Social Norms and Legal Design: Fault-Based vs Strict Liability Offences

Bruno Deffains*    Claude Fluet†‡

This version, March 2015

Abstract

The framing of offences is an important issue in legal or regulatory design. Should offences be fault-based, for instance involving considerations of recklessness or negligence, or should the doing of an act constitute an offence per se? We compare the performance, from a deterrence and enforcement cost perspective, of fault-based versus strict liability offences in the economic model of public enforcement of law, extended to incorporate informal motivations and social norms of conduct. We show that fault-based offences are generally more effective in harnessing social or self-image concerns for the purpose of inducing compliance. When enforcement relies on fines and assessing fault is not too costly, an optimal legal regime and enforcement policy entail fault-based offences when convictions would seldom occur and strict liability offences otherwise. We also discuss the optimal policies when legal sanctions are nonmonetary and when stigmatization imposes a deadweight loss.

KEYWORDS: Other-regarding behavior, social preferences, regulatory offences, law enforcement, strict liability, fault, legal standard, compliance, deterrence. (JEL: D8, K4, Z13)
1 Introduction

Illegal behavior ranges from crimes of great antiquity, such as murder or theft, carrying strong moral opprobrium down to lesser ‘quasi-crimes’, e.g., false or misleading advertising, income underreporting in tax filings, fishing out of season, and the like. An important issue in legal and regulatory design is the categorization of offences. Should they be criminalized or qualified as mere violations punished at most by a fine? Legal systems have been dealing with such questions since the mid 19th century owing to the multiplication of modern regulatory offences, e.g., in factory legislation, food and drug laws or sanitary and public health regulations. More recently, from the 1960s onwards, there has been a resurgence of the debate in the wake of the criminal law reforms in many countries. Whether some offences should be criminalized has also been contentious in the development of new fields of law, in particular competition law, financial regulations and environmental protection legislation.

The issue is in some respects related to the classical dichotomy between malum in se and malum prohibitum. Malum in se means wrong or reprehensible in itself. The expression refers to conduct viewed as inherently wrong independently of regulations or laws. Malum prohibitum refers to conduct that is wrong only because it is prohibited by law. Some acts are crimes not because they are inherently bad, but because they have been declared illegal by statute law.\(^1\) The distinction is important in most penal systems, if only implicitly. Obviously, many acts that are malum in se are also formally prohibited by legislation which then “expresses” the underlying social or moral norm. Independently of formal legal prohibitions, violators of the norm would face social sanctions through stigmatization, moral opprobrium and possibly other more direct means. The outcome would be different for

---

\(^1\)The difference is often described in terms of iussum quia iustum and iustum quia iussum, namely something that is commanded (iussum) because it is just (iustum) and something that is just (iustum) because it is commanded (iussum).
actions that are merely *malum prohibitum*, for example illegal parking.

A related question, although the distinction does not perfectly overlap, is whether offences should be fault-based — that is, involving considerations of knowledge, intention, recklessness or negligence — or defined on a strict liability basis, whereby the mere doing of an act constitutes a punishable offence *per se*. All legal systems qualify offences in terms of physical elements (e.g., conduct) and fault elements. “True crimes”, which include traditional crimes such as murder or theft, are always fault-based. Individuals are found guilty only when they have the requisite *mens rea* (“guilty mind”) and are therefore morally blameworthy. On the other hand, many lesser crimes and most regulatory offences do not require a *mens rea*. Selling alcohol to minors is a strict liability offence because one can be convicted even if one believed the customers were old enough. Similarly, traffic offences are usually on a strict liability basis.

In their survey of the theory of public law enforcement, Polinsky and Shavell (2007) discuss the various policy choices facing the state, one of which concerns the sanctioning rule: “The rule could be *strict* in the sense that a party is sanctioned whenever he has been found to have caused harm (or expected harm). Alternatively, the rule could be *fault-based*, meaning that a party who has been found to have caused harm is sanctioned only if he failed to obey some standard of behavior.” Whether there should be strict liability crimes is nevertheless a highly contentious issue. To give but two examples, the Model Penal Code of the American Law Institute in the 1990s rejected the principle of strict liability in criminal law. By contrast, in the 1990s Australia reformed its criminal code squarely on the basis of the fault-based versus strict liability dichotomy.\(^2\)

\(^2\)For a glimpse of the debates in other countries, see Law Reform Commission of Canada (1974), Faure and Heine (2005), Horder (2005), Simester (2005), Spencer and Pedain (2005), Wils (2007), and Law Commission (2010). For earlier influential discussions, see Kadish (1963) and Fitzgerald (1965). For recent assessments of the evolution in the US, see Singer (1989) and Brown (2012).
This paper analyzes legal design from the perspective of harnessing normative motivations. For this purpose, we extend the economic model of public enforcement of law to incorporate social preferences and pre-existing social norms of conduct. At one extreme, the pre-existing social norms concerning a particular act may be virtually nonexistent, so that the policy prescription would be the same as in the standard model without social preferences. At the other extreme, there is a strong social norm and individuals who are thought not to care would meet strong disapproval. We inquire how the strength of pre-existing social norms of conduct — together with social or self-image concerns with respect to deviations from the norms — affect the relative performance of fault-based versus strict liability offences from a deterrence and enforcement cost point of view.

We consider a general situation where violations of the law are not always detected because enforcing the law is costly. Some individuals nevertheless behave efficiently from a social point of view. Some may do so out of intrinsic moral or prosocial concerns. Others may have no such concerns but would like people to believe that they do or perhaps would want to perceive themselves as having such concerns; that is, they care about social approval or self image. To the extent that normative motivations suffice, legal enforcement is of course superfluous. We focus on situations where an individual’s actions are only vaguely observable by one’s reference group or would be only self-servingly recalled by the individual himself. However, legal sanctions provide hard information from which inferences can be drawn about the individuals’ intrinsic predispositions. Under either fault-based or strict liability offences, social and self image concerns then provide individuals with some incentives to mimic the virtuous.

A basic result is that fault-based offences tend to be more effective in harnessing image concerns. The reason is that convictions are then more informative. A strict liability offence merely ascertains that the offender committed a harmful action and says nothing about the circumstances in
which the action was committed. A fault-based offence unambiguously reveals reprehensible behavior, thereby providing more precise information about the individual’s character. When the social norm is a strong one, with potentially strong stigmatization of violators, socially useful incentives are therefore provided by the signaling role of fault, allowing greater deterrence or lower enforcement costs when ascertaining fault is not too costly. When the social norm is weak, however, it is not always the case that fault-based offences do better than strict liability in harnessing reputational concerns. Which regime performs better depends in a complex way on the underlying situation. The optimal legal regime and enforcement policy are interdependent and entail fault-based offences when convictions would seldom occur and strict liability offences otherwise.

The dichotomy between fault-based and strict liability offences partly captures the distinction between “criminalized” offences and purely “regulatory” offences. In our analysis, the legal design problem is approached from a standard utilitarian perspective. Opprobrium and the stigmatization effects of legal sanctions are considered for their incentive effects. Fault-based offences do better for acts that are clearly bad from a moral or social point of view. When there is no pre-existing norm, strict liability does as well and may be less costly. The analysis therefore provides an economic interpretation of the usefulness of the distinction between malum in se and malum prohibitum for the purpose of legal design and for deriving the optimal enforcement policy.

Section 2 reviews some of the related literature. Section 3 presents the basic setup. Section 4 compares the incentives under different legal regimes and enforcement policies. Section 5 derives the implications for efficient legal design when enforcement relies on fines. Section 6 extends the analysis to nonmonetary sanctions such as imprisonment and also discusses the possibility that stigmatization entails a deadweight loss. Section 7 concludes. Proofs are in the Appendix A.
2 Literature review

Our analysis belongs to a recent microeconomic literature on social preferences emphasizing that actions may reveal one’s unobservable predispositions and that some predispositions are socially valued (e.g., Bernheim, 1994; Bénabou and Tirole 2006, 2011; Daughety and Reinganum, 2010). Numerous experimental or field studies show that social image concerns are major motivators of prosocial behavior (Masclet et al. 2003, Dana et al. 2006, Ellingsen and Johannesson 2008, Andreoni and Bernheim 2008, Ariely et al. 2010, Funk 2010, Lacetera and Macis 2010, among others).

Relatedly, there is a growing literature on the interaction between formal legal sanctions and informal nonlegal sanctions. Much of this literature analyzes the substitutability between legal and nonlegal sanctions, emphasizing that stigma or loss of standing in a community may deter undesirable behavior just as or more effectively than formal legal sanctions (Macauley 1963; Ellickson 1991; Bernstein 1992). One strand of this literature also focuses on the potential complementarity between informal and formal sanctions, noting that legal penalties may influence the existence and impact of informal sanctions (Kahan, 1998, Posner 2000; Cooter 2000a, 2000b; Teichman 2005; Iacobucci 2014). This issue also relates to the role of stigma and shaming penalties in relation to criminal activity; see Rasmusen (1996), Harel and Clement (2007), and Zasu (2007) among others.

A subfield of the “norms and law” literature studies the relationship between morality and law. Cooter (1998a) discusses how law provides incentives to acquire morality and self-control. Posner (1997), Shavell (2002), and McAdams and Rasmusen (2007) provide a general discussion of legal sanctions versus informal motivations as regulators of conduct. Shavell compares the two in terms of the social costs of enforcement and the effectiveness in controlling behavior. He argues that, if the expected private gain from undesirable action and the expected harm due to the conduct are large, it
is optimal to have law supplement morality and, if morality does not function well, law alone is optimal. Mialon (2014) analyzes the effectiveness of moral norms in an evolutionary context. She shows that legal rules may be necessary when norms are easily swayed by social interaction in the long run. The issue of legal design also bears a relation to the concept of “expressive law”. According to this view even “mild law”, i.e., law backed by small sanctions or weakly enforced, can have desirable effects on behavior; see Cooter (1998b), Tyran and Feld (2006), and Galbiati and Vertova (2008, 2014).

The article most closely related to the present one, although in a civil litigation context, is Deffains and Fluet (2013). They model how tort rules and social pressure interact to provide incentives to take care. In their analysis, the extent to which liability rules are privately enforced and the parameters of the tort rules themselves are taken as given, e.g., if found liable the injurer must pay compensatory damages to the victim. The present paper, by contrast, considers legal design together with optimal public law enforcement. The design problem is whether offences should be strict or fault-based and the determination of the legal standard of fault in the latter case. Enforcement concerns the optimal investment in detecting violations and the setting of sanctions, whether monetary or nonmonetary.

3 Set-Up

We start with a simple version of the economic model of public law enforcement.3 The model analyzes the use of legal rules for preventing socially harmful behavior and of public agents to detect and punish offenders. In the standard model, individuals violate the law when their private net benefit from doing so is positive given the risk of legal sanctions. We use this frame-

---

3 Well known surveys are Polinsky and Shavell (2000, 2007). We differ slightly by introducing a costs of assessing fault.
work to define strict liability and fault-based offences. Next we extend the framework to incorporate social preferences.

