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

Bruno Deffains* Claude Fluet†

October 2014

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. However, this need not always be the case and depends in a complex way on the salience of social norms and on enforcement costs. An optimal legal regime and enforcement policy entails fault-based offences when detected violations of the law would seldom occur and strict liability offences otherwise.

KEYWORDS: Normative motivations, regulatory offences, other-regarding behavior, law enforcement, strict liability, fault, compliance, deterrence. (JEL: D8, K4, Z13)

*Université Panthéon Assas and Institut Universitaire de France. E-mail: Bruno.Deffains@u-paris2.fr

†Université Laval and CIRPEE. E-mail: Claude.Fluet@fsa.ulaval.ca

1 Introduction

Illegal behavior ranges from crimes of great antiquity, such as murder or rape, carrying strong moral opprobrium down to lesser 'quasi-crimes', e.g., false or misleading advertising, income underreporting in tax filings, discharges of pollutants, fishing out of season, and the like. An important issue in legal or regulatory design is the categorization of offences. Should they be criminalized or qualified as mere violations punished at most by a fine? The issue is one which legal systems have been dealing with 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. To give but one example, the Model Penal Code of the American Law Institute rejected the principle of strict liability in criminal law. 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*[1]. *Malum in se* means wrong or reprehensible in itself. The expression refers to conduct viewed as inherently wrong independently of regulations or laws governing the conduct. For example, murder and theft would be wrong regardless of the law. By contrast, *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 merely because they have been declared illegal by statute law. The distinction is important in most penal systems, if only implicitly. Obviously, many

[1] Another way to describe the underlying conceptual difference between *malum in se* and *malum prohibitum* is *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*).

acts that are *malum in se* are also formally prohibited by legislation. However, their wrongness — indeed their very "illegality" so to speak— exists independently of legal prohibitions. If, say, a country repelled its legal prohibition on murder, murderers would presumably remain wrongdoers from society's point of view because a fundamental social norm is at stake. Violators of this norm would face social sanctions through stigmatization, moral opprobrium and possibly other more direct means. The outcome would be different for 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. 'True crimes', which include traditional crimes such as murder or theft, are always fault-based. In most legal systems, criminal law provides that offences have physical elements (e.g., conduct) and fault elements such as intention and knowledge. Individuals are found guilty only when they have the requisite *mens rea* (i.e., "guilty mind") and are therefore morally blameworthy. By contrast, many lesser crimes and most regulatory offences do not require a *mens rea*. For example, selling alcohol to minors is a strict liability offence because a person can be convicted even if she believed the customers were old enough to consume alcohol. Another example is the contempt of court act: publishing information which can prejudice an active court case is a criminal offence regardless of whether or not the individual involved meant any harm. Similarly, traffic offences are usually on a strict liability basis: a driver still gets a speeding ticket even if he reasonably believed he was driving within the speed limit.

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". In the present paper, we discuss the issue of legal design from the perspective of harnessing normative motivations. For this purpose, we extend the economic model of public enforcement of law to incorporate normative motivation and pre-existing social norms of conduct. At one extreme, pre-existing social norms concerning a particular act are very weak, so that the policy prescription would be the same as in the standard model. 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 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.

Consider a situation where formal legal sanctions underdeter. Feasible fines are bounded and 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. Other individuals 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. Social enforcement requires one's actions to be observable by one's peers or reference group (family, friends and the like). To the extent that informal social sanctions suffice, legal enforcement is of course superfluous. We focus on situations where an individual's actions are not directly observable by society at large or one's peers. However, legal sanctions provide public information from which inferences can be drawn about the individuals' actions and therefore their intrinsic predispositions. Under either fault-based or strict liability offences, social and self image concerns provide the non

prosocial individuals with some incentives to mimic the virtuous. The issue is how this influences the design of offences and the enforcement policy, i.e., resources spent on detecting violations of the law.

A basic result is that fault-based offences tend to be more effective in harnessing reputational concerns. The reason is that being found guilty of an offence is then more informative. A strict liability offence merely ascertains that the violator committed a potentially harmful action and says nothing about the circumstances in which the action was committed. By contrast, 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" involves negligible additional costs. When the social norm is non-existent and legal sanctions are restricted to socially costless fines (pure transfers), fault-based and strict liability offences are equally efficient. However, in intermediate situations with weak social norms, it is not always the case that fault-based offences do better than strict liability offences in harnessing reputational concerns. Which regime is better depends in a complex way on the underlying situation. We show that the optimal legal regime and enforcement policy are interdependent and entail fault-based offences when detected violations of the law would seldom occur and strict liability offences otherwise.

The dichotomy between fault-based and strict liability offences captures an important distinction between "criminalized" offences and purely "regulatory" offences. In our analysis, the legal design problem is approached from a standard utilitarian perspective, in the sense that opprobrium and the reputational effects of legal sanctions are only considered for their incentive effects. When offences are fault-based, social sanctions tend to be harsher because of more precise reputational effects. It follows that fault-

based offences do better for acts that are clearly bad from a moral and/or social point of view. By contrast, when there is no pre-existing norm, strict liability does as well. 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 relevant litterature. 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. Section 6 concludes. Proofs are in the Appendix.

2 Literature review

A recent microeconomic literature has emphasized that one's actions may signal something about unobservable predisposition and that some predispositions are socially valued (see e.g. Bernheim, 1994; Bénabou and Tirole 2006, 2011; Daughety and Reinganum, 2010; Deffains and Fluet, 2013). Numerous experimental or field studies have also shown 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 substituability of legal and nonlegal sanctions, pointing out that various nonlegal sanctions, such as 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). Other aspects of the literature focus on the potential complementarity of 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). A specific field of the literature studies the relationship between morality and law. Cooter (1998) analyzes how law promotes individual incentive to acquire morality and self-control. Posner (1997) explains how law complements or substitutes for social norms. Shavell (2002) 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.

In a civil law context, Deffains and Fluet (2013) analyze how liability rules (under the form of damages paid to the plaintif) for misbehavior and social pressure interact to provide incentives to take care and what are the impacts of those interactions on the structure of the legal system. The main point made is that the negligence rule tends to be more efficient than strict liability. The logic is that under the negligence rule, courts assess a person's level of care. Hence, if he is found to be negligent, it will be impossible or unlikely that he is a prosocial person. Thus, for people who want to be viewed as prosocial, the negligence rule provides a reputational incentive to take optimal care. Under strict liability, however, no direct information is provided by the legal system about the level of care—a person who is truly prosocial and took care could have caused an accident nevertheless and be found strictly liable. Hence, the reputational incentive provided by strict liability to take more care, for people who value their reputation, is muted. The focus of that paper is the extent to which formal legal sanctions crowd-out or crowd-in informal motivations under different liability rules.

The question of the interaction between law and (moral or social) norms is also important in criminal law. For instance, McAdams and Rasmusen (2007) and Shavell (2002) provide a general discussion of legal sanctions versus informal motivation as regulators of conduct. Fault-based offences also bear a relation to the concept of "expressive law". According to this

view even "mild law", i.e., law backed by small sanctions or poorly enforced, can have desirable effects on behavior; see Cooter (1998) and the discussion in Tyran and Feld (2006). Finally, the discussion is also related 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. Iaccubuci (2014) shows that reputational sanctions are not independent of legal penalties but rather in part depend on them. Changing legal penalties to account for the reputational effects of misconduct changes the reputational effects. The analysis suggests that optimizing legal penalties in light of reputational effects is a complex task but a necessary one to improve compliance with legal rules and regulations. Mialom (2014) also demonstrates that morality can enhance the effectiveness of law enforcement and that legal rules are necessary when moral rules are easily swayed by social influence, especially when such a commitment to regulate behavior is desirable in the long run

Our paper aims to contribute to this strand of the literature which views incentives, legal standards, social norms in a encompassing framework. We develop a framework to interpret, from an economic perspective, some of the main characteristics of criminal law . Specifically, 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 approach this issue by comparing 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.

Tradionally, offences require proof of one or more "physical elements" – the fact that someone did, or failed to do, something, or that something did or did not, in fact, happen. Most offences also require proof of a "fault element", or what is sometimes called the "mental element", in relation to a person's conduct. This refers to the state of mind of the person charged

with the offence. Fault elements include intention, knowledge, recklessness or negligence.

From a "legal theory " perspective, there are strong arguments for requiring fault in criminal offences (for a synthesis see e.g. Simester, 2005). One important justification refers to the "rule-of-law" concept. For instance, Hart considers that it is wrong to convict anyone who has not had "a fair opportunity" to exercise the capacity for "doing what the law requires and abstaining from what it forbids". Raz adds that "respecting human dignity entails treating humans as persons capable of planning and plotting their future". Finaly, Duff explains "strict liability is therefore both unjust and dishonest: it portrays as proven culpable wrongdoers those who have not been proved to be that". Most legal traditions strongly supports the principle that people should not generally be subject to criminal action in the absence of at least some blameworthy mental state, ranging from the deliberate intention to do what they have done, to the negligent failure to take care.

