Stigmatization, Liability and Public Enforcement of Law

Clemens Buchen, Bruno Deffains, Alberto Palermo

Dans Revue d'économie politique 2019/2 (Vol. 129), pages 235 à 259

Éditions Dalloz

ISSN 0373-2630
DOI 10.3917/redp.292.0235
In the theory of public enforcement of law the choice of the liability rules is between strict liability and fault-based liability. In this paper, we study the determinants of compliance when in addition to standard economic incentives wrongdoers take into account stigmatization costs. In this context, this cost is not simply a transfer of resources. We show that a non-guiltiness standard — the fault standard equal to the deterrence level — is never optimal. In this scenario, we show how the optimal policy choice depends on the interplay between the magnitude of the harm and the stigmatization cost.

1. Introduction

Economists have traditionally emphasized a model of legal compliance where the key policy instruments are the severity and likelihood of formal
sanctions such as damages or fines. This view has been significantly broadened over time. For instance, there is considerable evidence that tax compliance or respect for environmental regulations is far greater than a purely self-interested cost-benefit calculation would suggest (Nyborg und Rege [2003]; Slemrod [2007]). Behavior is also shaped by personal values as well as by social rewards and sanctions (Daughety und Reinganum [2010]).

The paper contributes to an emerging field of research in the public enforcement of law which draws on the empirical findings of the “behavioral economics” or “experimental economics” approaches. This literature has much emphasized complementarities between incentives and other motives of behavior. Accordingly, pecuniary incentives may crowd out – or conversely crowd in – informal incentives based on moral and ethical norms, intrinsic motives, image concerns and other forms of social or other-regarding preferences (see e.g. the surveys by Frey und Jegen [2001] and Bowles und Polania-Reyes [2012]). From a law-and-economics perspective, the issue is how normative motivations interact with formal sanctions and the extent to which they are substitutes or complements (see, among others, Tyran und Feld [2006]; Lazzarini et al. [2004]; Zasu [2007]; Galbiati und Vertova [2008]). An obvious implication is that efficient legal sanctioning and enforcement should take these phenomena into account (Bowles und Hwang [2008]; Bénabou und Tirole [2011]). Crowding effects may also have implications for the choice of the “substantive” legal rule (Deffains und Fluet [2013]). Moreover, laws or regulations are more likely to be complied with if they are perceived as appropriate and fair (Posner [2000]; McAdams und Rasmusen [2007]).

Social preferences affect compliant behavior through many channels. The model described below explores some of these channels under different legal regimes. We adopt the same distinction about possible sanctioning rules as expressed by (Polinsky und Shavell [2007], 407) in their survey of the economic theory of public enforcement of law: “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 or regulatory requirement.”

Many recent cases show that an individual found to have caused harm faces not only the possibility of legal punishments such as fines but also the cost of suffering stigmatization such as boycotts or, more generally, disapproval. The law-and-economics literature has already studied stigma in relation to criminal activity (Rasmusen [1996]; Harel und Klement [2007]; Zasu [2007]; Iacobucci [2014]; Mungan [2016]). Our paper inquires more precisely how a concern for social disapproval and stigmatization interacts with legal incentives and how this affects the relative performance of different legal regimes.
A recent example is the Volkswagen emissions scandal. Government regulatory agencies initiated investigations on Volkswagen in many countries and in the days after the news, Volkswagen’s stock price fell in value by a third. The company is facing heavy fines for violating environmental laws. A survey by AutoPacific has compared consumers’ opinion of Volkswagen before and after the scandal and revealed that while before 3 out of 4 vehicle owners had a positive opinion about the brand, after the news only 1 out of 4 preserved such opinion. In an interview by Grieb [2015], Ed Kim, Vice President of Industry Analysis at AutoPacific, said: “this change in consumer opinion will put a significant dent in the brands’ overall sales.”

This suggests that the effect from a reputation damage resulting in stigmatizing or boycotting behavior is potentially large. Armour et al. [2017] report that the reputation effect of enforcement of violations of financial regulations and listing rules on firms is up to nine times larger than the actual fine.