**The standard model.** Risk-neutral individuals can obtain a private gain \( g \) from committing an act that causes an external harm of amount \( h \). The private gain — equivalently the opportunity cost of not committing the act — varies between individuals and depends on the circumstances.\(^4\) The probability distribution is \( F(g) \) with density \( f(g) \) on the support \([0, \overline{g}]\), where \( \overline{g} > h \). Social welfare is the sum of the gains individuals obtain from committing the act less the harm they cause to others. Denoting behavior by \( e \in \{0, 1\} \), where \( e = 1 \) means commission of the act, and denoting with \( e(g) \) the behavior in the circumstance \( g \), social welfare is

\[
\int_0^{\overline{g}} e(g)(g - h) f(g) \, dg.
\]

Socially optimal behavior is

\[
e^*(g) = \begin{cases} 
1 & \text{if } g \geq h, \\
0 & \text{otherwise.}
\end{cases}
\]

We assume that the harmful act is sometimes socially warranted. This allows a meaningful distinction to be made between strict liability and fault-based offences.

The harmful act is qualified as a strict liability offence if it is illegal irrespective of circumstances. For the time being, we focus on the case where the sanction for violating the law is a fine \( s \), a socially costless transfer of money. The enforcement cost is \( c(p) \) where \( p \) is the probability of detecting harmful acts and \( c(p) \) is the per capita expenditure with derivatives \( c' > 0, c'' \geq 0 \). An individual does not comply with the law if his private gain exceeds the expected fine, \( g \geq ps \). For a given enforcement policy, welfare

\(^4\)Acts can be interpreted from different perspectives, namely acts of “omission” (not complying with some regulation, e.g., fire detectors) versus “positive” acts (driving through red lights).
An optimal policy maximizes this expression with respect to the value of the fine and the probability of detection. Becker’s (1968) maximum sanction principle applies here. To economize on detection costs, the fine should be set at the highest feasible level, say the individuals’ wealth or some given upper bound on allowable fines which we denote by $s_M$. Given the maximal fine, welfare is maximized with respect to the probability of detection. Assuming an interior solution, the first-order condition is

$$ (h - p_{s_M}) \frac{dF(p_{s_M})}{dp} = c'(p). \tag{2} $$

The left-hand side is the marginal social benefit from deterrence, the right-hand side is the marginal enforcement cost. The first-order condition requires $h > p_{s_M}$, implying that optimal enforcement entails underdeterrence compared with first-best behavior. Some individuals, those for whom $p_{s_M} \leq g < h$, will commit the harmful act even though it is not socially warranted. Optimal enforcement trades-off some inefficiency in behavior against savings in enforcement expenses.

With fault-based offences individuals who cause harm are sanctioned only if they failed to obey some standard of behavior. The legal standard is in terms of the circumstances under which the harmful act is committed. An individual’s private benefit must be above some threshold $\hat{g}$ in order for him to avoid liability; otherwise, he is considered to be at fault. Committing the harmful act is illegal when the circumstances are $g < \hat{g}$, in which case the individual is subject to a fine if he is detected; when the circumstances are $g \geq \hat{g}$, the harmful act does not constitute an offence. Individuals therefore commit the act when $g \geq \min(p_s, \hat{g})$.

Enforcement costs include the cost of detecting harmful acts and the additional cost $k$ of assessing circumstances. For a given probability of
detection, the enforcement cost is now

\[ c(p) + kp \left[ 1 - F(\min(ps, \hat{g})) \right] \]

where the second term is the per capita cost of assessing the circumstances of the harmful acts committed by undeterred individuals. The optimal policy consists in choosing the fine, the probability of detection and the legal standard so as to maximize

\[ \int_{\min(ps, \hat{g})}^{\infty} (g - h) f(g) dg - c(p) - kp \left[ 1 - F(\min(ps, \hat{g})) \right]. \]

The maximum sanction principle still applies and it is easily seen that an optimal policy requires \( \hat{g} \geq ps_M \), otherwise enforcement costs could be reduced with no detrimental effect on deterrence. An interior solution yields the first-order condition

\[ (h + pk - ps_M) \frac{dF(ps_M)}{dp} = c'(p) + k [1 - F(ps_M)]. \] (3)

Optimal enforcement may now entail either underdeterrence or overdeterrence compared with first-best behavior. Overdeterrence would reduce the frequency of harmful acts and therefore the cost of ascertaining circumstances. As with strict liability offences, there is a trade-off between enforcement costs and some distortion of behavior.

Suppose assessing circumstances involves no additional cost, that is, \( k = 0 \). Enforcement costs are then the same under fault-based and strict liability offences. Condition (3) reduces to (2) and welfare is therefore the same under either legal regime. Thus, when assessing circumstances involves no additional cost, strict liability and fault-based offences are equally efficient. Clearly, when \( k > 0 \), the optimal legal regime is strict liability. Thus, when both the legal regime and the enforcement policy are optimally chosen, individuals are underdeterred to some extent.

A second observation is that, under a fault regime, any standard \( \hat{g} \geq ps_M \) yields the same level of deterrence. In other words, the legal standard is
irrelevant. Observe also that a standard above the upper bound of possible gains (i.e., $\bar{g} \geq \bar{g}$) is equivalent to strict liability. Committing the act is then illegal irrespective of possible circumstances and ascertaining circumstances then serves no purpose. By contrast, if sanctions were costly to impose, e.g., fines involve collection costs or sanctions are nonmonetary as with prison sentences, the standard of fault should satisfy $\bar{g} = p_s M$ (see Shavell 1987). Undeterred individuals should then not be found to be at fault even though they behave inefficiently from a social point of view, otherwise unnecessary sanction costs would be incurred. When sanctions are socially costly, the advantage of fault-based liability is to rely on the threat of sanctions while avoiding the cost of actually imposing them. With socially costless monetary sanctions fault-based liability has no useful role so far.

Social preferences. In the standard model behavior depends on private costs and benefits as conventionally defined. We now consider informal motivations. We assume that there are two types of individuals. A proportion $\lambda$, referred to as type $t = 1$, is intrinsically motivated to behave in a socially responsible manner. Such individuals are “good citizens” with moral or prosocial predispositions. The other group, referred to as type $t = 0$, has no such predispositions. Moreover, prosocial predispositions are socially valued and those who are thought to be good citizens earn social esteem or status, a source of utility.

The utility of a type-$t$ individual is

$$u_t = w - \gamma_t \max(e - e^*, 0) + \beta \mu, \quad t = 0, 1. \tag{4}$$

The first term, $w$, is net “material” payoff as in the conventional model. In the middle term, the parameter $\gamma_t$ is the disutility (“guilt”) suffered when one causes external harm while deviating from the socially responsible behavior $e^*$. Misbehavior occurs when $e = 1$ and $e^* = 0$. For the good citizens, $\gamma_1 > 0$ and is sufficiently large to intrinsically motivate the individual\(^5\); for

\(^5\)It suffices that $\gamma \geq h$, i.e., the good citizen “internalizes” the harm he causes.
the bad citizen, $\gamma_0 = 0$ and the middle term vanishes. As defined here, the social (or moral) norm of conduct is what everyone should be doing given the circumstances.\(^6\)

The third term in (4) is the utility from one’s social image or reputation. $\beta$ is a positive parameter and $\mu$ is society’s belief about the individual’s type. The belief will depend on information concerning the individual, i.e., $\mu$ equals the conditional expectation $E(t \mid I)$ where $I$ denotes publicly available information. Given our definition of types, the conditional expectation is simply the posterior probability that the individual is a good citizen. All individuals are assumed to care equally about social approval and $\beta$ may be interpreted as the utility of being perceived as a good citizen, given that the utility of being perceived as a bad citizen is normalized to zero. The parameter captures both the importance individuals attach to social approval and the importance (“social pressure”) society ascribes to being a good citizen in the case at hand. Both $\beta$ and the proportion $\lambda$ of good citizens reflect the salience of the social norm with respect to the kind of situation (and therefore possible acts) considered. For instance, when the harmful act is widely viewed as particularly reprehensible in ordinary circumstances, $\beta$ will be large and presumably so will be $\lambda$.

Welfare is defined in the usual way as the sum of utility over all individuals,

$$W = \int_0^\varpi [(1 - \lambda)u_0(g) + \lambda u_1(g)] f(g) \, dg \quad \text{(5)}$$

where $u_t(g)$ is the utility (or expected utility) of a type-$t$ individual in the circumstance $g$.

Before proceeding, we show that first-best behavior in the present set-up is the same as in the standard model without social preferences. Assume that individuals can both cause harm or suffer harm caused by others. Consider an omniscient regulator who can directly impose the action profile $e(g)$,\(^6\)

---

\(^6\)This is a simple version of Kant’s categorical imperative as in Brekke et al. (2003).
where $w_0$ is initial per capita wealth. Let the action profile $\tilde{c}(g)$ be welfare maximizing and suppose that the regulator has the option of either publicizing or preventing any information about the individuals’ types. If an optimum entails that no information is disclosed, then $\tilde{c}(g)$ maximizes $W$ subject to the resource constraint (6), given that beliefs satisfy $\mu = \lambda$ where $\lambda$ is the prior belief about types. This implies $\tilde{c}(g) = e^*(g)$ as defined in (1). Combining (4) and (5), the first-best welfare then equals

$$W^* = w_0 + \int_{0}^{\gamma} (g - h) f(g) dg + \beta \lambda.$$

Now the action profile $e^*(g)$ would also be optimal when full or imperfect information about types is disclosed. First, because the omniscient regulator is able to independently control the flow of information, there would be no reason for him to distort behavior from the wealth maximizing action profile. Secondly, welfare would also be as in (7) because reputational benefits and losses simply cancel out. The result follows from the law of iterated expectations\footnote{That is, $E(E(t | I)) = E(t) = \lambda.$} and the linearity of reputational utility in the beliefs about one’s type.

**Offences and labeling.** Society at large — equivalently one’s relevant reference group — is assumed not to be able to directly observe the circumstances faced by an individual nor his behavior. The assumption prevents social pressure from bearing directly on individuals independently of the legal system. Otherwise one’s type could be inferred directly from one’s behavior. When $\beta$ is large enough there would then be situations where the legal system plays no useful role. Direct informal reputational sanctions would suffice to induce socially appropriate behavior.
Public enforcers can detect harmful acts and can ascertain the circumstances; that is, they are able to apply the law when offences are fault-based. Legal proceedings against offenders then constitute public information from which society at large draws inferences about the individuals’ type. For simplicity, we assume that the only information publicly available about an individual — by which we mean in society at large — is either $G$ for “guilty”, in which case the individual is known to have been found guilty of an offence, or $N$ for “no news”. The latter means that either the individual did not commit an offence or that he did but was not detected by the public enforcers. The publicly available information affecting one’s social image is therefore the binary signal $I \in \{N, G\}$. Society’s belief about an individual will then be either $\overline{I}_N = E(t \mid N)$ or $\overline{I}_G = E(t \mid G)$. The significance of the signal depends on what the events “no news” and “guilty” reveal about one’s type in the social equilibrium. This will depend on the legal regime, namely whether offences are strict or fault-based, the legal standard in the latter case, and on the enforcement policy. Generally speaking, the event “guilty” will be detrimental to one’s reputation. Other things equal, individuals wish to avoid being labeled as offenders.