However, a significant and increasing number of offences imposes strict liability with respect to all physical elements, dispensing with any requirement of proof of fault. One could consider this evolution as the result of a "new principle" – different from the fault-based one. This principle recognises that certain kinds of offences – particularly those concerned with public safety and health – should not require that there be proof of any "blameworthy" mental state. Most strict liability crimes introduced to regulate such offences are created by statute.

Many scholars argue that there is a role for strict liability in criminal law, in relation to both regulatory offences and offences of social danger (see e.g. Horder, 2005). They argue that the interests of society as a whole can sometimes justify the imposition of liability without fault. There are at least two good reasons for this. Firstly, because one of the major functions of the criminal law is to deter certain forms of misconduct, and to make it clear

that conduct jeopardising public safety and health may be punishable in the absence of any blameworthy mental state, it sends a strong deterrence message ((Schaeffer and Muller-Langer, 2008). Secondly because, as a practical matter, it would be so extraordinarily difficult to prove that a person actually intended to act so as to jeopardise public safety or health, successful prosecutions would be so unlikely that the law would be meaningless.

One of the most common forms of strict liability offences of this kind are simple motor vehicle laws. These are intended to underpin public safety, and their utility depends upon the absence of a need to prove "fault" when they are contravened. Civil aviation safety regulations are strict liability offences for essentially the same reasons.

Consequently, the debate between strict liability and fault in criminal offences is a complex one. The legal doctrine sometimes looks for a compromise and tries to define the appropriate degree of fault according to different categories of crimes. For instance, Thomson (1994) explains how the Canadian Suprem Court relies on the notion of stigma as the primary factor to be considered in determining the constitutionally required degree of fault for crimes such as murders. These debates certainly shed some light on the fault-based vs strict liability controversy for criminal offences. However, one needs to go further by identifying the signaling effects of the different legal mechanisms when individuals are influenced by social norms of conduct and concerns about social approval and self-image.

3 Set-Up

We start by reviewing the basic public law enforcement model, borrowing from Polinsky and Shavell (2000, 2007). Next we extend this model to incorporate other-regarding behavior. The purpose of the model is to analyze the use of public agents to detect and to punish violators of legal rules. In this context, an individual will commit a harmful act if and only if his gain

from doing so exceeds the fine that is imposed by the public authority in case of violation.

The standard model. Risk-neutral individuals can obtain a private gain g from committing an act causing an external harm of amount h. The gain — equivalently the opportunity cost of not committing the act — varies among individuals and depends on the circumstances. 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 the individuals' behavior by $e \in \{0, 1\}$, where $e = 1$ means commission of the act[2], and interpreting $e(g)$ as behavior in the circumstance g, social welfare is

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

Socially optimal behavior is therefore

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

The harmful act is a strict liability offence if it is illegal irrespective of circumstances. We refer here to offenses such as traffic violations, fraudulent advertising, income tax underreporting, illegal parking... The sanction for violating the law is a fine s, a socially costless transfer of money. The enforcement policy is summarized by the probability p of detecting violations. The per capita enforcement expenditure is $c(p)$ with derivatives $c' > 0$, $c'' \geq 0$. An individual will commit the harmful action if his gain from doing so exceeds the expected fine, $g \geq ps$. For a given enforcement policy, welfare is therefore

$$\int_{ps}^{\overline{g}} (g-h)\,f(g)\,dg - c(p).$$

[2] Observe that we could intrepret the acts in different perspectives: acts of "omission" (not complying with some regulation, e.g. fire detectors) vs "positive" acts (driving through red light).

An optimal policy maximizes this expression with respect to the value of the fine and the probability of detection. As is well known, in such a framework the fine should be set at the maximum possible level, say the individuals' wealth or some given upper bound on allowable fines. Accordingly, I take s to be exogenous. Maximizing welfare with respect to the probability of detection and assuming an interior solution, the optimal probability of detection satisfies the first-order condition

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

Thus $ps < h$, implying that the optimal enforcement policy entails underdeterrence compared with first-best behavior. Some individuals, those for whom $ps \leq g < h$, will commit the harmful act even though it is not socially warranted. The optimal policy trades-off some inefficiency in behavior against savings in enforcement expenses.

Now consider fault-based offences, wereby an individual who causes harm is sanctioned only if he failed to obey some standard of behavior. In the present framework, legal standard of behavior is in terms of the circumstances under which the harmful act is committed. If an individual commits a harmful act, his gain must equal or exceed some threshold in order for him to avoid liability; otherwise, he is considered to be at fault. The legal standard is therefore defined by a threshold level of gain $\widehat{g}$. Committing the harmful act is illegal when the circumstances are $g < \widehat{g}$, in which case the violator is subject to a fine if he is detected. An individual will therefore commit the harmful act if $g \geq \min(ps, \widehat{g})$. The optimal policy consists in choosing the probability of detection and the fault standard so as to maximize

$$\int_{\min(ps,\widehat{g})}^{\overline{g}} (g - h)\, f(g)\, dg - c(p).$$

It is easily seen that an optimal policy requires $\widehat{g} \geq ps$, otherwise enforcement costs could be reduced with no detrimental effect on deterrence. The

optimal probability of detection then satisfies the first-order condition (2) and welfare is the same as with a strict liability offence.

In this simple set-up, any standard of fault $\widehat{g} \geq ps$ yields the same outcome. In particular, a standard equal to or greater than the upper bound of possible gains ($\widehat{g} \geq \overline{g}$) is equivalent to a strict liability offence: committing the act is then illegal irrespective of possible circumstances. By contrast, if fines are costly to impose (e.g., there are collection costs) or if sanctions are non-monetary as with prison sentences, one can make the argument that the standard of fault should satisfy $\widehat{g} = ps$. Undeterred individuals should not be found to be at fault even though they behave inefficiently from a social point of view, otherwise unnecessary sanction costs are incurred. When sanctions are costly, the advantage of fault-based liability is to rely on the threat of sanctions while avoiding the cost of actually imposing the sanctions. As we assume costless fines, fault-based liability plays no useful role, at least in the standard model.

Other-regarding preferences. So far we have assumed that behavior depends only on private costs and benefits as conventionally defined. We now consider informal motivations. We assume that there are two types of individuals. A proportion λ of individuals, referred to as type $t = 1$, are "good citizens" with prosocial predispositions. These individuals seek to behave in a socially or morally responsible manner. The other group of individuals, referred to as type $t = 0$, has no such predisposition. Prosocial predispositions are socially valued and individuals who are thought to be good citizens earn status or social esteem, a source of utility.

The utility of a type-t individual is

$$u_t = w - t\gamma \max(e - e^*, 0) + \beta \overline{t}_I, \quad t = 0, 1. \tag{3}$$

The first term, w, is net "material" payoff as in the conventional model. For the good citizens, $t = 1$ and the middle term is the disutility ("guilt") suffered when the individual causes external harm while deviating from what

he knows to be the socially responsible behavior e^* given the circumstances; γ is a positive parameter sufficiently large to intrinsically motivate the good citizen.[3] As defined here, the social (or moral) norm of conduct is what everyone should be doing.[4]

The third term in (3) is the utility from one's social image or reputation. All individuals are assumed to care equally about their social image. β is a positive parameter and $\bar{t}_I \equiv E(t \mid I)$ is the belief of society at large about the individual's type conditional on publicly available information which is denoted by I. The parameter β captures both the importance individuals attach to their social image and the importance ("social pressure") society ascribes to being a good citizen with respect to the kind of acts considered here[5]. Given our definition of types, the conditional expectation $E(t \mid I)$ is simply the posterior probability that the individual is a good citizen. Observe that status depends on beliefs about what the individual *is*. What he *does* matters only in so far as it affects these beliefs. In our analysis, an individual's type is private information and much will depend on what information about him is made available in society at large[6].

Welfare is the sum of utility over all individuals,

$$W = \int_0^{\overline{g}} \left[(1-\lambda)u_0(g) + \lambda u_1(g)\right] \, dg \tag{4}$$

where $u_t(g)$ is the expected utility of a type-t individual in the circumstance g. Before proceeding, we show that socially warranted behavior in the present set-up is the same as in the preceding section.

[3] It will suffice that $\gamma \geq h$, i.e., the good citizen's guilt disutility "internalizes" the harm he causes when $e = 1$ and $e^* = 0$.

[4] The middle term in (3) is a simple version of Kant's categorical imperative, as in Brekke et al. (2003).