Many experimental or field studies have shown that social image concerns are important motivators of prosocial behavior (Dana et al. [2006]; Ellingsen und Johannesson [2008]; Andreoni und Bernheim [2009]; Ariely et al. [2009]; Funk [2010]; Lacetera und Macis [2010]; Deffains et al. [2017]). In our framework, infractions are not directly observable by society at large. However, adverse court judgments provide public information about the defendants’ actions. Under either strict liability or the fault-based regime, moral concerns are shown to provide individuals with some deterrence incentives.

A basic result in Deffains und Fluet [2015] is that the fault-based rule is more effective than strict liability in harnessing illegal behavior in presence of intrinsic moral concerns. They model individuals who are intrinsically motivated in earning social esteem (“utility of being perceived as a good citizen”) and at the same time they make the difference between “good” and “bad” citizens, where the first internalize the harm they cause. The explanation for their finding is that trial outcomes are more informative allowing for a more precise inference about an infractor’s intrinsic predispositions. Socially useful incentives are therefore derived from the signaling role of “fault”. This is possible, because the intrinsic goal to earn social esteem can act as a positive externality, which can be more efficiently activated by a fault-based regime.

Complementary to this line of argument, we consider extrinsic motives instead of intrinsic predispositions. In our framework, offenders do not internalize the harm they cause, but rather face additional external costs. We consider a model of public enforcement of law where offenders are not always sued, e.g., it is not always feasible to prove harm or to identify the injurers. In our model, offenders consider the costs that would come from a possible stigmatization in the event of detection which, from their point of

---

1. In September 2015, the United States Environmental Protection Agency (EPA) reported that Volkswagen violated the Clean Air Act by installing a tampered emission control system on many cars with a diesel engine. This software was able to detect when the car was being tested, providing performance in adherence with all federal emissions levels, but switching to a separate mode and exceeding safe emission levels when driving normally.
view, constitutes an additional fine. In the case of strict liability the loss from stigmatization is independent of the level of care. Each offender is treated equally because the court ruling does not contain this piece of information. Under fault-based regime the court ruling provides higher informational content. This is so because under such a mechanism the court determines the circumstances of the harmful act. This allows a more fine-grained differentiation of the intensity of the stigmatization as now it is possible to distinguish between offenders “far away” and “close to” the prescribed fault standard. We show that when offenders can suffer external costs imposed by the society, and in contrast to the literature, the standard has some effects on the deterrence level.

Section 2 presents the literature. In section 3, we develop the basic setup, for which all proofs and derivations are relegated to the appendix. Section 4 discusses the optimality and compares strict liability and fault-based liability. Section 5 concludes.

2. Literature

To understand economic behavior, such as public goods contributions, employee relations, consumption of socially responsible products, and more, we must account for the role of social preferences in the choices people make. Economists have long recognized other-regarding preferences, including altruism (Becker [1974]; Andreoni [1989]), spitefulness, and reciprocity (Fehr und Gächter [2000]; Charness und Rabin [2002]; Sobel [2005]; Falk und Fischbacher [2006]). They have also recognized social image concerns (Holländer [1990]; Bernheim [1994]; Andreoni und Bernheim [2009]; Grossman [2015]). It is tempting to try to attribute social behavior (such as giving freely to others) to one of these motives as opposed to the others, but this is a false choice—there is evidence that both kinds of social preferences are at play together (Bowles und Gintis [2013]; DellaVigna et al. [2012]).

Shavell [2002] proposes a general discussion on legal sanctions versus informal motivation as regulators of conduct for accident law. The risk of lawsuits induces precautions to prevent accidental harm to third parties. In the economic model of legal liability, incentives to exercise care reduce to the “implicit prices” set by tort rules (see e.g. Brown [1973]; Landes und Posner [1987]; Shavell [1987]). Casual observation suggests that other motivations are often also at work. Most people exercise some care out of intrinsic concerns about hurting others or out of extrinsic concerns because they fear social disapproval. In this paper, we augment the standard model to include the extrinsic concerns.