We have stressed social signaling, which requires that convictions constitute public information. In practice convictions are often not publicized or receive little publicity. However, as in Bénabou and Tirole (2006), our framework can also be reinterpreted in terms of self-signaling through a concern for self-image. A simple formulation is Bodner and Prelec’s (2003) dual-self model. In the latter, an individual’s total utility is the sum of the “outcome utility” from choosing a particular course of action, which depends on one’s fuzzily known true inner predispositions, and of a “diagnostic utility” whereby an individual draws inferences about his true self from information about his past behavior. In terms of the utility function in (4), the first two terms correspond to outcome utility, the last term is the diagnostic utility based on objective information. In this interpretation,
individuals know they will have imperfect recall (or only self-serving recollections) of their past actions and motivational states and they wish to avoid future low diagnostic utility. For example, one may occasionally exceed the speed limit in a school zone or evade tax, but would feel shame from being labeled an offender even if convictions are not publicized.\footnote{See McAdams and Rasmusen (2007) on the distinction between guilt, social disesteem, and shame.}

4 Equilibrium under a Given Regime

We first describe the equilibria under given legal regimes and enforcement policies. A perfect Bayesian equilibrium is characterized by the individuals’ action profiles and the beliefs about individuals’ type conditional on the “guilty” and “no news” events. The legal regime is defined by the standard of fault when committing the harmful act. The regime is fault-based if the standard is less than the upper bound of possible gains, otherwise the regime is strict liability. The enforcement policy is defined by the fine for unlawful conduct and the probability of detecting such behavior.

We proceed in three steps. First we derive the action profiles taking as given the posterior beliefs conditional on the “guilty” and “no news” events. Next we derive these beliefs as a function of action profiles. Finally we solve for the equilibrium wherein action profiles and beliefs are consistent with one another.

**Incentives.** Denote the sanctioning rule by $\delta(g, \bar{g})$ where $\delta(g, \bar{g}) = 1$ if $g < \bar{g}$ and is otherwise zero. The expected utility of a type-$t$ individual in the circumstance $g$ is

$$
u_t = w + e \left[ g - p\delta(g, \bar{g})s \right] - \gamma_t \max(e - e^*(g), 0)$$

$$+ \beta \left[ pe\delta(g, \bar{g})\tilde{r}_G + (1 - pe\delta(g, \bar{g}))\tilde{r}_N \right], \quad e \in \{0, 1\}, \quad t \in \{0, 1\}.$$
The first two terms comprise material payoff as conventionally defined. The first term, \( w \), is the part of the individual’s wealth that he takes as given. This consists of initial wealth minus the average harm caused by others plus the per capita tax to finance the enforcement policy (expenditures minus fines collected). The second term is the expected net material payoff from committing or not committing the harmful act. The third term is the guilt disutility from committing the harmful act when it is socially unwarranted. The fourth term is the expected reputational utility. If the individual does not commit the act (i.e., \( e = 0 \)) or if he would not be legally at fault when he does (i.e., \( \delta(g, \bar{g}) = 0 \)), the belief about his type will be \( \bar{t}_N \) for sure, the posterior probability that he is a good citizen given “no news”. If he unlawfully commits the act, he is detected with probability \( p \) and the belief about his type is then \( \bar{t}_G \), the posterior probability conditional on “guilty”. If he is not detected, the belief is again \( \bar{t}_N \). These beliefs are determined at equilibrium but are taken as given by the individual.

For a nonprosocial individual, \( \gamma_0 = 0 \). If the harmful act is not committed, expected utility is \( u_t = w + \beta \bar{t}_N \). If it is committed and it is lawful, that is \( g \geq \bar{g} \), expected utility is \( u_t = w + g + \beta \bar{t}_N \). Hence it will then be committed. In circumstances where the act is unlawful, expected utility is

\[
u_t = w + (g - ps) + \beta(p\bar{t}_G + (1 - p)\bar{t}_N)
\]

and the act is then committed if \( g \geq p(s + \beta \Delta) \), where \( \Delta \equiv \bar{t}_N - \bar{t}_G \) is henceforth referred to as the legal stigma, i.e., the reputational loss from being convicted. The term \( ps \) is the standard material incentive to comply with the law, the term \( p\beta \Delta \) is the reputational motive. Altogether a nonprosocial commits the harmful act when

\[
g \geq \min[\bar{g}, p(s + \beta \Delta)] \equiv g_0,
\]

where \( g_0 \) is short-hand for the private gain threshold of nonprosocial individuals.
Good citizens are also motivated by legal sanctions and reputational concerns. In addition, their behavior reflects an intrinsic concern for complying with the social (as opposed to the legal) norm conduct. Given $\gamma_1$ sufficiently large, a good citizen never commits the harmful act in circumstances $g < h$. When $g \geq h$, the harmful act entails no guilt and the good citizen then behaves the same as the nonprosocial. The harmful act is therefore committed if
\[
g \geq \max(h, g_0) \equiv g_1
\]
where $g_1$ is the gain threshold for good citizens. The following summarizes the preceding discussion.

**Lemma 1** $g_0 \leq \widehat{g}$ and $g_1 = \max(h, g_0)$.

We shall say that type-$t$ individuals are underdeterred (resp. overdeterred) when the threshold $g_t$ is less (resp. greater) than the first best $h$. Whether individuals comply with the law is different. As noted, the legal standard may differ from the first best (and social norm). The interpretation of Lemma 1 is therefore that, if there is some overdeterrence (which requires $\widehat{g} > h$), then all individuals are equally overdeterred. Otherwise they either all efficiently behave or the good citizens do while bad citizens are underdeterred. Moreover, bad citizens never overcomply with the law but good citizens might.

**Beliefs.** The conditions (8) and (9) define the best response functions of an individual of either type given the behavior of others. How others behave affects the payoffs from one’s actions through its effect on the social significance of the “guilty-no news” events, as captured by the beliefs $\overline{t}_G$ and $\overline{t}_N$. The posterior beliefs and therefore the legal stigma are obtained from Bayes’ rule. We restrict attention to private gain thresholds satisfying Lemma 1.
Lemma 2 If \( \tilde{g} \leq h \), \( \Delta \geq \lambda \) and is decreasing in \( g_0 \) down to \( \Delta = \lambda \) when \( g_0 = \tilde{g} \). If \( \tilde{g} > h \) and \( g_0 < g_1 = h \), \( \Delta > 0 \) and is decreasing in \( g_0 \). If \( \tilde{g} > h \) and \( g_0 = g_1 \geq h \), \( \Delta = 0 \).

Unless both types behave the same, bad citizens are more likely to commit the harmful act. Therefore they are more likely to be convicted, implying that the event “guilty” is bad news concerning the individual’s type compared to “no news”. When the legal standard satisfies \( \tilde{g} \leq h \), good citizens are never found guilty. A conviction then reveals perfectly that the individual is nonprosocial, so that \( \overline{\tau}_G = 0 \) and \( \Delta = \tau_N \). The more the nonprosocial behave like good citizens, the smaller \( \tau_N \). When everyone behaves the same, the event “no news” is uninformative because it occurs with certainty, so the posterior probability then equals the prior \( \lambda \) that an individual is a good citizen.\(^9\) When \( \tilde{g} > h \), as would be the case with a strict liability offence, both good and bad citizens will at times be convicted, hence \( \overline{\tau}_G > 0 \). As long as violating the law is more likely for bad citizens, \( \tau_N > \tau_G \) and the legal stigma is positive.\(^10\) When both types behave the same, the events “guilty” and “no news” are uninformative and posterior beliefs equal the prior in either case. The legal stigma then vanishes.

Figure 1 provides examples of the legal stigma as a function of the bad citizens’ threshold under two different legal regimes, given first-best behavior on the part of the good citizens. The probability of detection is the same in both regimes. In case \( A \), the legal standard corresponds to the social norm, \( \tilde{g}_A = h \). The legal stigma is then bounded below by \( \lambda \). In case

---

\(^9\) \( \beta \lambda \) is the disutility from being perceived as a bad rather than an average citizen. The event “guilty” is then an out-of-equilibrium event with zero probability, implying that \( \overline{\tau}_G \) cannot be computed using Bayes’ rule. The reputational penalty is obtained from \( \lim_{g_0 \uparrow 1} \overline{\Delta} = \lim_{g_0 \uparrow 1} \tau_N = \lambda \). This can also be rationalized in terms of Cho and Kreps’ (1987) D\(_1\) criterion.

\(^10\) Strict liability disregards circumstances. Good citizens will then sometimes efficiently choose not to comply with the law given their knowledge of circumstances. See Shavell (2012).
$B$, the legal standard is above the social norm, $\tilde{g}_B > h$. The legal stigma then vanishes when all the nonprosocial conform to the social norm. When the non prosocial are only slightly underdeterred, the legal stigma under the regime $A$ is therefore larger than under the regime $B$. As depicted, the curves intersect. This need not occur but it is a possibility when the nonprosocial are sufficiently underdeterred. We will discuss this further when we turn to legal design.

Equilibrium. An equilibrium consists of private gain thresholds and of a legal stigma that are mutually consistent.

Proposition 1 Let the enforcement policy and legal regime satisfy $\hat{g} \geq ps$. Then there is a unique equilibrium with $g_0 \leq g_1$.

(i) If $ps \geq h$, $g_0 = g_1 = ps$.

(ii) If $ps < \hat{g} \leq h$, $g_1 = h$, $g_0 \in (ps, \hat{g}]$ and is increasing in $p$ and $s$ so long as $p(s + \beta \lambda) < \hat{g}$, otherwise $g_0 = \hat{g}$; in either case, $g_0$ is increasing in $\hat{g}$. 
(iii) If $ps < h < \hat{g}$, $g_1 = h$, $g_0 \in (ps, h)$ and is increasing in $p$ and $s$, it may be increasing or decreasing in $\hat{g}$.

The legal standard — that is, legal design — matters only when $ps < h$. The Figures 2 to 4 illustrate different equilibria for this case. Good citizens then conform to the social norm. In the figures, $\Delta(g_0)$ is the legal stigma as a function of the bad citizens’ threshold under a given legal regime and enforcement policy; $g_0(\Delta)$ is their threshold as a function of the legal stigma under the same regime and enforcement policy. The perfect Bayesian equilibrium is the intersection of the two curves (at point $E$).