[5] beta can be interpreted as the utility of being perceived as good citizen, given that the utility of being perceived as bad citizen is normalized to zero

[6] β and λ depend on social preferences with respect to the kind of situation (and therefore possible acts) considered here. For instance, when e = 1 is a particularly reprehensible action in ordinary circumstances , β will be large and presumably so will be λ

Individuals can both cause harm or suffer harm caused by others. Consider an omniscient regulator who can directly impose the action profile $e(g)$, $g \in [0, \overline{g}]$ on all individuals. The average net material payoff is then

$$w = w_0 + \int_0^{\overline{g}} e(g)\,(g - h)\,f(g)\,dg. \tag{5}$$

where w_0 is initial wealth. Let the action profile $\widehat{e}(g)$ 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 $\widehat{e}(c)$ maximizes W subject to the resource constraint (5) and to beliefs satisfying $\overline{t}_I = \lambda$, where λ is simply the prior belief about types. Clearly, this implies $\widehat{e}(g) = e^*(g)$ as defined in (1). Welfare then equals

$$W^* = w_0 + \int_h^{\overline{g}} (g - h)\,f(g)\,dg + \beta\lambda. \tag{6}$$

Now, the same result would also obtain when full or imperfect information about types is disclosed because the reputational benefits and losses would then simply cancel out.[7]Benabou and Tirole (2006, 2011): esteem is a zero sum game

Public information and branding. Society at large — equivalently an individual's relevant reference group — does not directly observe the circumstances faced by an individual nor his behavior. However, public enforcers are assumed to be able to ascertain the circumstances when they detect a harmful act. In other words, they are able to apply the law when the offence is fault-based. Legal proceedings against an individual constitute public information from which inferences can be made. Specifically, I assume that the only information "publicly" available about an individual — by which I 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

[7] This follows from the law of iterated expectations, $E(E(t \mid I)) = E(t) = \lambda$.

news". The latter means that either the individual did not commit an offence or that he did but was not detected. In terms of the previous notation, the publicly available information affecting one's reputation is therefore the binary signal $I \in \{G, N\}$. The significance of the signal will depend on the legal regime, in particular whether offences are strict or fault-based, and on the enforcement policy. We adopt here an interpretation of self-image that is developed by Bodner-Prelec (2003) and Benabou-Tirole (2006). These authors have incorporated concern for image into models of individual decisions by formalizing them as preference-signaling games and by applying the concept of signaling equilibrium to analyze behavior. These models feature a decision-maker with unobservable preferences over outcomes, who also derives value from the endogenously determined beliefs of an observer about those preferences.

4 Equilibrium under a Given Regime

This section describes 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 legal standard is less the upper bound of possible gains, otherwise the regime involves a strict liability offence. The enforcement policy is characterized 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 the posterior beliefs as given. Next we derive the beliefs as a function of action profiles. Finally we solve for the equilibrium wherein action profiles and beliefs are consistent with one another.

Incentives. Let the sanctioning rule be denoted by $\delta(g, \widehat{g})$, where

$\delta(g,\widehat{g}) = 1$ if $g < \widehat{g}$ and is otherwise zero. The expected utility of a type-t individual in the circumstance g is

$$u_t = w + e\left[g - p\delta(g,\widehat{g})s\right] - t\gamma\max(e - e^*(g), 0) \\ + \beta\left[pe\delta(g,\widehat{g})\bar{t}_G + (1 - pe\delta(g,\widehat{g}))\,\bar{t}_N\right], \quad e \in \{0,1\},\ t \in \{0,1\}.$$

The first term, w, is the part of the individual's wealth that he takes as given. This consist of initial wealth minus the average harm caused by others plus the per capita tax to finance the enforcement policy (surveillance 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 harmful act, $e = 0$, or if he would not be legally at fault when he does, $\delta(g,\widehat{g}) = 0$, 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 harmful 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.

Consider a non prosocial individual. 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 $g \geq \widehat{g}$, 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 $\Delta \equiv \bar{t}_N - \bar{t}_G$ will be referred to as the reputational penalty from being found guilty of an offence. Altogether a non prosocial commits the harmful act if and only if

$$g \geq \min[\widehat{g}, p(s + \beta\Delta)] \equiv g_0, \tag{7}$$

where g_0 is short-hand for the gain threshold of non prosocial individuals.

As the non prosocial, good citizens are motivated by the threat of legal sanctions and by reputational concerns. However, their behavior also reflects an intrinsic concern for complying with the *social* (as opposed to the *legal*) norm conduct. Given γ sufficiently large, a good citizen never commits the harmful act if $g < h$. When $g \geq h$, the harmful act entails no guilt and the good citizen then behaves the same as the non prosocial. The harmful act is therefore committed if an only if

$$g \geq \max(h, g_0) \equiv g_1 \tag{8}$$

where g_1 is the gain threshold for good citizens.

The proportion of type-t individuals who do not commit the harmful act is $F(g_t)$. It will be useful to focus on the ratio $y_t \equiv F(g_t)/F(h)$ of those who do not commit the act over those who *should not* from a social point of view. This ratio will be referred to as the compliance rate, where compliance is meant with respect to the *social* norm of conduct. Socially efficient compliance is $y_t = 1$. As noted in our discussion of the standard model, the legal norm of behavior may differ from the first best. Thus, a fault-based regime may appear to prescribe over- or undercompliance with the social norm of conduct. Let $\widehat{y} = F(\widehat{g})/F(h)$ denote the compliance rate implicitly prescribed by the legal standard of fault. The following is a direct implication of (7) and (8).

Lemma 1 *$y_0 \leq \widehat{y}$ and $y_1 = \max(1, y_0)$.*

The interpretation is that, if there is some overcompliance with respect to the social norm (which requires $\widehat{y} > 1$), then all individuals overcomply equally. Otherwise they either all efficiently comply or the good citizens do while bad citizens undercomply.

Beliefs. The conditions (7) and (8) define the best response functions of type 0 and type 1 individuals given the behavior of other individuals of either

type. The behavior of others affects the payoffs from one's actions thought its effect on the social significance of the "guilty-no news" events, as captured by the beliefs $\bar{t}_G$ and $\bar{t}_N$. Using Bayes' rule, the posterior beliefs conditional on "guilty" or "no news" — and therefore the reputational penalty — can be expressed as a function of the compliance rates. We restrict attention to compliance rates satisfying Lemma 1.

Lemma 2 *If $\widehat{y} \leq 1$, the reputational penalty as a function of y_0 satisfies $\Delta \geq \lambda$ and is decreasing in y_0 down to $\Delta = \lambda$ when $y_0 = \widehat{y}$. If $\widehat{y} > 1$ and $y_0 < y_1 = 1$, $\Delta > 0$ and is decreasing in y_0. If $\widehat{y} > 1$ and $y_0 = y_1 \geq 1$, $\Delta = 0$.*

The intuition is straightforward. Given Lemma 1, unless both types behave the same, bad citizens are more likely to commit the harmful act. Therefore they are more likely to be found guilty of an offence, implying that the event "guilty" is bad news concerning the individual's type compared to "no news".

When the fault-standard satisfies $\widehat{g} \leq h$, good citizens are never found guilty. An offence then reveals perfectly that the individual is non prosocial, so that $\bar{t}_G = 0$ and $\Delta = \bar{t}_N$. The more the non prosocial behave like good citizens, the smaller $\bar{t}_N$. When everyone behaves the same, the event "no news" is totally uninformative because it occurs with certainty. The posterior probability therefore equals the prior λ that an individual is a good citizen.[8] When $\widehat{g} > h$, both good and bad citizens will at times be found guilty, hence $\bar{t}_G > 0$. As long as violating the law is more likely for bad citizens, $\bar{t}_N > \bar{t}_G$ and the reputational penalty is positive. When both types behave the same, the events "guilty" and "no news" are uninformative and posterior beliefs equal the prior in either case.

[8] The event "guilty" is then an out-of-equilibrium event with zero probablity, implying that $\bar{t}_G$ cannot be computed using Bayes' rule. The reputational penalty is obtained from $\lim_{y_0 \uparrow 1} \Delta = \lim_{y_0 \uparrow 1} \bar{t}_N = \lambda$. This can also be rationalized in terms of Cho and Kreps' (1987) D1 criterion.

Figure 1 provides examples of the reputational penalty as a function of the bad citizens' compliance rate under two different legal regimes, given $y_1 = 1$. The enforcement policy is the same under both regimes. In case A, the standard of fault is the first-best $\widehat{g}_A = h$, equivalently $\widehat{y}_A = 1$. The reputational penalty is then bounded below by λ. In case B, the standard of fault is above the first best, $\widehat{y}_B > 1$. The reputational penalty then goes down to zero when all the non prosocial comply perfectly. For compliance rates sufficiently close to unity, the reputational penalty under regime A is therefore larger than under B. As depicted, the curves intersect. This need not occur but it is a possibility at sufficiently small compliance rates. I will discuss this further when we turn to legal design.

`Figure 1 about here`

Equilibrium. From the foregoing discussion, both types behave the same if they overcomply, so that the reputational penalty is then nil. Because reputational concerns then provide no incentives, overcompliance can arise only if the expected fine $ps > h$ as in the standard model with no informal motivations. We disregard policies with $ps > h$ because they would serve no purpose. In the cases considered, good citizens therefore always perfectly comply with the social norm of conduct. An equilibrium consists of a compliance rate for the non prosocial and of a reputational penalty that are mutually consistent, given that good citizens perfectly comply.