The paper is squarely in the tradition of models of public enforcement of law (see e.g. Polinsky und Shavell [2007] for a survey). The canonical model provides a utilitarian analysis of the use of governmental agents—inspectors, tax auditors, prosecutors—to detect and sanction viola-
tors of legal rules. Illegal acts—actions that impose negative externalities—are deterred by the threat of sanctions. In the simplest version of the model, the questions addressed are the severity of the sanction and the resources that should be spent on detecting violations, hence the degree of deterrence that will be achieved. The basic results are then: (i) when feasible, sanctions should take the form of fines and be set very high so as to minimize on detection costs; (ii) generally speaking, not all potential violators will be deterred and not all violations will be detected, and (iii) in the presence of costly sanctions, such as imprisonment, fault-based liability is preferred in that it not only leads to first-best deterrence, but also avoids anyone actually having to bear the cost of a sanction. The basic model has been further developed in many directions. In particular, introducing risk aversion on the part of potential violators tends to lower the amount of the prescribed fines and to increase the probability of detecting infractions (Polinsky und Shavell [1979]; Kaplow [1992]). “General” as opposed to “specific” enforcement – whether public monitoring is directed at a specific type of infraction or simultaneously covers many types – yields some proportionality between the social harm associated with infractions and the fines imposed on infractors.

The behavioral-economics literature has much emphasized the possibility that pecuniary incentives may undermine informal motivations; e.g., the much quoted study by Gneezy und Rustichini [2000] on the crowding-out effect of fines and the survey by Frey und Jegen [2001]. If crowding-out effects are sufficiently strong, legal liability in the tort context could well be counterproductive and reduce precautions to prevent accidental harm. Conversely, it could be that informal motivations and legal sanctions combine to generate too many incentives. For instance, Cooter und Porat [2001] ask whether courts should deduct “nonlegal sanctions” from legal damages to avoid overdeterrence.

Deffains und Fluet [2013] consider two benchmarks: no-liability versus perfectly enforced legal liability (both strict liability and the negligence rule). In the absence of legal liability, injurers take precautions, if at all, solely out of moral or image concerns. The main conclusion is that under the negligence rule, there may be motivational crowding-in. Because of the signal sent by a negligence ruling, image concerns tend to induce bunching on the legal due care standard. Thus, when enforcement is imperfect, the negligence rule may do much better than strict liability because of the individuals’ concern for social approval. Yet, Fluet und Mungan [2017], in a model where the defendants choose the level of care and potentially face stigmatization, focus on the possibility of the court making a mistake. They show that fault-based liability performs better than strict liability if courts do not make too many errors in assessing the level of care of the defendant.

We share with the previous literature the idea that the determinants of compliance with laws include moral considerations. However, our approach differs from previous papers in that the cost imposed by a stigmatization on the wrongdoer is an additional cost, which is not simply a transfer of resources: there are no other effects on the utility of neither the wrongdoer nor the individual who stigmatizes. In this scenario, we analyze in which way
an optimal liability rule depends on the magnitude of the harm and the cost of stigmatization.

3. Model

Consider a population of risk neutral individuals (firms). Each of them derives a private benefit \( g \) from performing a wrongful act which has a negative externality effect \( h \). This harmful damage is independent of the gains, which are uniformly distributed in the population over the support \([g, \bar{g}]\). Without loss of generality we set \( g = 0 \) and \( \bar{g} = 1 \). Moreover, we assume that \( h < 1 \); that is, some individuals have a benefit which exceeds the harm caused. The gains \( g \) determine the type of firms, that is, neither the value of the gain nor the level of the harm are choice variables of the potential wrongdoers.

A Social Planner (SP) chooses a probability of detection \( p \) and a fine \( s \) to maximize the social welfare. In addition either strict liability or fault-based regimes can be imposed. In the latter case the SP sets a fault standard. To clarify the distinction between strict and fault-based liability, suppose that a firm has to decide whether to properly recycle a given level of waste or not. In the latter case it would result in a harm \( h \). Recycling—and thereby avoiding the harm \( h \)—comes at a given opportunity cost \( g \). Strict liability in this context implies that the firm has to pay a fine if caught polluting. Fault-based liability, on the other hand, means that the firm would have to pay a fine if its opportunity costs \( g \) are lower than the prescribed standard. For any \( g \) above the standard it would be exempt.\(^2\)

By assumption the SP never observes ex ante the gain which is private information for individuals, but he knows the distribution of gains. The fine cannot exceed the individuals’ wealth which is assumed equal for all and normalized to 1. The welfare equals the gains and harms caused in the population minus the expenditure cost for detection.

