpha}{\alpha}$ so that

$$\begin{aligned}\frac{\partial F}{\partial \alpha} &= \frac{\partial m}{\partial \alpha}\frac{\pi-\theta}{(m+1)^2}\left[\frac{1-\frac{\theta}{\pi}m^2}{\left(1+\frac{\theta}{\pi}m\right)^2} - \frac{1-\frac{1-\theta}{1-\pi}m^2}{\left(1+\frac{1-\theta}{1-\pi}m\right)^2}\right] \\ &= \frac{\partial m}{\partial \alpha}\frac{(\pi-\theta)\left(\frac{\theta}{\pi}-\frac{1-\theta}{1-\pi}\right)m}{(m+1)^2\left(1+\frac{\theta}{\pi}m\right)^2\left(1+\frac{1-\theta}{1-\pi}m\right)^2}\left[\frac{\theta}{\pi}\frac{1-\theta}{1-\pi}m^3 - \left(1+\frac{\theta}{\pi}+\frac{1-\theta}{1-\pi}\right)m - 2\right]\end{aligned}$$

For $\alpha \in (0,1)$ we have that:

$\frac{\partial m}{\partial \alpha} = -\frac{1}{\alpha^2} < 0$ and $\frac{(\pi-\theta)\left(\frac{\theta}{\pi}-\frac{1-\theta}{1-\pi}\right)m}{(m+1)^2\left(1+\frac{\theta}{\pi}m\right)^2\left(1+\frac{1-\theta}{1-\pi}m\right)^2} < 0$

Let $f(m) = \frac{\theta}{\pi}\frac{1-\theta}{1-\pi}m^3 - \left(1+\frac{\theta}{\pi}+\frac{1-\theta}{1-\pi}\right)m - 2$

With informative signals, $\pi > \theta$, this function is such that

$$\begin{aligned}\lim_{m\to-\infty} f(m) &= -\infty < 0 \\ f(-1) &= -\left(\frac{1-\theta}{1-\pi}-1\right)\left(\frac{\theta}{\pi}-1\right) > 0 \\ f(1) &= \left(\frac{1-\theta}{1-\pi}-1\right)\left(\frac{\theta}{\pi}-1\right) - 3 < 0 \\ \lim_{m\to+\infty} f(m) &= +\infty > 0\end{aligned}$$

Since $f(m)$ is a continuous function of m, there is one solution for $f(m) = 0$ within $(-\infty,-1)$, and one solution within $(-1,0)$. Since there are at most three solutions for the function $f(m) = 0$, we can conclude that there is only one positive solution $\hat{m} \in (1,\infty)$, that is $\hat{\alpha} \in (0,\frac{1}{2})$. Q.E.D.

Proof Proposition 2. The two statements of the Proposition are proven separately:

- When $\alpha = 1$, society's beliefs match actual firm behavior: $f^v(1,\pi) = 1$ and $f^c(1,\pi) = 0$. This implies that $U^E(v;1) = U(v;1)$ and $U^E(c;1,\pi) = U(c;1)$. Since $U(v;1) > U(c;1)$, by the assumption $a > 0$, we have that $U^E(v;1,\pi) > U^E(c;1,\pi)$, which defines $x(1) = v$ for all firms as NE. The equilibrium is Stable since there always exists sufficiently small ϵ for which $U^E(v;1-\epsilon,\pi) < U^E(c;1-\epsilon,\pi)$

  When $\alpha = 0$, society's beliefs also match actual firms' behavior: $f^v(0,\pi) = 0$ and $f^c(0,\pi) = 1$. This implies that $U^E(c;0,\pi) = U(c;0) = -c$ and $U^E(v;0,\pi) = 0$. Since $U^E(v;0,\pi) > U^E(c;0)$, $x(0) = c$ for all firms is not an equilibrium.

- According to Definition 1, an interior equilibrium demands that $U^E(c;\alpha,\pi) = U^E(v;\alpha,\pi)$ or $F(\alpha,\pi) = a$. When $\alpha = 0$, then $F(\alpha,\pi) - a = -a < 0$. Thus, if there is α such that $F(\alpha,\pi) - a > 0$ there exists at least one α for which $F(\alpha,\pi) - a = 0$ (by the Bolzano's Theorem). Lemma 1 states that $F_\alpha > 0$ for $\alpha \in (0,\hat{\alpha})$, thus if $F(\hat{\alpha},\pi) - a > 0$ there exists one and only one $\alpha^k \in (0,\hat{\alpha})$ such that $F(\alpha^k,\pi) - a = 0$, which is the condition for a NE. Similarly, we know that, when $\alpha = 1$, $F(\alpha,\pi) - a = -a < 0$. From Lemma 1, $F_\alpha < 0$ for $\alpha \in (\hat{\alpha},1)$. Thus, when $F(\hat{\alpha},\pi) > a$ there exists one and only one NE, $\alpha^l \in (\hat{\alpha},1)$.