Proposition 1 *Let the enforcement policy and legal regime satisfy $ps \leq h$ and $\widehat{g} \geq ps$. Then there is a unique equilibrium with $y_0 \leq y_1 = 1$.*
(i) If $ps = h$, $y_0 = 1$ as well.
(ii) If $ps < \widehat{g} \leq h$, the equilibrium y_0 is increasing in p as long as $p(s+\beta\lambda) < \widehat{g}$, otherwise $y_0 = F(\widehat{g})/F(h)$; in either case, y_0 is increasing in $\widehat{g}$.
(iii)If $ps < h < \widehat{g}$, the equilibrium y_0 is increasing in p and may be increasing or decreasing in $\widehat{g}$.

Different equilibria are illustrated for the case $ps < h$. In the Figures 2 to 5, $y_S = F(ps)/F(h)$ denotes the compliance rate that would obtain in the standard model; $\Delta(y_0)$ is the reputational penalty as a function of the compliance rate under a given legal regime and enforcement policy; $y_0(\Delta)$ is the compliance rate as a function of the reputational policy under the same regime and enforcement policy. The perfect Bayesian equilibrium is the intersection of theses curves (point E).

Consider first the case $\widehat{y} < 1$. An individual found guilty of an offence is then for sure non prosocial. In Figure 2, the compliance rate of the non prosocial is increasing in the reputational penalty up to the upper bound $\widehat{y}$ entailed by the fault standard. Figure 3 depicts the case where $\widehat{y}$ does not bind. In either case, relaxing the fault standard (that is, increasing $\widehat{g}$ or equivalently $\widehat{y}$) yields an increase in the equilibrium compliance rate. The effect is obvious if the fault standard binds. When it does not, the effect follows from the fact that relaxing the standard shifts the reputational penalty curve to the right (the curve also rotates upwards, while remaining bounded below by λ when $y_0 = \widehat{y}$). The intuition is that relaxing the fault standard increases the significance of the "no news" event, so that reputational incentives have more bite. In Figure 4, the standard of fault is the first-best $\widehat{y} = 1$. In the case represented, all individuals comply perfectly. This arises if reputational concerns are sufficiently important (β is large) or if the expected fine is sufficiently large even though $ps < h$.

`Figure 2 to 5 about here`

In Figure 5, $\widehat{y} > 1$ and both types can be found guilty. Further relaxing the fault standard then has an ambiguous effect on the reputational penalty curve which may rotate upwards or downwards, so the equilibrium compliance rate may go either way. As before, relaxing the standard increases the significance of "no news". However, it also reduces the significance of an

offence because more good citizens are found guilty, so the net effect on the reputational penalty is ambiguous.

Increasing the probability of detecting harmful acts increases the significance of "no news", with no effect on the significance of the "guilty" event. In the figures, the reputational penalty curve therefore rotates upwards. Because offenders are now more likely to be apprehended, a larger probability of detection also shifts the positively sloped portion of the compliance curve to the right (and reduces the slope). Thus, compliance unambiguously increases except at corner solutions where $y_0 = 1$.[9]

5 Optimal Legal Regime and Enforcement

When the enforcement policy satisfies $ps \leq h$, welfare reduces to the first best W^* as defined in (6) minus the loss from undercompliance on the part of the non prosocial and the per capita enforcement expenditure:

$$W = W^* - (1-\lambda) \int_{g_0(\widehat{y},p)}^{h} (h-g)\, f(g)\, dg - c(p) \tag{9}$$

where $g_0(\widehat{y}, p)$ is the equilibrium threshold for the non prosocial, given the legal regime and enforcement policy. An optimal policy sets $\widehat{g}$ and p so as to maximize the above expression. This is equivalent to maximizing

$$V(\widehat{y}, p) \equiv (1-\lambda) \int_{g_0(\widehat{y},p)}^{h} (g-h)\, f(g)\, dg - c(p). \tag{10}$$

The structure of the problem is the same as in the standard model except that one now takes into account that different policies may be more or less efficient in harnessing reputational motivations.[10]

[9] When $\widehat{y} > 1$, the effect on the equilibrium reputational penalty is ambiguous. A negative effect may be interpreted as greater formal legal enforcement partially crowding out informal motivations.

[10] The maximum sanction principle still holds for the usual reasons. Thus we take s as given.

Efficient legal regimes. We first show that, by contrast with the indetermination in the standard model, there are now only two possibly optimal legal regimes.

Proposition 2 *The optimal legal regime is either a strict liability offence or a fault-based offence with the first-best standard of fault $\widehat{g} = h$. In either case, the optimal enforcement satisfies $p(s + \beta\lambda) < h$ and yields undercompliance on the part of the non prosocial.*

The reason for the underdeterrence result is the same as in the standard model. Starting from first-best compliance, a slight decrease in the probability of detection yields at most a second order welfare loss in terms of reduced compliance but a first-order decrease in enforcement expenditures. However, the inequality $ps < h$ does not necessarily entail undercompliance because individuals are now also motivated by reputational concerns. Thus, underterrence requires a sufficiently small expected fine, as shown in the proposition.

The first part of the proposition follows from the fact that compliance increases with the fault-standard up to $\widehat{g} = h$. If compliance can be increased further still, then the proposition states that the optimal standard is at the upper bound $\widehat{g} = \overline{g}$, which amounts to strict liability.[11] Compared with the conventional model, strict liability and fault-based offences are therefore not equivalent. Moreover, if the offence is to be fault-based, then the standard of fault must be the first-best threshold.

Choosing between regimes. We illustrate why one legal regime may perform better than the other. Figure 6 reproduces the reputational penalty curves of Figure 1. The subscript A refers to the fault-based regime with

[11] A strict liability offence disregard circumstances. Good citizens will then sometimes efficiently choose not to comply with the law given their knowledge of circumstances. This is reminiscent of Shavell (2012).

the standard $\widehat{y} = 1$, the subscript B to a strict liability regime. The enforcement policy is the same in both cases (and is therefore not necessarily optimal). I compare two situations, L and H, which differ in the intensity of reputational concerns with $\beta_L < \beta_H$. In situation L the compliance curve is less sensitive to reputational penalties. For the given enforcement policy, the highest feasible compliance rate is relatively small and the best regime is strict liability (the equilibrium at E_B^L). For the same enforcement policy, situation H yields the opposite: individuals are very sensitive to reputational penalties, the highest feasible compliance rate is relatively large and the best regime is fault-based (the equilibrium at E_A^H). Now, suppose the enforcement policy represented in the figure is in fact optimal for the strict liability regime. Because the marginal loss from undercompliance is smaller at E_A^H than at E_B^L, it would be welfare improving to somewhat reduce enforcement in situation H. Thus, in situation H, the optimal regime would be fault-based and the optimal enforcement policy would involve lower enforcement expenditures than in L.

Figure 6 about here

The argument illustrated in Figure 6 presumes that the reputational penalty curves intersect. As remarked in Section 3, this need not occur. We now discuss the reason why it can. Obviously, learning that an individual has been found guilty of an offence is more revealing about his type in a fault-based than in a strict liability regime. However, the information available to the general public under a given regime is the binary signal defined by the "guilty-no news" events. While the event "guilty" is *more unfavorable* news about an individual's type under the fault-based regime, it turns out that "no news" does not necessarily constitute *more favorable* news. In other words, from an informational point of view the signals cannot always be ranked. Hence the possibility that the curves intersect. More generally,

different fault standards determine different signals which cannot be ranked in the sense of one binary signal being more informative than the other.[12]

If the reputational sanction curves intersect, one can show that they do so at[13]

$$y_0 = \frac{1}{2(1-\lambda)}\left(1 - 2\lambda - \frac{1-p}{F(h)p}\right). \tag{11}$$

The right-hand side of (11) is positive only if $\lambda < 1/2$ and

$$p > \frac{1}{1 + (1-2\lambda)F(h)}.$$

For instance, if $\lambda = 1/4$, $F(h) = 2/3$ and $p = 9/10$, then the intersection is at $y_0 = 4/9$. When under strict liability the equilibrium compliance rate is less than the value at which the curves cross, as in situation L of Figure 6, one can check that the per capita frequency of detected offences will be greater than one half. As shown below, this turns out to be a necessary condition for a strict liability offence to be optimal.

The optimal legal regime and enforcement policy depend on the underlying situation and both must be jointly chosen. The underlying situation includes the importance of reputational concerns, the proportion of prosocial individuals, the severity of the harmful act, and the probability distribution of possible circumstances. Moreover, whether a fault-based or a strict liability regime performs better also depends on enforcement possibilities, as defined by the permissible fine and the enforcement cost function.

Proposition 3 *Suppose the legal regime and enforcement policy are optimal. If the legal regime is fault-based, then detected offences constitute a rare event,*

$$pF(h)(1-\lambda)(1-y_0) < \frac{1}{2}. \tag{12}$$

[12] There is an exception: for $\widehat{g} < h$, relaxing the fault-standard increases the informativeness of the binary "guilty-no news" signal, which explains why $\widehat{g} < h$ cannot be efficient.