We suppose that the per-capita expenditure by the government is a function of the probability of detection \( c(p) = \frac{p}{2} \). Moreover, we assume that the government budget constraint for financing \( p \) breaks even: \( \tau + \phi ps = c(p) \), with \( \tau \) defining the per-capita tax and \( \phi \) the proportion of guilty individuals. Hence, \( \tau \) and \( \phi \) are endogenous. There is an additional appraisal cost in the case of fault-based-liability. This cost arises, because in addition to detecting the infraction, the court has to establish whether the opportunity costs are within the standard, or not.

A person who commits the harmful act pays an expected fine \( ps \) and moreover faces a cost coming from the stigmatization \( C(\alpha) = z_1 \alpha \). In this function, \( \alpha \in [0, 1] \) is the fraction of individuals who are willing to express disapproval for the agent committing the harmful act (i.e. stigmatizing the

\(^2\) This is essentially Polinsky and Shavell’s [2007] leading example.

REP 129 (2) mars-avril 2019
wrongdoer) whenever they get information about the illegal behavior. In other words, \( \alpha \) can also be interpreted as the probability that the stigmatization occurs. The parameter \( z_i \in [0, 1] \) measures the intensity of the stigma. It is important that whereas \( \alpha \) is independent of the institutional choice of the liability rules, the intensity \( z_i \) will depend on whether the law foresees strict liability or fault-based liability, with \( i \in \{ S, F \} \), where \( S \) stands for strict liability and \( F \) for fault-based liability. In other words, there exists a potential stigmatization within the population given by \( \alpha \). The intensity \( z_i \) determines how much of this potential the individual wrongdoer can experience as stigma.

To summarize, there is one population of both potential injurers and potential condemners of the wrongdoer. The two subsets can overlap and all members are risk neutral. The assumption of risk neutrality of the judging individuals is without loss of generality for the comparison of policies (see explanations in footnote 11 below).

3. There is an alternative probabilistic interpretation of \( C(z_i ; \alpha) \). If \( \alpha \) is the probability that a member of the public hears about the infraction and \( z_i \) the probability that an individual decides to stigmatize the product of the two independent events gives the probability that the stigma occurs, so \( z_i \alpha \). The intensity would then be simply equal to 1.

4. There are other forms of non-transferable sanctions such as imprisonment, probation or electronic monitoring. They have in common with stigmatization that they impose a cost to the society, which is not compensated by any direct beneficial effect except for an increase in the deterrence level. Despite this similarity, there are few differences. First, imprisonment and other sanctions impose a disutility to the wrongdoer and a cost to the state for sustaining the system; stigmatization does not imply the latter cost. Second, whereas the imprisonment is a direct sanction for which the regulator can choose the amplitude, the existing stigmatization can only be used indirectly. In particular, as we model it, the strength of the stigmatization on the wrongdoer is a result of the policy. Since we focus on this, we do not consider possible extensions, such as the frequency of conviction having an impact on the level of stigmatization à la Harel und Klement [2007].

5. More precisely, we admit that the general public is assumed to be informed only of the court’s ruling, not of the detailed evidence disclosed at trial. It follows that “social esteem” depends on information available at large in the general public.
Substituting the government budget constraint, with \( \phi = 1 - \frac{p}{s} + \frac{h}{H} \), lets us reformulate the welfare as follows:

\[
W^S = 1 + \int_{p (s + \theta \alpha)}^{1} \left( g - h - p \theta \alpha \right) dg - \frac{p^2}{2} \tag{1}
\]

Standardly, the fine \( s \) does not enter directly in the social welfare because it is simply a transfer of resources in the case of a risk neutral specification. However, this is not the case for stigmatization costs, which have a double impact affecting the welfare directly and indirectly. These costs directly reduce the benefit of the injurer without being offset by an increase in the utility of a third party. Indirectly they work as a deterrent mechanism.