  $F(\alpha,\pi) \in (0,1)$ for $\alpha \in (0,1)$ since $0 > f^v > f^c > 1$ and $R(\alpha) = 1-\alpha \in (0,1)$. Since $a \in (0,1)$ there always exists a small enough a such that $F(\hat{\alpha}(\pi),\pi) > a$. $\frac{\partial F}{\partial \pi} > 0$ for $\alpha \in (0,1)$ (Appendix A) implies that $\frac{dF}{d\pi} > 0$ by the Envelope

Theorem.

To prove stability note that $F_\alpha(\alpha^k) > 0$ implies that for small enough ϵ, $F(\alpha^k+\epsilon)-a>0$ and $F(\alpha^k-\epsilon)-a<0$. That is $U^E(c,\alpha^k+\epsilon)>U^E(v,\alpha^k+\epsilon)$ and $U^E(v,\alpha^k-\epsilon)<U^E(c,\alpha^k-\epsilon)$. If a small mass of compliant firms deviate, the new violation level is $\alpha^k+\epsilon$. As shown above, their new best response is the same as the original equilibrium strategy, that is, compliance. If a small mass ϵ of violators deviate, the new violation level is $\alpha^k-\epsilon$. From the expressions above, it is clear that the deviants' new best response does not differ from their equilibrium response, that is, violation. Hence, α^k is a Stable NE. $F_\alpha(\alpha^l)<0$ implies that, for small ϵ, $U^E(v,\alpha^l+\epsilon)>U^E(c,\alpha^l+\epsilon)$ and $U^E(v,\alpha^l-\epsilon)U^E<(c,\alpha^l-\epsilon)$. Using the same line of reasoning, it is clear that small masses of compliant firms or violators have incentives to deviate at α^l so that it does not qualify as a Stable NE.

Total differentiation of the condition for interior equilibrium, $F(\alpha,\pi)-a=0$, with respect to π and a gives $\alpha_\pi=-\frac{F_\pi}{F_\alpha}$ and $\alpha_a=\frac{1}{F_\alpha}$. Since $F_\pi>0$, $F_\alpha>0$ at α^k and $F_\alpha<0$ at α^l, $\alpha^k_\pi<0$, $\alpha^k_a>0$, $\alpha^l_\pi>0$ and, $\alpha^l_a<0$. Q.E.D.

Proof Proposition 3. From equation 8, we know that for $\alpha \in (0,1)$,

$\lim_{(\theta,\pi)\to(0,1)} F(\alpha,\theta,\pi) = \lim_{\theta\to 0}(1-\theta)(1-\alpha)(\frac{\alpha}{\alpha+\theta(1-\alpha)} - 0) = 1-\alpha$.

This implies that $\hat{\alpha}(\theta,\pi) = \text{argmax}\, F(\alpha,\theta,\pi) \to 0$.

Thus, as signals become extremely informative, the condition for emergence of interior equilibria k and l, $F(\hat{\alpha}(\theta,\pi),\theta,\pi) > a$ (Proposition 2), is met: Note that $1-\hat{\alpha}(\theta,\pi) \to 1$ while $a \in (0,1)$. Further, since $\alpha^k \in (0,\hat{\alpha}(\theta,\pi))$, it must also be case that $\alpha^k \to 0$. According to Definition 2, at interior equilibrium l, $F(\alpha^l,\theta,\pi)-a = 0$. From the discussion above, it follows that $\lim_{(\theta,\pi)\to(0,1)} F(\alpha^l,\theta,\pi)-a = 1-\alpha^l-a = 0$. Hence, $\alpha^l \to 1-a$ for this equality to hold. Q.E.D.

References

[1] Afsah S., Blackman A., García J.H., T Sterner (2013). *Environmental Regulation and Public Disclosure: The Case of PROPER in Indonesia*, Routledge - RFF Press, Washington D.C.

[2] Akerlof G. (1970) The Market for 'Lemons': Quality Uncertainty and the Market Mechanism, *Quarterly Journal of Economics* 84(3), 488-500.