When $\hat{g} < h$, an offender is for sure nonprosocial. Figure 2 compares the equilibria under two different legal standards. With the standard $\hat{g}_A$ the stigma and deterrence curves are respectively $\Delta_A(g_0)$ and $g_0^A(\Delta)$. The equilibrium is at $E_A$, a corner equilibrium where everyone complies with the law. With the standard $\hat{g}_B > \hat{g}_A$, we have an interior equilibrium at $E_B$. In either case, deterrence increases with a strengthening of the legal standard.
because this shifts the stigma curve to the right (so long as $\hat{g} < h$). The
intuition is that strengthening the standard increases the significance of the
"no news" event, so that reputational effects have more bite.

In Figure 3, the standard is the first-best $\hat{g} = h$. In the case represented
all individuals comply with the law and therefore are optimally deterred.
This requires $p(s + \beta \lambda) \geq h$. In the figure, the latter condition holds as a
strict inequality, hence the detection probability could be reduced while still
preserving first-best deterrence. Thus, under a fault-based regime, first-best
deterrence is feasible even though $ps < h$. Indeed, when $\beta \lambda$ is sufficiently
large, specifically $\beta \lambda \geq h$, first-best deterrence is feasible with purely sym-

bolic convictions with no fines, an example of “mild law”.

In Figure 4, $\hat{g} > h$ and both types will then sometimes not comply
with the law. First-best deterrence then cannot be achieved with $ps < h$.
Strengthening the legal standard further may have an ambiguous effect on
the legal stigma curve. As before, it increases the significance of “no news”.

In Figure 3: A corner equilibrium with $\hat{g} = h$
However, it also reduces the significance of the “guilty” outcome because more good citizens are convicted, so the net effect on deterrence may be ambiguous. In the situation represented in Figure 4, $\hat{g}_B > \hat{g}_A$ and weakening the standard from $\hat{g}_B$ to $\hat{g}_A$ shifts the stigma curve upwards, so deterrence increases.

![Fig. 4: Equilibria with $\hat{g} > h$](image)

Finally, increasing the probability of detecting illegal acts increases the significance of “no news”, with no effect on the significance of the “guilty” event. In the figures, the legal stigma curves would shift (or rather rotate) upwards. Because offenders are now more likely to be convicted, a larger probability of detection also shifts the deterrence curve to the right. Thus, greater detection unambiguously increases deterrence, except at corner solutions where all bad citizens are efficiently deterred as in Figure 2. Similarly, a larger fine unambiguously increases deterrence because it shifts the deterrence curves to the right.\(^{12}\)

\(^{11}\)The conditional expectation $\hat{E}_G$ does not depend on $p$. See the proof of Lemma 2.

\(^{12}\)Greater detection has an ambiguous effect on the equilibrium legal stigma. A negative effect may be interpreted as greater legal enforcement partially crowding out informal
5 Optimal Legal Regime and Enforcement

Welfare equals the first best $W^*$ as defined in (7) minus the loss from socially inefficient behavior and minus the per capita enforcement expenditure:

$$W = W^* - \{ (1 - \lambda) \int_{g_0}^h (h - g) f(g) \, dg + \lambda \int_{g_1}^h (h - g) f(g) \, dg \} - C(p, g_0, g_1, \bar{g})$$

(10)

where $g_0$ and $g_1$ are equilibrium thresholds as derived in Section 4. The expression inside the brackets is the loss from inefficient behavior; the formulation allows for the possibility of overdeterrence. The third term is the enforcement cost function. When $\bar{g} = \bar{g}$, the legal regime is strict liability and enforcement expenditures reduce to the cost of detecting harmful acts:

$$C(p, g_0, g_1, \bar{g}) = c(p).$$

(11)

When $\bar{g} < \bar{g}$, offences are fault-based and enforcement expenditures also includes the cost of assessing the circumstances of the harmful acts that are detected. The total cost is then

$$C(p, g_0, g_1, \tilde{g}) = c(p) + pk [1 - (1 - \lambda)F(g_0) - \lambda F(g_1)], \quad \tilde{g} < \bar{g}. \quad (12)$$

**Deterrence maximizing legal design.** To start, we take the enforcement policy as given and compare different legal designs in terms of deterrence.

**Proposition 2** When $ps < h$, the legal regime maximizing deterrence of the nonprosocial is either strict liability or the fault-based regime with the standard $\bar{g} = h$. motivations; a positive effect reflects complementarity between legal enforcement and informal sanctions. By contrast, a larger fine unambiguously reduces the legal stigma.
This contrasts with the indetermination of the legal standard in the model without social preferences. Strict liability and fault-based offences are not equivalent in terms of deterrence because they yield different legal stigmas. Moreover, if deterrence is maximized with a fault regime, the efficient legal standard equals the social norm. When $ps \geq h$, the standard is irrelevant so long as $\hat{g} \geq ps$. Individuals then behave as in the standard model and are either efficiently deterred or equally overdeterred.

The intuition for Proposition 2 is that deterrence of the nonprosocial increases with the legal standard when $\hat{g} \leq h$. If it can be increased further still with $\hat{g} > h$, then it reaches its maximum at the upper bound $\hat{g} = \zeta$. For an enforcement policy satisfying $ps < h$, deterrence is therefore maximized either with the legal standard $\hat{g} = h$ or with the standard $\hat{g} = \zeta$, where the latter amounts to strict liability.

**Legal design and reputational incentives.** We illustrate why one regime may perform better than the other. In Figure 5 the regime $A$ is fault-based with the standard $\hat{g} = h$, the regime $B$ is strict liability. The fine and the probability of detection are the same, hence enforcement need not be optimal.

The stigma curves intersect at $D$. We compare two situations, $L$ and $H$, which differ in the intensity of reputational concerns with $\beta_L < \beta_H$. In situation $L$ the deterrence curve is not very sensitive to beliefs about one’s type and deterrence under either legal regime is relatively low. As shown, it is greater under strict liability. Situation $H$ yields the opposite. The social norm has high salience and individuals are very sensitive to reputational penalties. Deterrence is then relatively high and is greater with the fault regime.

Suppose the enforcement policy in Figure 5 is in fact optimal for the strict liability regime in situation $L$, with equilibrium at $E^L_B$. The same enforcement policy would not be optimal in situation $H$ even if assessing fault is costless. Because the marginal loss from underdeterrence is smaller.
at \( E^H_A \) than at \( E^L_S \), it would be welfare improving to somewhat reduce enforcement. Thus, in situation \( H \) the optimal regime would be fault-based and enforcement expenditures would be smaller than in situation \( L \).

![Fig. 5: Legal design when \( \beta_L < \beta_H \)](image)

The foregoing presumes that the stigma curves intersect. As remarked in Section 3, this need not occur. Obviously, convictions are more revealing about intrinsic predispositions in a fault-based than in a strict liability regime. However, what matters for reputational incentives is the difference in beliefs between the “guilty” and “no news” events. It can be shown that the binary signals under strict liability and fault cannot be ranked in terms of informational content. While the event “guilty” constitutes more unfavorable news about an individual’s type under the fault regime, the event “no news” is not more favorable under fault than under strict liability. This explains why the stigma curves may intersect.

**Lemma 3.** The stigma curves under strict liability and the fault regime with the legal standard \( \tilde{g} = h \) intersect at most once and do so if and only if

\[
p > \frac{1}{1 + (1 - 2\lambda)F(h)}
\]  

(13)
The condition (13) implies \( p > 1/2 \). Moreover it cannot hold when \( \lambda \geq 1/2 \) because the right-hand side would be greater than unity. In other words, the situation depicted in Figure 6 cannot arise when good citizens constitute a majority. For any enforcement policy with \( ps < h \), a fault regime then always induces greater deterrence than strict liability.

**Optimal policies.** Generally speaking, the optimal legal regime and enforcement policy must be jointly chosen and will depend on the features of the underlying situation, e.g., the proportion of good citizens, the salience of the social norm, the likelihood of the circumstances under which harmful acts would be socially warranted, and the cost of detecting offenders and assessing circumstances.

**Proposition 3** Under an optimal legal regime and enforcement policy, the fine is maximal and the probability of detection satisfies \( ps_M < h \).

(i) If liability is fault-based, the legal standard is \( \hat{g} = h \), the probability of detection satisfies \( p(s_M + \beta\lambda) \leq h \) and the nonprosocial are underdeterred or efficiently deterred; convictions constitute a rare event,

\[
p(1 - \lambda)(F(h) - F(g_0)) < \frac{1}{2}. \tag{14}
\]

(ii) If liability is strict, the nonprosocial are underdeterred.

When ascertaining circumstances is not too costly,

(iii) liability is fault-based if \( \lambda \geq 1/2 \) or if \( s_M \) or \( \beta \) are sufficiently large; if liability is strict, convictions (and therefore offences) constitute a frequent event,

\[
p[1 - \lambda F(h) - (1 - \lambda)F(g_0)] \geq \frac{1}{2}. \tag{15}
\]

The left-hand side of (14) is the frequency of convictions at equilibrium under the fault regime. The left-hand side of (15) is the frequency of convictions under strict liability given the equilibrium deterrence under that
The maximum sanction principle still holds for the usual reason: a larger fine allows the same level of deterrence to be achieved with a smaller probability of detection, thus saving on enforcement costs. By contrast with the standard model, a fault regime may now be optimal even though assessing fault is costly. Moreover, first-best deterrence is then a possibility.

Because the legal regime is optimally chosen, overdeterrence is never optimal. It would require \( ps_M > h \), in which case reputational incentives vanish and strict liability does as well as the fault regime and strictly better if \( k > 0 \). But then this is dominated by strict liability with \( ps_M = h \). The latter in turn is dominated by strict liability with some degree of underdeterrence and possibly also, if \( k \) is not too large, by the fault regime with either first-best deterrence or some degree of underdeterrence.

\(^{13}\)Everyone commits the harmful act when \( g \geq h \); for \( g \) in \([g_0, h)\) the non prosocial also do so.
In Figure 6, the corner equilibrium $E_A$ is the optimal policy under the fault regime given the maximal fine $s_{M}^{A}$. The probability of detection is then $p_A$ such that $p_A(s_{M}^{A} + \beta \lambda) = h$. Optimal enforcement then satisfies the condition
\[ p_A k f(h) \left( \frac{dg_0}{dp} \right)_{p=p_A} \geq c'(p_A) + k \left[ 1 - F(h) \right]. \] (16)
The right-hand side is the increase in detection and fault assessment costs from a marginal increase in the probability of detection up to $p_A$. The left-hand side is the resulting savings in fault assessment costs from the marginal increase in deterrence. The derivative is a left derivative. The right derivative is zero because increasing detection slightly beyond $p_A$ has no effect on deterrence, as is obvious from Figure 6.