The SP maximizes (1) choosing \( s \) and \( p \) facing the countervailing effect of the latter. Then, the fine is maximum and set by assumption equal to the wealth, so \( s = 1 \). The proof for maximality is standard: keep \( p (s + \theta \alpha) \) —the level of deterrence—constant increasing (costlessly) \( s \) and decreasing \( p \) to decrease the enforcement expenditures; that is, the standard Becker [1968]-argument is still valid in this context. Furthermore, now the reasoning about the maximal magnitude of a fine is even stronger since a higher probability would have an additional negative effect represented by the direct cost \( \rho \theta \alpha \).

This argument implies that \( p \) is the only choice variable left and the necessary and sufficient condition for the SP’s maximization problem can be expressed as follows:

\[
(1 + \theta \alpha)(h - p) = \int_{p (1 + \theta \alpha)}^{1} \theta \alpha dg + p \tag{2}
\]

The left-hand side represents the marginal benefit of increasing the probability of detection, which is a marginal increase in the deterrence. The right-hand side gives the marginal costs, which consist of two components. The first is the above-mentioned additional burden on the injurers, whereas the second part is the increased cost of detection for the society. Solving (2) to find the probability of detection, we have:

\[
p^S = \frac{(1 + \theta \alpha)h - \theta \alpha}{2 - (\theta \alpha)^2} \tag{3}
\]

To guarantee an interior solution, we assume that the harm is always relatively large enough: \( h > \frac{\theta}{1 + \theta} \).

In this specification, underdeterrence (overdeterrence) is defined as \( p (1 + \theta \alpha) < ( > )h \). Clearly, in the absence of stigmatization cost ( \( \theta \alpha = 0 \) ) there is always underdeterrence. Conversely, introducing stigmatization cost leads to the following conditions for the level of deterrence.

---

6. Note that \( \frac{\partial W^2}{\partial p} = (\alpha \theta^2) - 2 < 0 \) for the relevant intervals for \( \alpha \) and \( \theta \).

REP 129 (2) mars-avril 2019
Proposition 1. Under strict liability rule, if the stigmatization cost is large enough \( \theta \alpha \in \left( \frac{\sqrt{3} - 1}{2}, 1 \right) \), there exists a threshold \( \hat{h} \) for the negative externality such that for \( h > \hat{h} \) an efficient solution always requires over deterrence; the opposite is true for \( h < \hat{h} \). Conversely, if the stigmatization cost is small enough \( \theta \alpha \in \left[ 0, \frac{\sqrt{3} - 1}{2} \right] \) an optimal solution always implies underdeterrence.

The existence of stigmatization costs is a necessary but not sufficient condition to leave the standard world of underdeterrence. It is the confluence of a high enough cost of the stigma and a large externality, which leads to overdeterrence. If both conditions are met, the SP aims to curtail infractions as much as possible for two reasons. First, the harm is large and therefore should be avoided. Second, since the stigmatization costs are non-transferable, these costs should be avoided as well. Hence, overdeterrence follows.

3.2. Fault-based liability

In the case of fault-based liability the conviction depends on the fault standard applied by the court, \( \hat{g} \). In what follows, we define fault-based liability as a tuple \( \langle p^F, \hat{g} \rangle \), where \( p^F \) is the probability of detection under fault-based liability. We assume that different standards produce different intensities of the stigmatization cost. More precisely, we assume now \( z_F = \theta + \kappa (\hat{g}, g) \), which results in the following cost function:

\[
C(\theta + \kappa (\hat{g}, g) ; \alpha) = [\theta + \kappa (\hat{g}, g) ] \alpha
\]

The informational content of fault-based rulings is higher than strict liability. In order to determine the proofs of guiltiness, the offender's opportunity costs \( g \) become known during trial. Therefore both the wrongdoer's \( g \) and the fault standard \( \hat{g} \) are public information. In contrast to strict liability, the intensity of the stigmatization is not constant anymore because it depends on how “far” a guilty person’s gain is from the standard. This is captured by the function \( \kappa (\hat{g}, g) \), which will be explained in more detail below.