[13] Solve (17) and (18) in the Appendix for y_0 yielding the same Δ under both legal regimes.

If the legal regime involves a strict liability offence, then detected offences constitute a frequent event,

$$p\left[1 - F(h) + F(h)(1-\lambda)(1-y_0)\right] \geq \frac{1}{2}. \tag{13}$$

The left-hand side of (12) is the frequency of detected violations under the fault-based regime with standard $\widehat{g} = h$ and enforcement policy p, given the equilibrium compliance rate on the part of the non prosocial. The left-hand side of (13) is the frequency of detected violations under a strict liability regime given the equilibrium compliance rate under that regime. Everyone then commits the harmful act when $g \geq h$; when $g < h$, a fraction $1 - y_0$ of the non prosocial do.

The condition (12) is more likely to hold when the permissible fine is large, as this allows deterrence with a relatively small probability of detection. It is also more likely if detecting violations is very costly so that the probability of detection is small. Thus, the choice between fault-based and strict liability offences will depend on enforcement considerations. In particular, the following condition is sufficient.

Corollary 1 *The optimal legal regime is fault-based if socially unwarranted acts would constitute a rare event in the absence of legal sanctions,*

$$F(h)(1-\lambda) < \frac{1}{2}. \tag{14}$$

The condition is satisfied if the harmful act is usually socially warranted (i.e., $1 - F(h)$ is greater than one half) or if good citizens constitute a majority.

6 Discussion and Extensions

In many situations, socially unwarranted behavior will be a rare event because most individuals are socially minded. Hence illegal behavior will also

be rare. It may also be that *detected* illegal behavior is rare because substantial deterrence is achieved with a large fine and a low probability of apprehension. A legal regime that seeks to harness reputational incentives should then seek to reduce apparent unlawfulness. This is achieved by a fault-based regime. Not finding fault may then be banal, therefore posteriors conditional on "no news" do not differ too much from the prior. But then to be found guilty of an offence yields substantial disesteem. By contrast, when detected illegal behavior would be a frequent event under a fault-based regime, offences are banal and not finding fault may yield significant esteem. It will then be better to switch to a strict liability regime, as this increases the salience or visibility of offences, thereby increasing the significance of "no news".

Our results are 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 being guilty of an offence imposes significant stigma or whether not having been found guilty confers significant honor depends on the underlying situation but also on the legal regime itself together with enforcement possibilities.

7 Concluding Remarks

Violating the law does not have the same social meaning under strict liability and fault-based offences. A fault-based offence is a stronger signal about one's character than a strict liability offence. Fault-based regimes will therefore often perform better in harnessing reputational concerns for the purpose

of inducing socially appropriate behavior. However, the result does not always follow because the social meaning of legal sanctions depends on the frequency of detected offences. This in turn will depend on the underlying situation, the enforcement policy and the legal regime.

We emphasized the information conveyed by offences under different legal regimes. 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-based regime, the social norm can be perfectly "expressed" by the duty or obligation with respect to which fault is defined. Strict liability is fuzzier in this respect. Strict liability and fault-based offences may also differ in other ways with respect to expressive content. When individuals are imperfectly informed of the harm they may cause, the legal standard of behavior conveys information, as in D'Antoni and Galbiati (2007). The prescriptive content of fault may help socially minded individuals to coordinate on the socially appropriate conduct (see Cooter 1998). Imitative behavior due to reputational concerns then induces some bunching by the non prosocial on the socially appropriate behavior.

Appendix

Proof of Lemma 1. From (7), $y_0 = F(g_0)/F(h) \leq F(\widehat{g})/F(h) \equiv \widehat{y}$. From (8),

$$y_1 = F(g_1)/F(h) = \max[F(h), F(g_0)]/F(h) = \max(1, y_0). \blacksquare$$

Proof of Lemma 2. Let $y_0 \leq \widehat{y}$ and $y_1 = \max(1, y_0)$ as in Lemma 1. Applying Bayes' rule,

$$\bar{t}_N \equiv E(t \mid N) = \frac{\lambda \left[1 - pF(h) \max(\widehat{y} - y_1, 0)\right]}{1 - pF(h) \left[\lambda \max(\widehat{y} - y_1, 0) + (1 - \lambda)(\widehat{y} - y_0)\right]}, \qquad (15)$$

$$\bar{t}_G \equiv E(t \mid G) = \frac{\lambda \max(\widehat{y} - y_1, 0)}{\lambda \max(\widehat{y} - y_1, 0) + (1-\lambda)(\widehat{y} - y_0)}, \tag{16}$$

where (16) is undefined when $y_0 = y_1 = \widehat{y}$.

If $y_0 < \widehat{y} \leq 1$, $y_1 = 1$ implies $\bar{t}_G = 0$ and therefore

$$\Delta \equiv \bar{t}_N - \bar{t}_G = \frac{\lambda}{1 - pF(h)(1-\lambda)(\widehat{y} - y_0)}. \tag{17}$$

This is decreasing in y_0. For $y_0 = \widehat{y}$, I take the lower bound $\Delta = \lambda$. If $\widehat{y} > 1$ and $y_0 < y_1 = 1$,

$$\begin{aligned} \Delta &= \frac{\lambda \left[1 - pF(h)(\widehat{y} - 1)\right]}{1 - pF(h)\left[\lambda(\widehat{y} - 1) + (1-\lambda)(\widehat{y} - y_0)\right]} \\ &\quad - \frac{\lambda(\widehat{y} - 1)}{\lambda(\widehat{y} - 1) + (1-\lambda)(\widehat{y} - y_0)}, \end{aligned} \tag{18}$$

which is positive and decreasing in y_0. If $\widehat{y} > 1$ and $y_0 = y_1 < \widehat{y}$, (15) and (16) yield $\bar{t}_N = \bar{t}_G$ so that $\Delta = 0$. ■

Proof of Proposition 1. I first show that $ps \leq h$ implies $y_0 \leq y_1 = 1$. By Lemma 1, if the foregoing does not hold, $y_0 = y_1 > 1$. This requires $\widehat{y} > 1$ (equivalently $\widehat{g} > h$) and Lemma 2 then implies $\Delta = 0$. Thus, $y_0 = y_1 > 1$ is possible only if $ps > h$.

By Lemma 2, $\Delta \geq 0$. For $\widehat{g} \geq ps$, (7) then implies $y_0 \geq y_S \equiv F(ps)/F(h)$. Given $ps \leq h$ and $\widehat{g} \geq ps$, the relevant domain for y_0 is therefore the interval $[y_S, \min(\widehat{y}, 1)]$. The equilibrium y_0 in this interval is a solution to

$$y_0 = \min\left[\widehat{y}, \frac{F(p(s + \beta\Delta(y_0, \widehat{y}, p)))}{F(h)}\right], \tag{19}$$

where $\Delta(y_0, \widehat{y}, p)$ denotes the reputational penalty satisfying (17) or (18) for the cases $\widehat{y} \leq 1$ and $\widehat{y} > 1$ respectively. Equivalently, the equilibrium y_0 is a solution to

$$\varphi(y_0) \equiv \min\left[\widehat{y}, \frac{F(p(s + \beta\Delta(y_0, \widehat{y}, p)))}{F(h)}\right] - y_0 = 0. \tag{20}$$

$\varphi(y_0)$ is a continuous function. Because $\Delta(y_0, \widehat{y}, p)$ is strictly decreasing in y_0 in the relevant domain, $\varphi(y_0)$ is also strictly decreasing in the same domain. This ensures uniqueness of the equilibrium.

(i) Let $ps = h$. If $\widehat{y} = 1$, obviously $y_0 = 1$. For $\widehat{y} > 1$, Lemma 2 implies $\Delta(y_0, \widehat{y}, p) > 0$ for all $y_0 < 1$ and $\Delta(1, \widehat{y}, p) = 0$, hence $y_0 = 1$ is again the unique solution to (20).

(ii) Let $ps < \widehat{g} \leq h$, equivalently $y_S < \widehat{y} \leq 1$. I show that the equilibrium $y_0 \in (y_S, \widehat{y}]$. Obviously $\varphi(y_S) > 0$. By Lemma 2, $\Delta(\widehat{y}, \widehat{y}, p) = \lambda$. If $p(s+\beta\lambda) \geq \widehat{g}$, $\varphi(\widehat{y}) = 0$ and the equilibrium satisfies $y_0 = \widehat{y}$. If $p(s+\beta\lambda) < \widehat{g}$, $\varphi(\widehat{y}) < 0$ and the equilibrium satisfies $y_0 < \widehat{y}$. In the latter case, differentiating (19) totally with respect to $\widehat{y}$ and p yields

$$\frac{\partial y_0}{\partial \widehat{y}} = \frac{p\beta f(g_0)\Delta_{\widehat{y}}}{F(h) - p\beta f(g_0)\Delta_{y_0}}, \tag{21}$$

$$\frac{dy_0}{dp} = \frac{f(g_0)(s + \beta\Delta + p\beta\Delta_p)}{F(h) - p\beta f(g_0)\Delta_{y_0}}, \tag{22}$$

where $g_0 = p(s+\beta\Delta(y_0, \widehat{y}, p))$. The reputational penalty is decreasing in y_0, hence the denominator is positive. From (17), $\Delta(y_0, \widehat{y}, p)$ is increasing in $\widehat{y}$ and in p. Hence (21) and (22) are both positive. To complete the argument, when $p(s + \beta\lambda) \geq \widehat{g}$, $y_0 = \widehat{y}$ and is then also increasing in $\widehat{y}$.