The equilibrium at $E_B$ illustrates an interior optimum for the case of a smaller maximal fine $s_{M}^{B}$. The optimal probability of detection is then $p_B$ such that $p_B(s_{M}^{B} + \beta \lambda) < h$, hence the nonprosocial are underdeterred. The marginal social benefit from greater detection is smaller than in case $A$ because of the smaller fine. The first-order condition is now
\[ (h + p_B k - g_0) f(g_0) \left( \frac{dg_0}{dp} \right)_{p=p_B} = c'(p_B) + k \left[ 1 - (1 - \lambda)F(g_0) - \lambda F(h) \right]. \] (17)
When $k = 0$, the optimum must be an interior solution as in (17).

Whether the optimum is interior or at a corner, fault-based offences may do better than strict liability even though assessing fault is costly because of the larger legal stigma attached to convictions. The condition (14) is then necessary. If it did not hold deterrence could be increased by switching to strict liability under the same enforcement policy, as this would then yield a larger stigma (see the proof of Proposition 3). Because this would also save on the fault assessment costs, strict liability would be unambiguously better.

Part (iii) of the proposition provides sufficient conditions for a fault-regime to be optimal when assessing fault is not too costly. The condition
that good citizens are sufficiently numerous follows directly from Lemma 3. Even when good citizens are not a majority, fault-based offences do better when either the maximum permissible fine or image concerns are sufficiently large. In either case, appropriate deterrence (including first-best deterrence) can be achieved with a relatively small probability of detection, which ensures that the fault regime is deterrence maximizing. By contrast, condition (15), which implies $p > 1/2$, is necessary for strict liability to be optimal when assessing fault is not too costly. If the condition did not hold, switching to fault-based offences would increase deterrence under the same enforcement policy. The condition need not hold when ascertaining circumstances would impose significant costs.

6 Costly Sanctions and Stigmatization

We consider two extensions of the foregoing analysis. First, we inquire how the optimal policies differ when enforcement relies on nonmonetary sanctions rather than fines. Next we relax the assumption that reputational consequences only serve to motivate and examine the possibility that they also entail a social cost.

**Nonmonetary sanctions.** We take imprisonment as an example but the results would carry over to other forms of nonmonetary sanctions such as community service, withdrawal of a licence and the like. Let $s$ now denote the length of prison sentence, where $s_M$ is the maximum allowable. The disutility is assumed to be proportional to the sentence. While a fine is a pure transfer involving no social costs, imprisonment is a net loss in the utilitarian calculus. In addition society may bear a resource cost which we represent by $qs$ where $q$ is the administrative cost per unit of sentence.

Consider first the standard framework without social preferences. Compared with the formulation in Section 3, enforcement costs are now aug-
mented by the addition of
\[ \int_{\min(p,s,\tilde{g})}^{\tilde{g}} p(s + qs) f(g) \, dg, \]
the per capita costs of the sanctions imposed on convicted offenders. Whether
the regime is strict liability or fault-based, the optimal policy is again to set
the sanction at the maximum permissible level.\(^{14}\) However, under strict
liability and by contrast with the case of fines, the optimal probability of
detection may now entail overdeterrence relative to first-best behavior. The
possibility arises because increasing detection may reduce prison costs: while
offenders are more likely to be sanctioned, there will also be a smaller num-
ber of them. Under a fault regime, the optimal legal standard satisfies
\( \tilde{g} = ps_M \) and everyone then complies with the law. Although sanctions are
never actually imposed, because assessing fault is costly, there may also be
overdeterrence compared with the first best in order to economize on fault
assessment costs. If assessing circumstances imposes little costs, a fault
regime is obviously preferable.\(^{15}\)

In the model with social preferences, the properties of the equilibria
derived in Section 4 remain the same. The only change is with respect to
the social welfare function. Welfare is now modified to
\[
W = W^* - (1 - \lambda) \left\{ \int_{g_0}^{h} (h - g) f(g) \, dg + p \int_{\min(g_0,\tilde{g})}^{\tilde{g}} (s + qs) f(g) \, dg \right\} \\
- \lambda \left\{ \int_{g_1}^{h} (h - g) f(g) \, dg + p \int_{\min(g_1,\tilde{g})}^{\tilde{g}} (s + qs) f(g) \, dg \right\} \\
- C(p, g_0, g_1, \tilde{g}). \tag{18}
\]
\(^{14}\)For any given level of deterrence \( ps \), detection costs are reduced if \( s \) is raised and \( p \) is
reduced proportionally, with no effect on sanction costs.

\(^{15}\)A more general formulation would write the disutility of imprisonment as \( d(s) \),
\( d(0) = 0, d'(s) > 0 \); see Polinsky and Shavell (2007). In a strict liability regime, the
optimal sentence may then be less than the maximum feasible when \( d(s) \) is concave, i.e.,
individuals are “risk lovers” with respect to prison sentences. We abstract from this pos-
sibility because reputational concerns will introduce additional reasons for sanctions to be
less than maximal, both under strict liability and in a fault regime.
The first term inside the brackets is the social loss from inefficient behavior, allowing for the possibility of underdeterrence or overdeterrence. The second term inside the brackets is the social cost of sanctions. We focus on the main differences with Proposition 3.

**Proposition 4** With nonmonetary sanctions:

(i) If a fault-based regime is optimal, the legal standard may be above or below the social norm. When \( \hat{g} > h \), the sanction is maximal, \( p s_M = \hat{g} \) and all comply with the law; overdeterrence with a fault regime is optimal only if the cost of assessing fault is large but \( k \leq (1 + q)s_M \). When \( \hat{g} \leq h \), either the sanction is maximal, \( p(s_M + \beta \lambda) = \hat{g} \) and all comply with the law; or \( s \leq s_M, p(s + \beta \lambda) < \hat{g} \) and some of the nonprosocial do not comply.

(ii) If strict liability is optimal, the sanction may be less than \( s_M \) when the nonprosocial are undeterred, it is maximal otherwise; overdeterrence with a strict liability regime is optimal only if sanction costs are large but \( k \geq (1 + q)s_M \).

The equilibrium threshold of the nonprosocial solves

\[
g_0 = p(s + \beta \Delta(g_0, p)),
\]

where the legal stigma is written as a function of \( g_0 \) (as in the figures of Section 4) and of the probability of detection. For a given level of deterrence, the trade-off between the probability of detection and the sanction is

\[
\eta \equiv -\left. \frac{dp}{ds} \right|_{g_0=ct} = \frac{s}{s + \beta \Delta_p}
\]

where \( \Delta_p \) denotes the partial derivative. When there are no reputational concerns, \( \beta = 0 \) and therefore \( \eta = 1 \). This yields the argument in the standard model, i.e., it is always desirable to increase the sanction and reduce the probability of detection proportionally. When \( \beta > 0 \) but the optimal policy entails overdeterrence, the same argument applies because all individuals are then equally overdeterred, hence the legal stigma vanishes and \( \Delta_p = 0 \).
However, when the nonprosocial are underdeterred, the legal stigma is positive and $\Delta_p > 0$ because greater detection increases the information content of “no news”. In this case $\eta < 1$. Increasing the sanction and reducing the probability of detection so as to keep deterrence constant then reduces detection costs but also increases sanction costs. The net effect may be to increase costs, implying that the optimal sanction may be less than maximal when reputational effects are important.

Under a fault regime and by contrast with the case of monetary sanctions, the optimal legal standard may be above or below the social norm. In the corner solutions where everyone complies with the law, sanctions are at the maximum allowable level because they are never actually imposed. The nonprosocial are underdeterred if the standard $\hat{g} < h$. This possibility may arise because, while a higher standard would increase deterrence, the number of convictions would also increase.\(^{16}\) This does not matter when sanctions consist of fines, but with costly sanctions more convictions imply that sanction costs increase.

When the cost of assessing fault satisfies $k < (1 + q)s_M$, a strict liability regime with overdeterrence cannot be optimal. It would then be better to switch to fault-based offences, thereby substituting fault assessment costs to sanction costs. Conversely, when $k > (1 + q)s_M$, a fault regime with overdeterrence cannot be optimal because strict liability is cheaper.

**Socially costly stigmatization.** Historically one of the main arguments against strict liability welfare offences was the risk of stigmatizing respectable entrepreneurs.\(^{17}\) Similar arguments have been made in the wake of the recent criminal law reforms. Aversion to stigmatization risks can be captured by expressing the individuals’ reputational utility as a concave

\(^{16}\)Specifically, $\partial [F(\hat{g}) - F(g_0)]/\partial \hat{g} > 0$. See Figure 2.

\(^{17}\)See Paulus (1977) on the debates about regulations to counter food adulteration in mid 19th century Britain.
function of the beliefs about one’s type.\footnote{A similar approach has been used in other contexts, e.g., Köszegi (2006) and Dal Bó and Terviö (2013).} To facilitate comparison with our previous formulation, we write the reputational term as $\beta v(\mu)$ where $v$ is increasing and strictly concave with $v(0) = 0$ and $v(1) = 1$. The overall utility function of a type-$t$ individual is now

$$u_t = w - \gamma_t \max(e - e^*, 0) + \beta v(\mu), \quad t = 0, 1.$$  \hspace{1cm} (21)

A benevolent omniscient regulator would impose the same wealth maximizing action profile, but he would not disclose information about the individuals’ type because reputational gains and losses no longer cancel out. Accordingly the first-best welfare level is

$$W^* = w_0 + \int_h^\bar{\gamma} (g - h) f(g) \, dg + \beta v(\lambda).$$  \hspace{1cm} (22)

All of the results of Section 4 continue to hold provided the legal stigma is rewritten as $\Delta = v(\bar{\gamma}_N) - v(\bar{\gamma}_G)$. Assuming that enforcement relies on fines, welfare is now

$$W = W^* - \{(1 - \lambda) \int_{g_0}^h (h - g) f(g) \, dg + \lambda \int_{g_1}^h (h - g) f(g) \, dg\}$$

$$- \beta [v(\lambda) - \lambda \bar{\sigma}_1 - (1 - \lambda) \bar{\sigma}_0] - C(p, g_0, g_1, \bar{\gamma})$$  \hspace{1cm} (23)

where

$$\bar{\sigma}_t = p \min[F(\bar{\gamma}) - F(g_t), 0] v(\bar{\gamma}_N) + \{1 - p \min[F(\bar{\gamma}) - F(g_t), 0]\} v(\bar{\gamma}_N), \quad t = 0, 1.$$  \hspace{1cm} (24)

In equation (23) the term inside the brackets is the loss from inefficient behavior, again allowing for the possibility of overdeterrence. The third term is the deadweight loss from legal stigmatization. In the first best, reputational utility is $\beta v(\lambda)$ for all individuals. Under a given legal regime and enforcement policy, the average reputational utility is equal to $\beta \bar{\sigma}$ where

$$\bar{\sigma} \equiv (1 - \lambda) \bar{\sigma}_0 + \lambda \bar{\sigma}_1.$$
Because \( v \) is concave, \( \tau \leq v(\lambda) \) with strict inequality unless everyone behaves the same. Welfare as defined in (23) incorporates social aversion to stigmatization, introducing a tension between the usefulness of legal stigma for motivating appropriate behavior and the deadweight loss from stigmatization.