Under fault-based liability, all individuals with \( g > \hat{g} \) are within the fault standard and therefore not guilty. Hence, they commit the harmful act. The remainder will commit the harmful act if \( g \geq \hat{g} \), where \( \hat{g} \) is the deterrence level which solves implicitly:

\[
\hat{g} = p \{ 1 + [\theta + \kappa (\hat{g}, \hat{g}) ] \alpha \}
\]

Note that the fine \( s \) has been set equal to 1, the wealth, since the same reasoning about choosing the maximum possible fine holds. The main purpose of introducing the function \( \kappa (\hat{g}, g) \) is to model more fine-grained stigmatization costs. While in strict liability the stigmatization occurs indiscriminately, now the cost can be higher or lower under the fault-based rule.
Whenever a wrongdoer's gain is within the prescribed fault standard, the stigma costs are zero (as is the fine). The further away (closer to) the individual gain is from the fault standard, the more (less) intense the stigma becomes. Arguably, the more blatantly someone behaves in a negligent way the more angered the public reacts to hearing this news. On the other hand, a “narrow miss” of the fault standard evokes a less intense reaction. In this sense we argue that the fault standard is a “reference point” for the society.

We can use our earlier example to illustrate how we think about the stigmatization here. For a given fault standard that sets the minimum opportunity costs for which firms would not be fined for polluting, suppose that out of two firms violating the standards with an opportunity costs “close” (resp. “far”) to the standard, we argue that society would stigmatize more intensely the firm further away from the standard. As yet another example, suppose a tax policy foresees a tax break to a firm with a certain number of employees, say \( x \). A violating firm which has a number of employees much smaller than the fault standard \( x \) is (or, at least is more likely to be) stigmatized more than another violating firm which avoids paying taxes, but with a number of employees much closer to the standard fixed by law.

Formally, the characteristics of \( \kappa(\hat{g},g) \) are succinctly summarized in the following assumptions:

(A1) \( \text{possible negative societal reaction} \) \( \theta + \kappa(\hat{g},g) \in (0,1] \) \( \forall g \in [\hat{g},\hat{g}) \), \( \forall \hat{g} \)

(A2) \( \text{zero within the standard} \) \( \kappa(\hat{g},g) = 0 \) \( \forall g \in [\hat{g},\hat{g}], \forall \hat{g} \)

(A3) \( \text{proclivity to the standard} \) \( \kappa(\hat{g},g) < 0 \) \( \forall g \in [g,\hat{g}) \), \( \forall \hat{g} \)

(A4) \( \text{accordance to the standard} \) \( \kappa(\hat{g},g) > 0 \) \( \forall g \in [g,\hat{g}), \forall \hat{g} \)

Hence, the stigma cost is different for each benefit \( g \) and in particular: (A1) the society can punish less under fault-based liability as compared to strict liability; (A2) whenever an individual is within the fault standard and therefore not guilty, the stigmatization cost is zero; (A3) the society punishes less whenever the opportunity cost is closer to the standard; (A4) increasing the standard increases the stigma cost for each guilty person. We assume a monotonic relationship between the gain \( g \), the fault standard \( \hat{g} \) and the total stigmatization cost under the fault-based policy. Corresponding to the pollution example above, we say that the closer the pollution emission is to the standard defined by the law, the lower is the stigmatization cost (A3). However, the higher is the standard defined by the law, the stronger \( \text{ceteris paribus (i.e., for a given pollution emission)} \) is the reaction from the society at large (A4). Observe that under assumptions (A2) and (A3), (A1) is a corollary. Under (A2)-(A4), we can claim:

**Lemma 1.** Other things being equal, an increase of the probability of detection, an increase of the fault standard, or an increase of the number of people willing to stigmatize, increases the deterrence level.