[3] Akerlof G. (1980) A Theory of Social Custom, of which Unemployment May Be One Consequence, *Quarterly Journal of Economics* 94(4), 749-75.

[4] Amacher, G. , E. Koskela, M. Ollikainen, (2004). Environmental Quality Competition and Eco-Labeling, *Journal of Environmental Economics and Management* 47(2), 284-306.

[5] Arrow K., (1971). Political and Economic Evaluation of Social Effects and Externalities, in M. Intriligator (Ed.), *Frontiers of Quantitative Economics*, North-Holland, Amsterdam, pp. 3-25.

[6] Cremer H., J-F. Thisse, (1999). On the Taxation of Polluting Products in a Differentiated Industry, *European Economic Review* 43(3), 575-94.

[7] Bardhan P., (1997). Corruption and Development: A Review of Issues, *Journal of Economic Literature* 35(3), 1320-46.

[8] Bernheim B.D., (1994). A Theory of Conformity, *Journal of Political Economy* 102(5), 841-77.

[9] Baron D.P., (2009). A Positive Theory of Moral Management, Social Pressure, and Corporate Social Performance, *Journal of Economics & Management Strategy*, 18(1), 7–43

[10] Clark A.E.,(2003). Unemployment as a Social Norm: Psychological Evidence from Panel Data, *Journal of Labor Economics* 21(2), 323-52.

[11] Cropper M., W. Oates, (1992).Environmental Economics: A Survey, *Journal of Economic Literature* 30(2), 675-740.

[12] Elhauge E., (2005). Corporate Managers Operational Discretion to Sacrifice Corporate Profits in the Public Interest, in B. Hay, R. Stavins, R. Vietor (Eds.), *Environmental Protection and the Social Responsibility of Firms*, RFF Press, Washington D.C.,pp. 13-76.

[13] El-Gamal M.A., D.M. Grethner, (1995) Are People Bayesian? Uncovering Behavioral Strategies' Journal of the American Statistical Association 90, 1137-1144

[14] Elster J., (1989). Social Norms and Economic Theory, *Journal of Economic Perspectives* 3(4), 47-74.

[15] Foulon, J., Lanoie P., B. Laplante, (2002), Incentives for Pollution Control: Regulation or Information?, Journal of Environmental Economics and Management 44, 169-187

[16] Fudenberg D., J. Tirole, (1998). *Game Theory*, The MIT Press, Cambridge MA.

[17] Fershtman C., U. Gneezy, J.A. List, (2008).Equity Aversion, Centre for Economic Policy Research, Discussion PaperNo.6853.

[18] Herrendorf R., Valentinyi A., R. Waldman, (2000) Ruling out Multiplicity and Indeterminacy: The Role of Heterogeneity, *The Review of Economic Studies* 67(2), 295-307

[19] Garvie D., A. Keeler,(1994). Incomplete Enforcement with Endogenous Regulatory Choice, *Journal of Public Economics* 55(9), 141-62.

[20] Harrington W.,(1988). Enforcement Leverage when Penalties Are Restricted, *Journal of Public Economics* 37(1), 29-53.

[21] Khan M.A., Y.N. Sun, (2002). Non-cooperative Games with Many Players, in R.J. Aumann, S. Hart (Eds.), *Handbook of Game Theory with Economic Applications* Volume III, North-Holland, Amsterdam, pp. 1761-1808.

[22] Kaplow L., S.M. Shavell, (2007). Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System, *Journal of Political Economy* 115(3), 494-514.

[23] Kübler, D.F. (2001). On the Regulation of Social Norms, *Journal of Law, Economics, and Organization* 17(2), 449-476.

[24] Levitt S.D., J.A. List,(2007). What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?, *Journal of Economic Perspectives* 21(2), 153-74.

[25] Lindbeck A., S. Nyberg, J. Weibull, (1999). Social Norms and Economic Incentives in the Welfare State, *Quarterly Journal of Economics* 144(1), 1-35.

[26] Lai C., C. Yang, J. Chang, (2003). Environmental Regulations and Social Norms, *International Tax and Public Finance* 10(1), 63-75.

[27] Lyon T.P., J.W. Maxwell, (2011). Greenwash: Corporate Environmental Disclosure under Threat of Audit, *Journal of Economics & Management Strategy* 20(1), 3-41.