**Proposition 5** Under stigmatization aversion, an optimal legal regime and enforcement policy sets the fine at the maximum permissible.

(i) If liability is fault-based, the legal standard satisfies \( \bar{g} \leq h \) and the probability of detection satisfies \( p(s_M + \beta \lambda) \leq \bar{g} \).

(ii) If liability is strict, the nonprosocial are underdeterred.

In a strict liability regime, both good and bad citizens will at times choose not to comply with the law. Because the nonprosocial are underdeterred in the optimal policy, \( \bar{t}_N > \bar{t}_G \) at equilibrium and therefore \( \tau < v(\lambda) \), meaning that there is a social loss from legal stigmatization. This loss could be reduced by increasing the probability of detection because the equilibrium \( \tau \) is increasing in \( p \). Indeed the loss would vanish if the nonprosocial were made to behave like good citizens with an expected fine \( p s_M \geq h \). However

\[
\frac{\partial \tau}{\partial p}\bigg|_{p s_M = h} = 0,
\]

implying that the loss from stigmatization is of the second order when enforcement is marginally reduced from the level ensuring first-best deterrence.\(^{19}\)

Under a fault regime, the optimal enforcement policy is similar to part (i) of Proposition 3, except that the legal standard may be below the social

\(^{19}\)A parallel can be made with Polinsky and Shavell (1979) who discuss the optimal fine and enforcement policy when individuals are risk averse with respect to income. Optimal enforcement may then be non stochastic, thereby eliminating income risks. In the situation considered here, a policy with \( p = 1 \) would not eliminate the social loss from stigmatization unless of course \( s_M \geq h \).
norm. In the case of the corner solution, the loss from stigmatization disappears because everyone complies with the law. The legal standard may be less than the first-best because strengthening the standard then reduces the average reputational utility: while a stronger standard would increase deterrence, the consequence would also be that some of the nonprosocial no longer comply with the law. A standard $\hat{g} < h$ is also possible in the interior solution where some of the nonprosocial do not comply with the law because strengthening the standard increases the deadweight loss.

Figure 7: Legal standards and stigmatization

Figure 7 illustrates the advantage of a fault regime together with the possibility of a standard less than the social norm. Suppose the probability of detection $p$ is the best enforcement policy under a strict liability regime, yielding the deterrence curve $g_0(\Delta)$. We denote the stigma curves under the same probability of detection as $\Delta(g_0, \hat{g})$. The curve for strict liability is $\Delta(g_0, \bar{\gamma})$ and the equilibrium deterrence level is therefore $g'_0$. Switching to
a fault regime with the standard $\bar{g} = h$ would increase deterrence up to $g''_0$. However, this need not increase welfare because the loss from stigmatization may be much greater than in the initial situation under strict liability. In the Appendix B the Figures 8 and 9 provide examples of situations where, at comparable levels of deterrence, the social loss from stigmatization is substantially larger under a fault regime than under strict liability.\textsuperscript{20} Nevertheless, as shown in Figure 7, a fault regime does unambiguously better (assuming assessing fault is costless) if the legal standard is weakened to $\bar{g}_A$, which corresponds to the stigma curve $\Delta(g_0, \bar{g}_A)$. All individuals then comply with the law, so there is no stigmatization. Of course this fault regime and enforcement policy could presumably be improved further.

7 Concluding Remarks

Violating the law does not have the same social meaning under strict liability and fault-based offences. The latter is a stronger signal about one’s character. Fault-based offences will therefore usually perform better in harnessing reputational concerns for the purpose of motivating socially appropriate behavior. Nevertheless, the result does not always follow because the social meaning and incentive effects depend on the frequency of convictions.

In many situations, socially unwarranted behavior will be a rare event because most individuals are socially minded. It may also be that offenders are rare because the enforcement policy achieves substantial deterrence. A fault-based regime that seeks to harness reputational incentives should aim at reducing apparent unlawfulness. Not finding fault may then be banal, therefore posterior beliefs conditional on “no news” do not differ too much from the prior. But then convictions yield substantial disesteem. By contrast, when convictions would be a frequent event under a fault regime,

\textsuperscript{20}The upside is that fault provides greater incentives the more individuals are averse to reputational risks.
offences are banal. A strict liability regime would then perform better, as this increases the salience of offences, thereby increasing the significance of “no news”.

The argument is reminiscent of Bénabou and Tirole’s (2006, 2011) discussion of how acceptable behavior arises from the interplay of “honor” and “stigma”. High stigma is attached to a behavior that “is just not done”, only the worst type will do it. Alternatively, when “everyone does it”, the same behavior carries little stigma. But then “not doing it” yields prestige. In the case of legal regimes, whether a conviction imposes significant stigma or whether “no conviction” confers significant honor depends on the underlying situation but also on the legal regime itself together with enforcement possibilities.

We emphasized the information conveyed by offences under different legal regimes given a pre-existing social norm. One could also remark that different regimes have different “expressive content”. In our analysis, the underlying social norm was that individuals should be socially minded and behave accordingly. Under a fault regime, the norm can be “expressed” by the duty or obligation with respect to which fault is defined whereas strict liability is fuzzier in this respect. However, when enforcement relies on socially costly sanctions such as imprisonment or when stigmatization entails a deadweight loss, the legal norm may be weaker than the social norm. In either case there is a trade-off between preventing socially inefficient behavior and the social cost of finding fault. With costly legal sanctions, the legal norm may also be stronger than the social norm.

Strict liability and fault-based offences may differ in other ways with respect to expressive content. In particular, when individuals are imperfectly informed of the harm they may cause, a legal standard of behavior conveys information. Its prescriptive content may then help socially minded individuals to coordinate on socially appropriate conduct. Imitative behavior due to social or self image concerns then induces some bunching by the
nonprosocial on the socially appropriate behavior.

Appendix A

Proof of Lemma 1. The claim follows directly from (8) and (9). 

Proof of Lemma 2. Applying Bayes’ rule,

\[ T_N = \frac{\lambda [1 - p \max(F(\hat{g}) - F(g_1), 0)]}{1 - p [\lambda \max(F(\hat{g}) - F(g_1), 0) + (1 - \lambda)(F(\hat{g}) - F(g_0))]}, \tag{25} \]

\[ T_G = \frac{\lambda \max(F(\hat{g}) - F(g_1), 0)}{\lambda \max(F(\hat{g}) - F(g_1), 0) + (1 - \lambda)(F(\hat{g}) - F(g_0))}, \tag{26} \]

where (26) is undefined when \( g_0 = g_1 = \hat{g} \).

If \( g_0 < \hat{g} \leq h = g_1, \ T_G = 0 \) and therefore

\[ \Delta = T_N = \frac{\lambda}{1 - p(1 - \lambda)(F(\hat{g}) - F(g_0))}. \tag{27} \]

This is decreasing in \( g_0 \) with \( \Delta = \lambda \) when \( g_0 = \hat{g} \). If \( g_0 \leq g_1 = h < \hat{g} \),

\[ \Delta = \frac{\lambda [1 - p(F(\hat{g}) - F(h))]}{1 - p [\lambda(F(\hat{g}) - F(h)) + (1 - \lambda)(F(\hat{g}) - F(g_0))]} - \frac{\lambda(F(\hat{g}) - F(h))}{\lambda(F(\hat{g}) - F(h)) + (1 - \lambda)(F(\hat{g}) - F(g_0))}, \tag{28} \]

which is positive and decreasing in \( g_0 \) with \( \Delta = 0 \) when \( g_0 = h \). If \( h < g_0 = g_1 < \hat{g} \), (25) and (26) yield \( T_N = T_G = \lambda \) so that \( \Delta = 0 \). For \( h < g_0 = g_1 = \hat{g} \), we take the limit of the preceding result, so that again \( \Delta = 0 \). 

Proof of Proposition 1. Let \( \hat{g} \geq ps \). We first show uniqueness of the equilibrium. From Lemma 1 either \( g_0 = g_1 > h \) or \( g_0 \leq g_1 = h \). By Lemma 2, the first case implies \( \Delta = 0 \). Thus, it arises only if \( ps > h \) and the equilibrium is then simply \( g_0 = g_1 = ps \). A policy with \( ps \leq h \) therefore yields the second case. The relevant domain for \( g_0 \) is then the interval \([ps, \min(\hat{g}, h)]\). If \( ps = \min(\hat{g}, h) \), the equilibrium is trivially \( g_0 = ps \), so let \( ps < \min(\hat{g}, h) \). From (8) the equilibrium \( g_0 \) is a solution to

\[ g_0 = \min[\hat{g}, p(s + \beta \Delta(g_0, \hat{g}, p))] \tag{29} \]
where $\Delta(g_0, \hat{g}, p)$ is defined by (27) or (28) for the cases $\hat{g} \leq h$ or $\hat{g} > h$ respectively. Equivalently, the equilibrium $g_0$ solves

$$
\varphi(g_0) \equiv \min [\hat{g}, p(s + \beta\Delta(g_0, \hat{g}, p))] - g_0 = 0, \ g_0 \in [ps, \min(\hat{g}, h)],
$$

(30)

where $\varphi(g_0)$ is a continuous function. By Lemma 2, $\Delta(ps, \hat{g}, p) > 0$ and therefore $\varphi(ps) > 0$. For the case $\hat{g} \leq h$, $\Delta(\hat{g}, \hat{g}, p) = \lambda > 0$ and therefore $\varphi(\hat{g}) \leq 0$. Because $\Delta(g_0, \hat{g}, p)$ is strictly decreasing in $g_0$ in the relevant domain, so is $\varphi(g_0)$ and the equilibrium is therefore unique. For the case $\hat{g} > h$, $\Delta(h, \hat{g}, p) = 0$ and $\varphi(h) < 0$. Again $\varphi(g_0)$ is strictly decreasing, ensuring uniqueness.

(i) The claim for the case $ps \geq h$ follows directly from the above argument.

(ii) For $ps < \hat{g} \leq h$, the above argument shows that $g_1 = h$ and $g_0 \in (ps, \hat{g})$. If $p(s + \beta\lambda) \geq \hat{g}$, $\varphi(\hat{g}) = 0$ and the equilibrium satisfies $g_0 = \hat{g}$. If $p(s + \beta\lambda) < \hat{g}$, $\varphi(\hat{g}) < 0$ and the equilibrium is $g_0 < \hat{g}$ solving $g_0 = p(s + \beta\Delta(g_0, \hat{g}, p))$. Differentiating totally with respect to $\hat{g}$ and $p$ yields

$$
\frac{\partial g_0}{\partial \hat{g}} = \frac{p\beta \Delta_{\hat{g}}}{1 - p\beta \Delta_{g_0}},
$$

(31)

$$
\frac{dg_0}{dp} = \frac{s + \beta \Delta + p\beta \Delta_p}{1 - p\beta \Delta_{g_0}}.
$$

(32)

From (27), $\Delta_{g_0}$ is negative while $\Delta_p$ and $\Delta_{\hat{g}}$ are positive. Hence (31) and (32) are both positive. To complete the argument, when $p(s + \beta\lambda) \geq \hat{g}$, $g_0 = \hat{g}$ and is then also increasing in $\hat{g}$.