From Lemma 1 it follows that the standard itself is a deterrence tool. This is in contrast with the literature where the fault standard neither works as complement nor as substitute of the probability of detection. It means that even under the assumption of risk neutrality, the fault standard loses its passive role in harnessing illegal behavior.
Lemma 2. An optimal policy \((p^F, \hat{g})\) always requires \(\hat{g} \geq \bar{g}\) and the minimum fault standard equals the probability of detection \(\hat{g}_{\text{min}} = p^F\).

The following function for \(\kappa(\hat{g}, g)\) satisfies (A2)-(A4) and allows an easy tractability\(^7\) of the problem:

\[
\kappa(\hat{g}, g) = (\hat{g} - g) - \theta
\]

Given this functional form all individuals with an opportunity cost close enough to the standard \((\hat{g} - g < \theta)\) pay a stigmatization cost lower than the one they would pay under a strict liability policy. The linear example of \(\kappa(\hat{g}, g)\) gives rise to linear stigma costs as a function of the gain \(g\), which are shown in comparison with the constant cost \(\theta\alpha\) under strict liability in Figure 1. The graph clearly shows the main distinction we have in mind for the stigmatization cost. Whereas for strict liability cases the gain of the infraction has no influence on the magnitude of the stigma costs, under fault-based liability stigmatization depends on how bad the infraction is in relation to the fault standard.

Moreover, we assume that besides the higher informational content that the fault-based regime can provide, it requires an expenditure to determine the level of the wrongdoer’s fault.\(^8\) In particular, we assume a fixed appraisal cost for each detected individual equal to \(F\). Accordingly, the total expected

---

\(^7\) The assumption about linearity w.r.t. the two variables can be easily replaced. However, as it will be clearer later, this would simply add algebraic computations without affecting the main claim.

\(^8\) Determining whether a party has been negligent is often a highly fact-intensive process. Adducing proof of negligence may therefore be economically costly and time consuming. For example, showing how carefully the defendant acted on a particular occasion may involve more cost and time than showing that the defendant was involved in a particular activity for which strict liability is imposed. Expert testimony is more likely to be required or permitted to prove negligence than to prove the defendant’s involvement in such an activity. Evidence of compliance with, or violation of, a customary practice is admissible because it is relevant to negligence, but it is unlikely to be relevant in a strict liability action. Because an unexcused violation of an applicable safety statute constitutes negligence, whether a statute
cost for defining the fault standard is simply \( \int_\tilde{g}^1 pFdg \). In what follows, we assume that \( F < h \) avoiding cases in which the mere appraisal of fault would be more costly than the harm. As will become clear later, this would lead to a probability of detection of zero.

The government budget constraint becomes \( \tau + \phi ps = c(p) + \int_\tilde{g}^1 pFdg \). Welfare is now the sum of gains of infractions net of harm minus the expected fine and stigma costs paid by those not within the fault standard minus the tax burden:

\[
WF = 1 + \int_\tilde{g}^1 (g - h)dg - \int_\tilde{g}^1 p(s + [\theta + \kappa (\tilde{g}, g)]\alpha)dg - \int_{\tilde{g}}^1 pFdg - \frac{\tilde{F}^2}{2}
\]

Substituting the government budget constraint, with \( \phi = \tilde{g} - \tilde{g} \), welfare can be written as:

\[
WF = 1 + \int_\tilde{g}^1 (g - h)dg - \int_\tilde{g}^1 p(s + [\theta + \kappa (\tilde{g}, g)]\alpha)dg - \int_{\tilde{g}}^1 pFdg - \frac{\tilde{F}^2}{2} \tag{5}
\]

**Lemma 3.** The necessary and sufficient conditions for the unique interior solution are:

\[
\frac{\partial W^F}{\partial p} = 0 \iff \frac{\partial \tilde{g}}{\partial p} (h - p + pF) = \int_\tilde{g}^1 (\tilde{g} - \tilde{g})\alpha dg + F(1 - \tilde{g}) + p \tag{6}
\]

\[
\frac{\partial W^F}{\partial g} = 0 \iff \frac{\partial \tilde{g}}{\partial g} (h - p + pF) = \int_\tilde{g}^1 p\alpha dg \tag{7}
\]