[28] Nyborg K., K. Telle, (2004). The Role of Warnings in Regulation: Keeping Control with Less Punishment, *Journal of Public Economics* 88(12), 2801-16.

[29] Ostrom, E.(1990) *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge Univ. Press, New York.

[30] Patacchini E., Y. Zenou, (2012). Juvenile Delinquency and Conformism *Journal of Law, Economics, and Organization* 28(1), 1-31

[31] Sethi R., E. Somanathan, (1996). The Evolution of Social Norms in Common Property Resource Use, *American Economic Review* 86(4), 766-88.

[32] Schmeidler, D. (1973). Equilibrium Points of Nonatomic Games, *Journal of Statistical Physics* 7(4), 295-300.

[33] Smith, A. (1790). *The Theory of Moral Sentiments*, 6th edition, Oxford University Press, Oxford.

[34] Young, H.P. (2008). Social Norms, in S.N. Durlauf, L.E. Blume (Eds.), *New Palgrave Dictionary of Economics*, Palgrave Macmillan, London.

Figure 1: Perfect information equilibria

(● Stable ● Non-Stable)

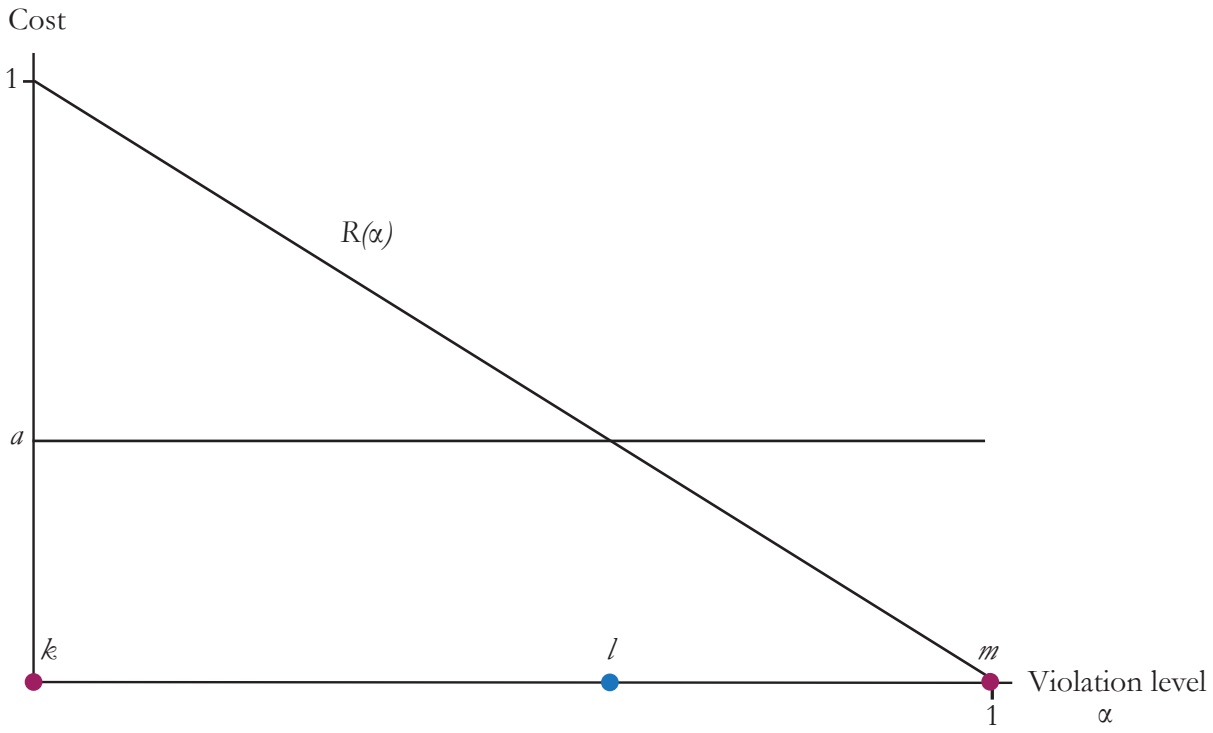
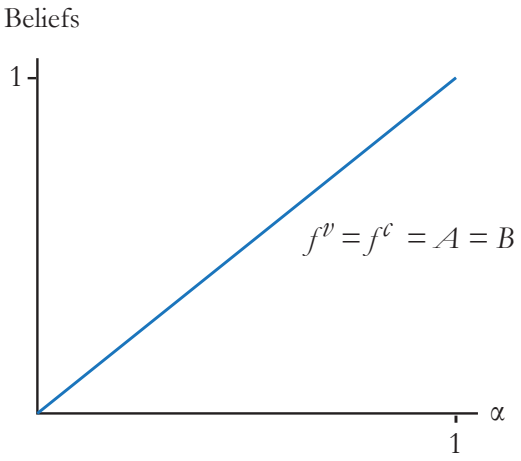
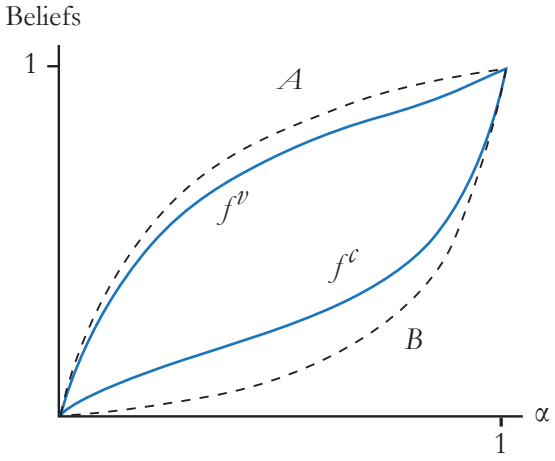