(iii) For $ps < h < \hat{g}$, the argument is similar except that the solution now satisfies $g_0 \in (ps, h)$. Differentiating (29) totally with respect to $\hat{g}$ and $p$ again yields (31) and (32) but with $\Delta$ now defined by (28). The signs of $\Delta_{g_0}$ and $\Delta_p$ are as before but that of $\Delta_{\hat{g}}$ is now ambiguous (see the proof of Proposition 2). Thus (32) is positive but the sign of (31) is ambiguous.

How $g_0$ varies with $s$ is derived similarly and is left to the reader.
Proof of Proposition 2. Let \( g_0(\hat{g}) \) denote the equilibrium as derived in Proposition 1 for some \( ps < h \), so that \( g_0(\hat{g}) \leq g_1 = h \). The deterrence maximizing legal regime solves \( \max \hat{g} g_0(\hat{g}) \). We consider separately the possibility that the solution satisfies \( \hat{g}^* \leq h \) or \( \hat{g}^* > h \).

If \( \hat{g}^* \leq h \), the function \( g_0(\hat{g}) \) satisfies part (ii) of Proposition 1 and is strictly increasing, therefore \( \hat{g}^* = h \). If \( \hat{g}^* > h \), the function \( g_0(\hat{g}) \) satisfies part (iii) of Proposition 1, hence \( g_0(\hat{g}) < h \). Either \( \hat{g}^* = \bar{g} \) or \( \hat{g}^* \) is an interior solution in \((h, \bar{g})\). In the latter case, recalling (31), the solution must satisfy the first-order condition

\[
\frac{\partial g_0(\hat{g})}{\partial \hat{g}} \bigg|_{\hat{g}=\hat{g}^*} = \frac{p\beta \Delta_{\hat{g}}(g_0(\hat{g}^*), \hat{g}^*)}{1 - p\beta \Delta_{g_0}(g_0(\hat{g}^*), \hat{g}^*)} = 0, \tag{33}
\]

where the right-hand-side is as in (31) but with \( \Delta \) defined as in (28). The second-order necessary condition is that

\[
\frac{\partial^2 g_0(\hat{g})}{\partial \hat{g}^2} \bigg|_{\hat{g}=\hat{g}^*} = \frac{p\Delta_{\hat{g}^2}(g_0(\hat{g}^*), \hat{g}^*)}{1 - p\Delta_{g_0}(g_0(\hat{g}^*), \hat{g}^*)} \leq 0. \tag{34}
\]

be non positive, where the expression is obtained given that (33) holds and therefore \( \Delta_{\hat{g}}(g_0(\hat{g}^*), \hat{g}^*) = 0 \). Because the denominator in (34) is positive, the second-order condition requires \( \Delta_{\hat{g}^2}(g_0(\hat{g}^*), \hat{g}^*) \leq 0 \). From (28),

\[
\Delta_{\hat{g}^2}(g_0(\hat{g}^*), \hat{g}^*) = \frac{2p\lambda(1 - \lambda)(F(h) - F(g_0(\hat{g}^*)))}{\Phi(1 - p\Phi)^2} > 0 \tag{35}
\]

where

\[
\Phi = F(\hat{g}) - \lambda F(h) - (1 - \lambda)F(g_0(\hat{g}^*)).
\]

Thus, the necessary condition does not hold, implying that the corner solution \( \hat{g}^* = \bar{g} \) is the only possibility. \( \blacksquare \)

Proof of Lemma 3. Solve (27) and (28) in the proof of Lemma 2 for the value of \( g_0 \) consistent with the same \( \Delta \) under either strict liability regime or the fault regime with \( \hat{g} = h \). This yields

\[
F(g_0) = \frac{1}{2(1 - \lambda)} \left((1 - 2\lambda)(1 - \frac{1 - \frac{p}{\Phi}}{p})\right). \tag{36}
\]
The equation has a solution \( F(g_0) > 0 \) only if the condition in Lemma 3 holds.

**Proof of Proposition 3.** We first show that \( ps < h \). Suppose to the contrary that \( ps \geq h \). Proposition 1 then implies \( g_0 = g_1 = ps \). If assessing fault is costless the optimal regime is then indifferent, otherwise it must be strict liability. In either case, using (10),

\[
\frac{\partial W}{\partial p} = (1 - \lambda)(h - g_0) f(g_0) \frac{\partial g_0}{\partial p} + \lambda (h - g_1) f(g_1) \frac{\partial g_1}{\partial p} - c'(p). 
\] (37)

For \( ps > h \), \( \partial g_0/\partial p > 0 \) and \( \partial g_1/\partial p > 0 \) and (37) is negative. At \( ps = h \), the preceding derivatives are discontinuous but then

\[
\left. \frac{\partial W}{\partial p} \right|_{ps=h} = -c'(p) < 0.
\]

Thus, an optimal policy entails \( ps < h \). By Proposition 1, this implies \( g_0 \leq g_1 = h \).

Next we show that \( s = s_M \). From (10), under any legal regime, a policy change that reduces \( p \) with no change in \( g_0 \) is beneficial because

\[
\left. \frac{\partial W}{\partial p} \right|_{g_0=ct} = -C_p < 0.
\]

If \( \hat{g} > h \), \( p \) and \( s \) solve

\[
g_0 = p(s + \beta \Delta(g_0, \hat{g}, p)).
\] (38)

where \( \Delta_p > 0 \). If \( \hat{g} \leq h \), either \( g_0 < \hat{g} \) and \( p \) and \( s \) solve (38); or \( g_0 = \hat{g} \) and \( p(s + \beta \lambda) \geq \hat{g} \). In all of these cases, it is possible to reduce \( p \) and increase \( s \) with no change in \( g_0 \). Hence the optimal fine is maximal.

Finally, we now show that, given \( ps_M < h \), the optimal legal regime is deterrence maximizing. From (10),

\[
\frac{\partial W}{\partial g_0} = (1 - \lambda)(h - g_0) f(g_0) - C_{g_0}
\]
where $C_{g_0}$ is zero under strict liability and is negative if the regime is fault-based and assessing fault is costly. Because welfare is strictly increasing in $g_0$ for all $g_0 \leq h$, Proposition 2 implies that the optimal regime is either strict liability or the fault regime with $\bar{g} = h$.

(i) Suppose the optimal regime is fault-based. The possibility of first-best deterrence with $p(s_M + \beta \lambda) = h$ is discussed in the text. Otherwise, $p(s_M + \beta \lambda) < h$ and $g_0 < h$. We now prove (14). If fault-based liability is optimal for the probability of detection $p$, we must have
\[
\frac{\partial g_0}{\partial g} \bigg|_{\bar{g} = h} = \frac{p\beta \Delta \bar{g}(g_0, h, p)}{1 - p\beta \Delta \bar{g}(g_0, h, p)} \leq 0.
\]
where the expression is a right-derivative and is the same as (33) in the proof of Proposition 2. Now, the inequality must be strict because, as shown in the proof of the same proposition,
\[
\Delta \bar{g}(g_0, h, p) > 0.
\]
Recalling that $\Delta(h, \bar{g}, p) = 0$, $\varphi(h) \equiv p \Delta \bar{g}(g_0, h, p) - g_0 = 0$. Recalling that $\Delta(h, \bar{g}, p) = 0$, $\varphi(h) \equiv p \Delta(g_0, h, p) - g_0 = 0$. Because $\varphi(g_0)$ is strictly decreasing, the equilibrium satisfies $g_0 < h$.

(ii) $p s_M < h$ implies $g_0 \leq g_1 = h$. Under strict liability $g_0$ solves
\[
\varphi(g_0) \equiv p(s_M + \beta \Delta(g_0, \bar{g}, p)) - g_0 = 0.
\]
Recalling that $\Delta(h, \bar{g}, p) = 0$, $\varphi(h) \equiv p s_M - h < 0$. Because $\varphi(g_0)$ is strictly decreasing, the equilibrium satisfies $g_0 < h$.

(iii) If $k = 0$, for any $p$ the best regime is the one that maximizes deterrence. By Lemma 3, this is always the fault regime if $\lambda \geq 1/2$. When $s_M$ or $\beta$ are sufficiently large, $g_0 = h$ obtains with $p \leq 1/2$ satisfying $p(s_M + \beta \lambda) = h$, e.g., when $s_M \geq 2h$ or $\beta \lambda \geq 2h$. An optimal policy then necessarily satisfies $p \leq 1/2$. Condition (13) in Lemma 3 requires $p > 1/2$. Therefore,
for \( p \leq 1/2 \), the fault regime dominates strict liability in terms of deterrence. By continuity, the same argument applies if \( k \) is positive but not too large. To conclude, we prove (15). Using the same argument as in (i), if the optimal regime is strict liability with \( \hat{g} = \overline{g} \), it must be the case that \( \Delta_{\hat{g}}(g_0, \overline{g}, p) \geq 0 \), otherwise deterrence would be maximized with the fault regime \( \hat{g} = h \). From (28), this is equivalent to condition (15).

Before proving the next propositions, we derive a result for corner equilibria.

**Lemma 4** Let \( p^*(s_M + \beta \lambda) = \hat{g}^* \leq h \) so that \( g_0(p, \hat{g}, s_M) = \hat{g}^* \) when \((p, \hat{g}) = (p^*, \hat{g}^*)\). Then the right and left derivatives satisfy

\[
\begin{align*}
\frac{\partial g_0}{\partial \hat{g}} & \bigg|_{(p^*, \hat{g}^*)}^{+} = \frac{p^* \beta (1 - \lambda) f(\hat{g}^*)}{1 + p^* \beta (1 - \lambda) f(\hat{g}^*)}, \\
\frac{\partial g_0}{\partial \hat{g}} & \bigg|_{(p^*, \hat{g}^*)}^{-} = 1, \\
\frac{\partial g_0}{\partial p} & \bigg|_{(p^*, \hat{g}^*)}^{+} = 0, \\
\frac{\partial g_0}{\partial p} & \bigg|_{(p^*, \hat{g}^*)}^{-} = \frac{s_M + \beta \lambda}{1 + p^* \beta (1 - \lambda) f(\hat{g}^*)}.
\end{align*}
\]

**Proof:** By Proposition 1, when \( \hat{g} < h \) and \( p(s + \beta \lambda) \geq \hat{g} \), \( g_0 = \hat{g} \) which implies (41) and (42). When \( p(s + \beta \lambda) < \hat{g} \), \( \partial g_0 / \partial \hat{g} \) and \( \partial g_0 / \partial p \) satisfy (31) and (32) respectively where \( \Delta \) satisfies (27). From the latter it is easily seen that

\[
- \Delta_{g_0} = \Delta_{\hat{g}} = p(1 - \lambda) f(\hat{g}) \quad \text{and} \quad \Delta_{\hat{g}} = 0 \quad \text{for} \quad g_0 = \hat{g} \leq h.
\]

Substituting in (31) and (32) then yields (40) and (43).