(iii) Let $ps < h < \widehat{g}$, equivalently $y_S < 1 < \widehat{y}$. The argument is similar except that the solution now satisfies $y_0 \in (y_S, 1)$. As before, $\varphi(y_S) > 0$. By Lemma 2, $\Delta(1, \widehat{y}, p) = 0$ and therefore $\varphi(1) < 0$. Differentiating (19) totally with respect to $\widehat{y}$ and p again yields (21) and (22). However, the reputational penalty is now defined by (18). The signs of $\partial\Delta/\partial y_0$ and $\partial\Delta/\partial p$ are positive but that of $\partial\Delta/\partial\widehat{y}$ is now ambiguous (see the proof of Proposition 2). Thus y_0 is increasing in p but may now be increasing or decreasing in $\widehat{y}$. ■

Proof of Proposition 2. Let $y_0(\widehat{y}, p)$ and $g_0(\widehat{y}, p)$ denote equilibrium values as derived in Proposition 1, $y_0(\widehat{y}, p) \equiv F(g_0(\widehat{y}, p))/F(h)$. An optimal legal regime and enforcement policy maximizes

$$V(\widehat{y}, p) := (1 - \lambda) \int_{g_0(\widehat{y},p)}^{h} (g - h)\, f(g)\, dg - c(p). \tag{23}$$

The partial derivatives are

$$V_{\widehat{y}}(\widehat{y}, p) = (1 - \lambda) \left(\frac{h - g_0(\widehat{y}, p)}{F(h)} \right) \frac{\partial y_0(\widehat{y}, p)}{\partial \widehat{y}},$$

$$V_p(\widehat{y},p) = (1-\lambda)\left(\frac{h-g_0(\widehat{y},p)}{F(h)}\right)\frac{\partial y_0(\widehat{y},p)}{\partial p} - c'(p).$$

From the proof of Proposition 1, the function $y_0(\widehat{y},p)$ may be discontinuous at $\widehat{y}=1$; even when the function is continuous, the derivative $\partial y_0(\widehat{y},p)/\partial\widehat{y}$ is discontinuous at $\widehat{y}=1$. Moreover, for $\widehat{y}<1$, $\partial y_0(\widehat{y},p)/\partial p$ is discontinuous when $p(s+\beta\lambda)=\widehat{g}$. As need be, derivatives will be understood to mean left or right derivatives.

Let $(\widehat{y}^*,p^*)$ denote an optimal policy, assuming $p^*>0$. I consider the possibilities that $\widehat{y}^*\leq 1$ or $\widehat{y}^*>1$. Let $\overline{y}=F(\overline{g})/F(h)$, so that $\widehat{y}=\overline{y}$ amounts to strict liability.

Case 1: $\widehat{y}^*\leq 1$.

For $\widehat{y}\leq 1$, the function $y_0(\widehat{y},p)$ satisfies part (ii) of Proposition 1. When $\widehat{y}<1$, $g_0(\widehat{y},p)<h$ and $\partial y_0(\widehat{y},p)/\partial\widehat{y}>0$, hence $V_{\widehat{y}}(\widehat{y},p)>0$. If $\widehat{y}^*\leq 1$, it must therefore be that $\widehat{y}^*=1$, equivalently $\widehat{g}^*=h$. When $p(s+\beta\lambda)\geq h$, $g_0(1,p)=h$ so that $V_p(1,p)<0$. The optimal p^* must therefore satisfy $p^*(s+\beta\lambda)<h$, implying underterrence, i.e., $y_0(1,p^*)<1$.

Case 2: $\widehat{y}^*>1$.

For $\widehat{y}>1$, the function $y_0(\widehat{y},p)$ satisfies part (iii) of Proposition 1, implying $g_0(\widehat{y},p)<h$ whenever $ps<h$. I first show that the optimal p^* must satisfy $p^*(s+\beta\lambda)<h$ as in Case 1. Suppose not. Then, for all $\widehat{y}>1$,

$$V(\widehat{y},p^*) = (1-\lambda)\int_{g_0(\widehat{y},p^*)}^{h}(g-h)\,f(g)\,dg - c(p^*) < -c(p^*) = V(1,p^*)$$

where the right-hand side is the welfare level reached by setting the standard $\widehat{y}=1$ instead and inducing the first-best behavior because $p^*(s+\beta\lambda)\geq h$.

When $p^*(s+\beta\lambda)<h$, $y_0(\widehat{y},p^*)<1$ in the closed interval $[1,\overline{y}]$ and the function is then continuous on this interval. The latter follows from the fact that, when $y_0<1$, the reputational penalty is continuous in $\widehat{y}$ at $\widehat{y}=1$. It follows that

$$\lim_{\widehat{y}\downarrow 1}\Delta(y_0(\widehat{y},p^*),\widehat{y},p^*) = \frac{\lambda}{1-p^*F(h)(1-\lambda)(1-y_0(1,p^*))},$$

where the right-hand side is the penalty as defined in (17). Thus we may look for an optimal policy $\widehat{y}^* \in [1, \overline{y}]$. If $\widehat{y}^* \neq 1$, then either $\widehat{y}^* = \overline{y}$ or $\widehat{y}^*$ is an interior solution.

In the latter case, the solution must satisfy the first-order condition $V_{\widehat{y}}(\widehat{y}^*, p^*) = 0$, implying

$$\left.\frac{\partial y_0(\widehat{y}, p^*)}{\partial \widehat{y}}\right|_{\widehat{y}=\widehat{y}^*} = 0, \tag{24}$$

and the second-order condition $V_{\widehat{y}\widehat{y}}(\widehat{y}^*, p^*) \leq 0$ which, when (24) holds, implies

$$\left.\frac{\partial^2 y_0(\widehat{y}, p^*)}{\partial \widehat{y}^2}\right|_{\widehat{y}=\widehat{y}^*} \leq 0. \tag{25}$$

From (21), the condition (24) requires

$$\Delta_{\widehat{y}}(y_0^*, \widehat{y}^*, p^*) = 0, \tag{26}$$

where $y_0^* = y_0(\widehat{y}^*, p^*)$. Differentiating (21) with respect to $\widehat{y}$, given (24) and (26), yields

$$\left.\frac{\partial^2 y_0(\widehat{y}, p^*)}{\partial \widehat{y}^2}\right|_{\widehat{y}=\widehat{y}^*} = \frac{p\beta f(g_0^*)\Delta_{\widehat{y}\widehat{y}}(y_0^*, \widehat{y}^*, p^*)}{F(h) - p\beta f(g_0^*)\Delta_{y_0}(y_0^*, \widehat{y}^*, p^*)}.$$

Therefore (25) requires

$$\Delta_{\widehat{y}\widehat{y}}(y_0^*, \widehat{y}^*, p^*) \leq 0. \tag{27}$$

The reputational penalty as defined in (18) can be rewritten as

$$\Delta(y_0, \widehat{y}, p) = \frac{\lambda(1-\lambda)(1-y_0)}{(\widehat{y} - \lambda - (1-\lambda)y_0)\left[1 - pF(h)(\widehat{y} - \lambda - (1-\lambda)y_0)\right]}$$

so that

$$\Delta_{\widehat{y}}(y_0, \widehat{y}, p) = -\frac{\lambda(1-\lambda)(1-y_0)\left[1 - 2pF(h)(\widehat{y} - \lambda - (1-\lambda)y_0)\right]}{\left\{(\widehat{y} - \lambda - (1-\lambda)y_0)\left[1 - pF(h)(\widehat{y} - \lambda - (1-\lambda)y_0)\right]\right\}^2}. \tag{28}$$

Given (26), it is then easily seen that

$$\Delta_{\widehat{y}\widehat{y}}(y_0^*, \widehat{y}^*, p^*) = \frac{\lambda(1-\lambda)(1-y_0^*)2pF(h)}{\left\{(\widehat{y} - \lambda - (1-\lambda)y_0^*)\left[1 - pF(h)(\widehat{y} - \lambda - (1-\lambda)y_0^*)\right]\right\}^2},$$

Thus, the necessary condition (27) does not hold. ■

Proof of Proposition 3. Suppose p^* and $\widehat{g}^*$ (equivalently $\widehat{y}^*$) are optimal and yield y_0^*. By Proposition 2, either $\widehat{y}^* = 1$ or $\widehat{y}^* = \overline{y}$.