Figure 2: Beliefs under imperfect and perfect information

a) Uniformative signals



b) Informative signals



c) Perfect information

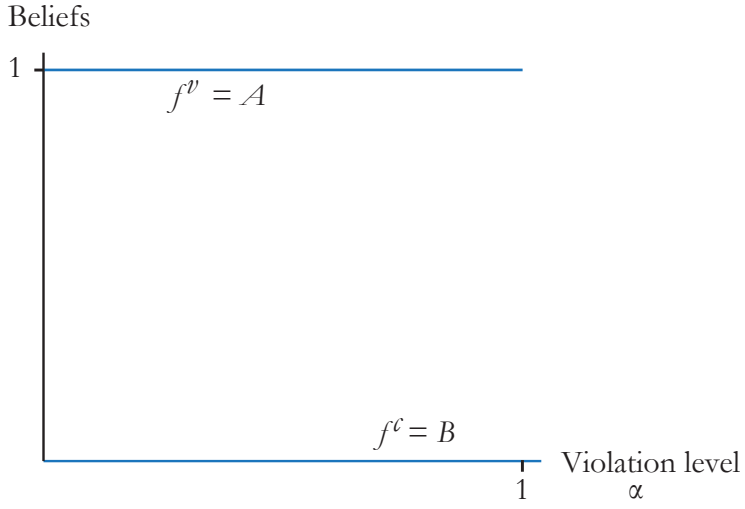


Figure 3a: Imperfect information equilibria with uniformative signals, $\pi = \theta$

(● Stable ● Non-Stable)