**Proof of Proposition 4.**

(i) If the optimal regime is fault-based and overdeters, the policy is the same as in the conventional model so that \( s = s_M \) and \( p s_M = \hat{g} \). Sanction costs are not incurred but assessing fault imposes the cost \( pk(1 - F(p s_M)) \).
If \( k > (1+q)s_M \), this cannot be optimal because strict liability would impose the smaller sanction cost \( p(1+q)s_M(1-F(p_s)) \).

If the optimal regime does not overdeter, \( \tilde{g} \leq h \) otherwise unnecessary sanction costs are incurred. At a corner solution, with \( g_0 = \tilde{g} \) and \( p(s+\beta \lambda) = \tilde{g} \), increasing \( s \) and reducing \( p \) while preserving the preceding equality has no effect on sanctions (which are not incurred) but reduces the enforcement cost

\[
c(p) + pk[\lambda(1 - F(h)) + (1 - \lambda)(1 - F(\tilde{g}))].
\]

Hence the sanction must be maximal.

To see that \( \tilde{g} < h \) is a possibility, differentiate welfare in (18) with respect to \( \tilde{g} \). At the corner solution \( p(s_M + \beta \lambda) = \tilde{g}^* < h \),

\[
\frac{\partial W}{\partial \tilde{g}} = (1-\lambda)f(\tilde{g}^*) \left[ (h+kp - \tilde{g}^*) \frac{\partial g_0}{\partial \tilde{g}} - p(1+q)s_M \left( 1 - \frac{\partial g_0}{\partial \tilde{g}} \right) \right].
\]

Substituting from Lemma 4 yields

\[
\text{sign} \left. \frac{\partial W}{\partial \tilde{g}} \right|_{\tilde{g}^*} = \text{sign} \left\{ p(1-\lambda)(h+kp - \tilde{g}^*)f(\tilde{g}^*)\beta - (1+q)s_M \right\}.
\]

If \( k = 0 \), this is negative for \( \tilde{g}^* \) sufficiently close to \( h \). With \( k \) positive, it also negative if \( \beta \) is small and \( s_M \) is large.

At an interior solution where \( g_0 < \tilde{g} \), the argument for the possibility that \( \tilde{g} < h \) is similar. To see that \( s < s_M \) is then also a possibility, set \( k = 0 \) for simplicity, so that at the policy \((p, \tilde{g}, s)\)

\[
\frac{\partial W}{\partial p} = (1-\lambda) \left[ (h+p(1+q)s - g_0) f(g_0) \frac{\partial g_0}{\partial p} - (1+q)s \int_{g_0}^{\tilde{g}} f(g) \, dg \right] - c'(p)
\]

\[
= 0.
\]

The derivative with respect to \( s \) is

\[
\frac{\partial W}{\partial s} = (1-\lambda) \left[ (h+p(1+q)s - g_0) f(g_0) \frac{\partial g_0}{\partial s} - p(1+q) \int_{g_0}^{\tilde{g}} f(g) \, dg \right].
\]
Substituting from (45) in (46) yields

\[ \frac{\partial W}{\partial s} = \frac{\theta pc'(p)}{s} - (1 - \theta)(1 - \lambda)p(1 + q) \int_{g_0}^{\hat{g}} f(g) dg \]  

(47)

where

\[ \theta = \frac{s \partial g_0/\partial s}{p \partial g_0/\partial p} = \frac{s}{s + \beta \Delta + p \beta \Delta_p}. \]

The equality follows from (31) and (32); (27) implies \( \Delta_p > 0 \) when \( g_0 < \hat{g} \). Substituting back in (47) yields

\[ \text{sign} \left( \frac{\partial W}{\partial s} \right) = \text{sign} \left[ c'(p) - \beta(\Delta + p \Delta_p)(1 - \lambda)(1 + q) \int_{g_0}^{\hat{g}} f(g) dg \right]. \]

Thus, when \( \beta \) is large, the sign may be negative at \( s = s_M \).

(ii) The arguments for \( s = s_M \) when strict liability overdeters and for \( s \leq s_M \) are similar to that under a fault regime. The argument that strict liability with overdeterrence can only be optimal if \( (1 + q)s_M \leq k \) is symmetric to the one in (i). □

Proof of Proposition 5.

(i) We only prove the possibility that \( \hat{g} < h \). When \( \hat{g} \leq h, g_0 \leq g_1 = h \).

Setting \( k = 0 \) for simplicity, welfare in (23) reduces to

\[ W = W^* - (1 - \lambda) \int_{g_0}^{h} (h - g) f(g) dg - \beta[v(\lambda) - \overline{v}] - c(p) \]

where

\[ \overline{v} = [\lambda + (1 - \lambda)(1 - p(F(\hat{g}) - F(g_0))] v(\overline{T}_N) \]  

(48)

and \( \overline{T}_N \) satisfies (27).

Consider the set of corner policies \((p, \hat{g})\) with \( g_0 = \hat{g} \) and

\[ p(s_M + \beta \lambda) = \hat{g}. \]  

(49)

A necessary condition for a policy in this set to be optimal is

\[ \frac{dW}{dp} = (1 - \lambda)(h - \hat{g})f(\hat{g})(s_M + \beta \lambda) - c'(p) = 0. \]  

(50)
Let \((p^*, \tilde{g}^*)\) satisfy (49) and (50) and note that \(\tilde{g}^* < h\). For this policy to be optimal it must also not be beneficial to move to an interior solution by changing either \(p\) or \(\tilde{g}\).

The gain from a marginal change in \(\tilde{g}\) while keeping \(p = p^*\) is

\[
\frac{\partial W}{\partial \tilde{g}} = (1 - \lambda)(h - \tilde{g}^*)f(\tilde{g}^*)\frac{\partial g_0}{\partial \tilde{g}} + \beta \frac{\partial \pi}{\partial \tilde{g}}. \tag{51}
\]

where the notations refer to either the right or left derivatives. From (48) and (27)

\[
\frac{\partial \pi}{\partial \tilde{g}} = -p^*(1 - \lambda)f(\tilde{g}^*) \left(1 - \frac{\partial g_0}{\partial \tilde{g}}\right)v(\lambda) + v'(\lambda)\frac{\partial T_N}{\partial \tilde{g}}
\]

\[
= -p^*(1 - \lambda)f(\tilde{g}^*) \left(1 - \frac{\partial g_0}{\partial \tilde{g}}\right) \left(v(\lambda) - \lambda v'(\lambda)\right). \tag{52}
\]

Substituting from (52) in (51) and recalling Lemma 4,

\[
\left.\frac{\partial W}{\partial \tilde{g}}\right|_{(p^*, \tilde{g}^*)} = (1 - \lambda)(h - \tilde{g}^*)f(\tilde{g}^*)\frac{\partial g_0}{\partial \tilde{g}} > 0,
\]

\[
\text{sign} \left.\frac{\partial W}{\partial \tilde{g}}\right|_{(p^*, \tilde{g}^*)} = \text{sign} \left\{p^*(1 - \lambda)(h - \tilde{g}^*)f(\tilde{g}^*) - (v(\lambda) - \lambda v'(\lambda))\right\}.
\]

Thus a marginal change in \(\tilde{g}\) is not beneficial if

\[
p^*(1 - \lambda)(h - \tilde{g}^*)f(\tilde{g}^*) \leq v(\lambda) - \lambda v'(\lambda), \tag{53}
\]

where the right-hand side is positive by the concavity of \(v\).

Similarly the gain from a marginal change in \(p\) while keeping \(\tilde{g} = \tilde{g}^*\) is

\[
\frac{\partial W}{\partial p} = (1 - \lambda)(h - \tilde{g}^*)f(\tilde{g}^*)\frac{\partial g_0}{\partial p} + \beta \frac{\partial \pi}{\partial p} - c'(p^*) \tag{54}
\]

where

\[
\frac{\partial \pi}{\partial p} = p^*(1 - \lambda)f(\tilde{g}^*) \left[v(\lambda) - \lambda v'(\lambda)\right] \frac{\partial g_0}{\partial p}. \tag{55}
\]

Then

\[
\left.\frac{\partial W}{\partial p}\right|_{(p^*, \tilde{g}^*)} = -c'(p^*) < 0.
\]
Substituting from (55) and (50) in (54) and again using Lemma 4,

\[ \text{sign} \left( \frac{\partial W}{\partial p} \right)_{(p^*, \hat{\beta}^*)} = \text{sign} \left\{ -p^*(1 - \lambda)(h - \hat{g}^*)f(\hat{g}^*) + (v(\lambda) - \lambda v'(\lambda)) \right\}. \]

Thus (53) also ensures that it is not beneficial to marginally change the probability of detection.

(ii) Under strict liability with the policy \( ps_M = h \), the effect of a marginal decrease in \( p \) is given by the left derivative

\[ \frac{\partial W}{\partial p} \bigg|_{ps_M = h} = \beta \frac{\partial \varphi}{\partial p} - c'(p) \]  \hspace{1cm} (56)

where

\[ \varphi = p[(1 - \lambda)(1 - F(g_0) + \lambda(1 - F(h))] \nu(\bar{r}_G) \]
\[ + \{1 - p[(1 - \lambda)(1 - F(g_0) + \lambda(1 - F(h))]\} \nu(\bar{r}_N). \]

\( \bar{r}_N \) and \( \bar{r}_G \) are defined in (25) and (26) with \( \hat{g} = \varphi \) and \( g_0 \) solves (29). It is then easily verified that

\[ \frac{\partial \varphi}{\partial p} \bigg|_{ps_M = h} = 0, \]

hence (56) is negative, implying that \( p \) should be reduced.
Appendix B

The Figures 8 and 9 illustrate the effect of aversion to reputational risk on reputational incentives and on the social loss from stigmatization.

Fig. 8: $\bar{\xi}_N - \bar{\xi}_G$, $\nu(\bar{\xi}_N) - \nu(\bar{\xi}_G)$ and $\nu(g_0)/\nu$ as a function of deterrence ($\lambda = .2$, $p = .8$, $F(h) = .8$, $\nu(\mu) = \sqrt{\mu}$)
Fig. 9: $\ell_{m} - \ell_{G}$, $\nu(\ell_{m}) - \nu(\ell_{G})$ and $\nu(g_{0})/\nu$ as a function of deterrence
$(\lambda = .4, \ p = .6, \ F(h) = .8, \ \nu(\mu) = \sqrt{\mu})$
References