If $\widehat{y}^* = 1$, we must have

$$\left.\frac{\partial y_0(\widehat{y}, p^*)}{\partial \widehat{y}}\right|_{\widehat{y}=1} \leq 0,$$

where the expression denotes the right derivative. Using the same argument as in the proof of Proposition 2, this implies $\Delta_{\widehat{y}}(y_0^*, 1, p^*) \leq 0$,again a right derivative, so that $\Delta_{\widehat{y}}$ is given by (28). Therefore

$$1 - 2pF(h)(1-\lambda)(1-y_0^*) \geq 0.$$

Now, the inequality must be strict. Otherwise, as shown in the proof of Proposition 2, we would have $\Delta_{\widehat{y}\widehat{y}}(y_0^*, 1, p^*) > 0$ and therefore

$$\left.\frac{\partial^2 y_0(\widehat{y}, p^*)}{\partial \widehat{y}^2}\right|_{\widehat{y}=1} > 0,$$

implying that $y_0(\widehat{y}, p^*)$ would be increasing in a neighborhood of $\widehat{y} = 1$.

Similarly, if $\widehat{y}^* = \overline{y}$, it must be the case that $\Delta_{\widehat{y}}(y_0^*, \overline{y}, p^*) \geq 0$, implying

$$1 - 2pF(h)(\overline{y} - \lambda - (1-\lambda)y_0^*) \leq 0.$$

The latter is equivalent to condition (13) in the proposition given that $\overline{y} \equiv F(\overline{g})/F(h) = 1/F(h)$. ■

References

[1] Andreoni, J. and B.D. Bernheim (2009). "Social Image and the 50-50 Norm." *Econometrica* 77, 1607-1636.

[2] Ariely, D., A. Bracha and S. Meier (2009). "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99, 544-555.

[3] Bénabou, R. and J. Tirole (2006). "Incentives and Prosocial Behavior." *American Economic Review* 96, 1652-1678.

[4] Bénabou, R. and J. Tirole (2011). "Laws and Norms." NBER wp 17579.

[5] Bernheim, B.D. (1994), "A Theory of Conformity." *Journal of Political Economy* 102, 905-953.

[6] Bodner, R. and D. Prelec (2002). "Self-Signaling in a neo-Calvinist Model of Everyday Decision-Making." In I. Broca and J. Carillo (eds), *Psychology and Economics*, Vol. II, Oxford University Press.

[7] Brekke, K.A., Kverndokk, S. and K. Nyborg (2003), "An Economic Model of Moral Motivation." *Journal of Public Economics* 87, 1967-1983.

[8] Brown, D. K. (2012). "Criminal Law Reform and the Persistence of Strict Liability." *Duke Law Journal* 62, 285-338.

[9] Cho, I. K. and D. Kreps (1987), "Signaling games and stable equilibria." *Quarterly Journal of Economics* 102, 179-221.

[10] Cooter, R. (1998a), "Expressive Law and Economics." *Journal of Legal Studies* 27, 585-608.

[11] Cooter, R. (1998b), "Models of Morality in Law and Economics: Self-Control and Self-Improvement for the Bad Man of Holmes", Boston University Law Review, 78.

[12] Cooter, R. (2000a), "Do Good Laws Make Good Citizens? An Economic Analysis", *Virginia Law Review,* 86, 1577–1601.

[13] Cooter, R. (2000b), "Three Effects of Social Norms on Law: Expression, Deterrence and Internalization", *Oregon Law Review,* 79, 1–22.

[14] Cooter, R. and A. Porat (2001), Should Courts Deduct Nonlegal Sanctions from Damages? Journal of Legal Studies 30, 401–22.

[15] Dal Bó, E. and M. Terviö (2013), "Self-Esteem, Moral Capital and Wrong-Doing." *Journal of the European Economic Association* 11, 599-633.

[16] Dana, J., D.M. Cain and R. Dawes (2006), "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes* 100, 193-201.

[17] D'Antoni, M. and R. Galbiati (2007), "A Signalling Theory of Non-monetary Sanctions." *International Review of Law and Economics* 27, 204-218

[18] Daughety, A. and J. Reinganum (2010), "Public Goods, Social Pressure, and the Choice Between Privacy and Publicity." *American Economic Journal: Microeconomics* 2, 191-222.

[19] Deffains, B. and C. Fluet (2013), "Legal Liability when Individuals Have Moral Concerns." *Journal of Law, Economics, and Organization*, 29, 930-955.

[20] Duff, A. (2007), *Answering for Crime, Responsibility and Liability in the Criminal Law (Legal Theory Today)*, Hart Publishing.

[21] Duber, M. and T. Hörle (2014), *Criminal Law: A Comparative Approach*, Oxford University Press.

[22] Elllingsen, T. and M. Johannesson (2008), "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98, 990-1008.

[23] Faure, M. and G. Heine (2005), *Criminal Enforcement of Environmental Law in the European Union*, Kluwer International.

[24] Fitzgerald, P.J. (1965), "Real Crimes and Quasi-Crimes". *Natural Law Forum* 10, 21-53.

[25] Frey, B. and R. Jegen (2001), "Motivation Crowding Out Theory." *Journal of Economic Surveys* 15, 589-611.

[26] Funk, P. (2010), "Social Incentives and Voter Turnout: Theory and Evidence." *Journal of the European Economic Association* 8, 1077-1103.

[27] Harel, A. and A. Klement (2007), "The Economics of Stigma: Why More Detection of Crime May Result in Less Stigmatization." *Journal of Legal Studies* 36, 355-378.

[28] Hart, H.L.A. (1994), The Concept of Law, 2nd edn., Oxford: Clarendon Press.

[29] Honoré, T. (1993), "The Dependence of Morality on Law", Oxford Journal of Legal Studies, 13 (1), pp. 1-17.

[30] Horder, J. (2005), Whose Values Should Determine When Liability is Strict? in *Apraising Strict Liability*, A. Simester (ed), Oxford University Press.

[31] Iacobucci, E.M. (2014). "On the Interaction Between Legal and Reputational Sanctions." *Journal of Legal Studies*, 43, 2014.

[32] Kadish, S.H. (1963). "Some Observations on the Use of Criminal Sanctions in Enforcing Economic Regulations." *University of Chicago Law Review* 30, 423-442.

[33] Kahan, D.M. (1998), "Social Meaning and the Economic Analysis of Crime." *Journal of Legal Studies* 27, 709-622.

[34] Lacetera, N. and M. Macis (2010), "Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme." *Journal of Law, Economics, and Organization* 76, 225-237.

[35] Law Commission (2010), *Criminal Liability in Regulatory Contexts*, Consultation Paper No 195, London.

[36] Law Reform Commission of Canada (1974), *Studies on Strict Liability*, Information Canada, Ottawa.

[37] Masclet, D., C. Noussair, S. Tucker and M.C. Villeval (2003), "Monetary and Non-Monetary Punishment in the Voluntary Contribution Mechanism", *American Economic Review*, 93(1), pp 366-380.

[38] McAdams, R.H. and E. B. Rasmusen (2007), "Norms in Law and Economics." In Polinsky, A. M. and S. Shavell (eds.), *Handbook of Law and Economics*, Vol. 1, New York: North-Holland.

[39] Mialom, S. (2014), Declining Moral Standards and The Role of Law, Working Paper.

[40] Paulus, I. (1977). "Strict Liability: Its Place in Public Welfare Offences." *Criminal Law Quarterly* 20, 445-467.

[41] Polinsky, M.A. and S. Shavell (2007), "The theory of public enforcement of law." In Polinsky, A. M. and S. Shavell (eds.), *Handbook of Law and Economics*, Vol. 1, New York: North-Holland.

[42] Posner, R. (1997), "Social Norms and the Law: An Economic Approach", American Economic Review, 87, pp. 365-369.

[43] Posner, E. (1998), "Symbols, Signals, and Social Norms in Politics and the Law." *Journal of Legal Studies* 27, 765-798.

[44] Posner, E. (2000), *Law and Social Norms*. Cambridge, MA: Harvard University Press.

[45] Prelec, D. and R. Bodner (2003). "Self-Signaling and Self-Control." In G. Loewenstein, D. Read, and R. Baumeister, *Time and Decisions*, Russell Sage Foundation.

[46] Rasmusen, E. (1996), "Stigma and Self-Fulfilling Expectations of Criminality." *Journal of Law and Economics* 39, 519-544.

[47] Raz, J. (1999), Practical Reason and Norms, Oxford University Press.

[48] Schaefer, H. B.and Mueller-Langer (2008), Strict Liability Versus Negligence (2008). Available at SSRN: http://ssrn.com/abstract-2062787 or http://dx.doi.org/10.2139/ssrn.2062787

[49] Shavell, S. (2002), "Law versus Morality as Regulators of Conduct." *American Law and Economics Review* 4, 227-257.

[50] Shavell, S. (2012), "When is Complying with the Law Socially Desirable?" *Journal of Legal Studies* 41, 1-36.378-405.

[51] Simester A. (2005), Is Strict Liability always wrong? in *Apraising Strict Liability*, A. Simester (ed), Oxford University Press.