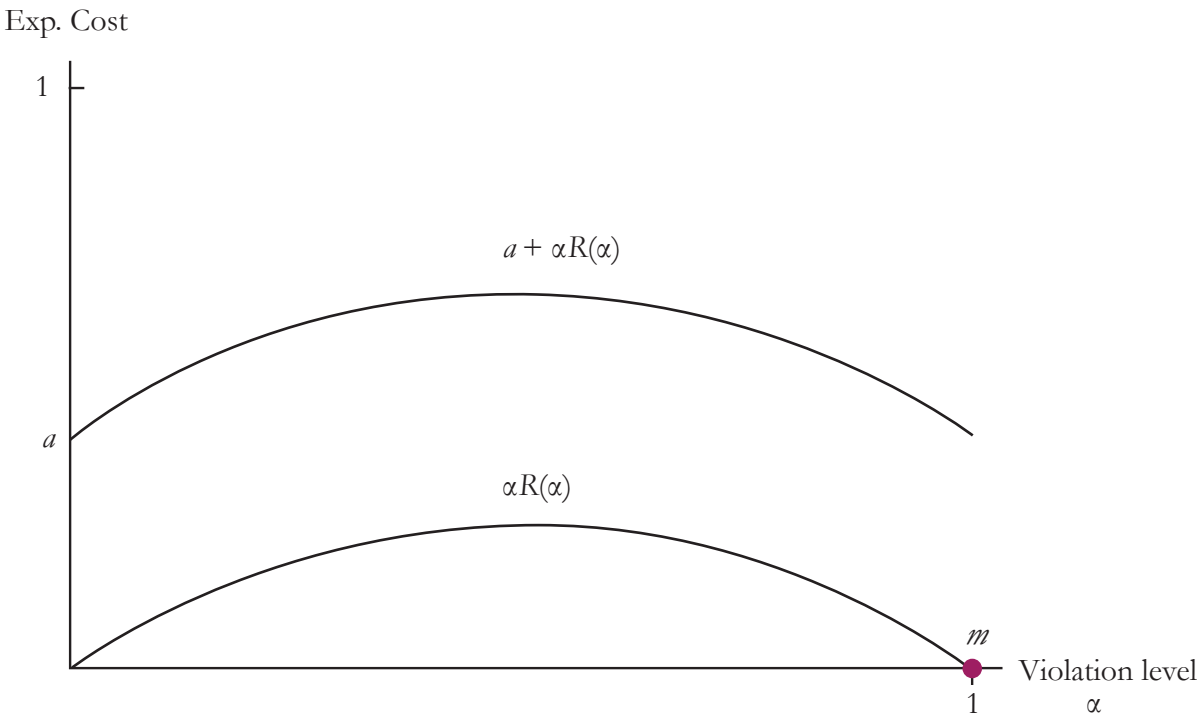


Figure 3b: Imperfect information equilibria with informative signals, $\pi > \theta$

(● Stable ● Non-Stable)

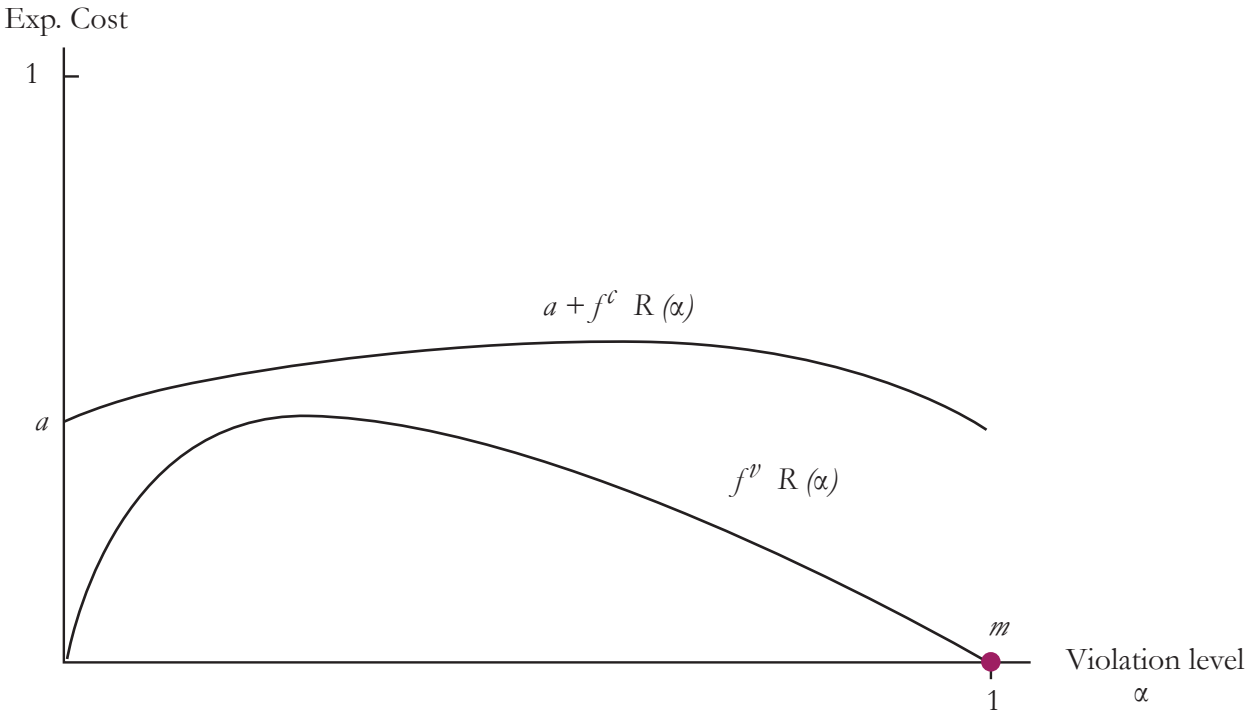


Figure 3c: Imperfect information equilibria with very informative signals, $\pi >> \theta$

(● Stable ● Non-Stable)