[52] Singer, R. (1989). "The Resurgence of Mens Rea: The Rise and Fall of Strict Criminal Liability." *Boston College Law Review* 30, 337-408.

[53] Smith, A. (1759), *The Theory of Moral Sentiments*. Reedited (1997), Washington, D.C., Regnery Publishing.

[54] Spencer, J.R. and A. Pedain (2005), "Approaches to Strict and Constructive Liability in Continental Criminal Law." In: Simester, A. P. (ed), *Appraising Strict Liability*, Oxford Monographs on Criminal Law.

[55] Teichman (2005), "Sex, Shame, and the Law: An Economic Perspective on Megan's Laws", *Harvard Journal on Legislation*, 42, 355–415.

[56] Thomson, N. (1994), "Fundamental Justice, Stigmas and Fault", University of Toronto Faculty Law Review, 52, 379-404.

[57] Tyran, J. and L. Feld (2006), "Achieving Compliance When Legal Sanctions are Non-Deterrent." *Scandinavian Journal of Economics* 108, 135-156.

[58] Wils, W. (2006). "Is Criminalization of EU Competition Law the Answer?" In: Cseres, K.J., Schinkel, M.-P. and F. Vogelaar (eds), *Criminalization of Competition Law Enforcement*, Edward Elgar Publishing.

[59] Zasu, Y. (2007), "Sanctions by Social Norms and the Law: Substitutes or Complements?" *Journal of Legal Studies* 36, 379-396.

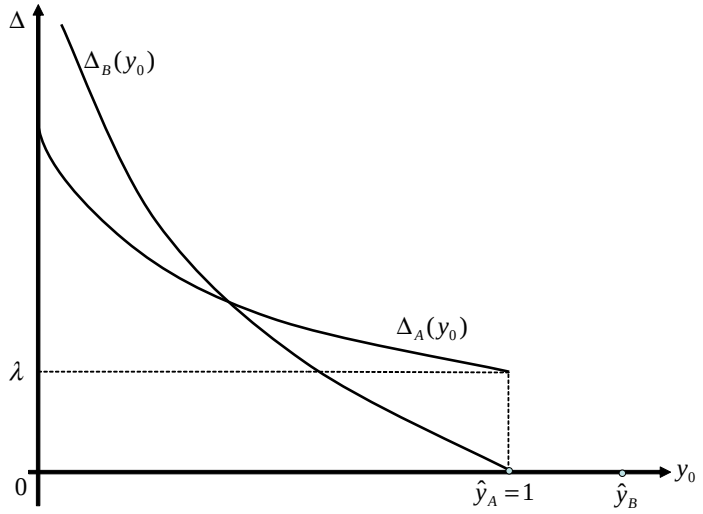

Figure 1. Reputational penalty curves

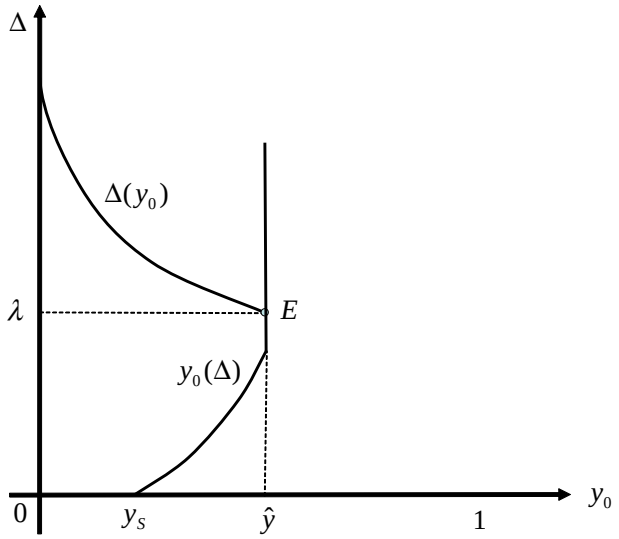

Figure 2. Equilibrium with binding $\widehat{y} < 1$

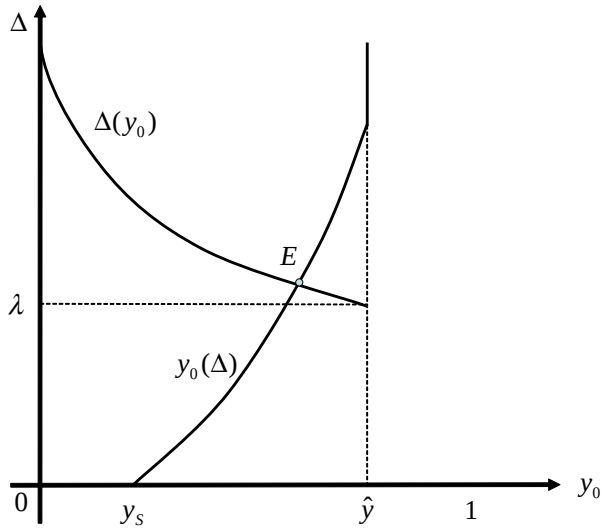

Figure 3. Equilibrium with non binding $\widehat{y} < 1$

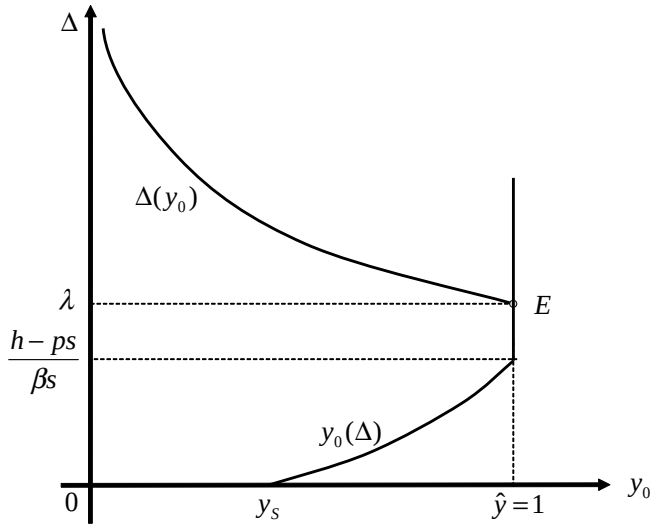

Figure 4. First-best equilibrium with $\widehat{y} = 1$

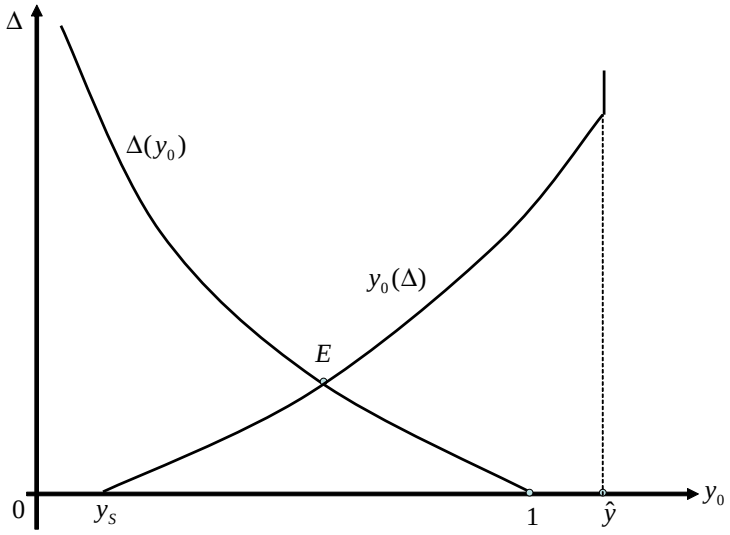

Figure 5. Equilibrium with $\widehat{y} > 1$

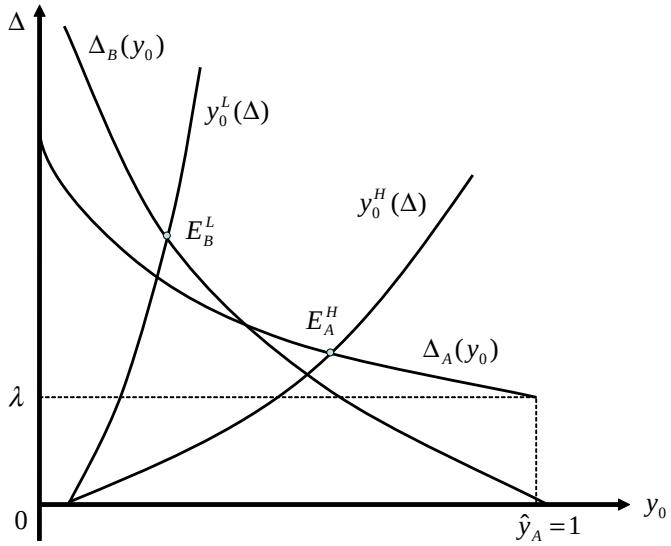

Figure 6. Comparing regimes when $\beta_L < \beta_H$