The first equation (6) expresses the marginal benefit (left-hand side) and marginal cost (right-hand side) for the probability of detection keeping the standard constant. The marginal costs consist of three terms. The first accounts for an increase in stigmatization because of the higher detection rate, the second represents the increase in appraisal costs because more infractions are detected\(^9\) and the third is the linear marginal cost of a higher probability of detection. The second equation (7) records the first-order condition for the fault standard. Again, the left-hand side is the marginal benefit of the standard keeping the probability of detection constant. The right-hand side represents the marginal increase in the stigmatization costs, because the higher the fault standard, the higher the cost for all infractions that are not within the fault standard (see also Assumption (A4) above).

---

\( ^9 \) The presence of the appraisal cost has the effect of shifting down the value function for fault-based liability. Applying the envelope theorem to (5): \( \frac{\partial W^F}{\partial F} \bigg|_{\tilde{g}, \tilde{g}} = -pF(1 - \tilde{g}) < 0 \). Therefore, a large enough fixed cost \( F \) would imply superiority of strict liability. However, we are able to show this superiority even in absence of appraisal cost.

**REP 129 (2) mars-avril 2019**
Including the costs from the stigmatization leads to a solution in contrast with standard arguments about the superiority of fault-based with non-transferable costs. In the standard framework the resulting standard can be characterized as non-guiltiness standard.\textsuperscript{10} By this we mean an outcome where the society is split in two parts: those individuals who do not commit the infraction because the expected cost outweighs the benefit and those whose benefit is within the fault standard. In our model, this would translate in $\hat{g} = \tilde{g}$. However, as it is clear from the first-order condition (7) this is only a special case in our model.

Solving (7) gives an expression for the fault standard:

$$\hat{g}^* = h + p^F F$$ \hspace{1cm} [8]

In principle, using (8) it is possible to solve (6) for $p^F$. However, since this is a cubic equation, we proceed without a closed-form solution for $p^F$.

**Corollary 1.** Under a non-guiltiness standard policy ($\hat{g} = \tilde{g}$) the optimal probability of detection and the fault standard are:

$$p^F = \frac{h - F}{2(1 - F)} = \hat{g}$$

Moreover, an optimal fault-based policy $\langle p^F, \hat{g} \rangle$ under stigmatization costs never requires a non-guiltiness standard and it holds that $\hat{g} \neq h$ for $F > 0$.

Intuitively, non-guiltiness is never optimal, because implementing it would always imply foregoing the opportunity to use the standard as an additional tool to affect the deterrence level and in order to more efficiently spread the non-transferable cost from the stigmatization. In a model of public enforcement of law the first-best is where $\hat{g} = h$ (Polinsky and Shavell [2007], 409), because then the legal system “allows” infractions for types, for whom the gain is greater than the harm. We see that this is still true once the external stigmatization costs are introduced, unless there are appraisal cost $F > 0$. In the following, we show that indeed under fault-based liability, we can expect underdeterrence; that is, $\hat{g} < h$.

**Proposition 2.** Under fault-based rule, an optimal solution always requires underdeterrence.

To better understand the intuition assume for the moment that $F = 0$. Then (8) and underdeterrence imply $\tilde{g} < \hat{g} = h$. Observe that without stigmatization any standard $\hat{g} \in [\tilde{g}, h]$ can be chosen, because the fault standard alone has no impact on the deterrence. Once the stigmatization costs are introduced decreasing the fault standard from its first-best would impose a welfare loss, because the deterrence level would decrease.

\textsuperscript{10} Shavell [1987] is the first to recognize the relevance of the non-guiltiness standard, while not so named.
4. Discussion

Having determined the probability of detection for each policy (and the fault standard) we now define under which conditions one policy performs better than the other. Denote the value functions of the respective welfares as $W^S(\ast) = W^S\mid_{p,\hat{\theta}}$, $W^F(\ast) = W^F\mid_{p,\hat{\theta}}$, and the value function of fault-based regime under the non-guiltiness standard as $W^F\mid_{p,\hat{\theta}}$. Whereas the welfare of a fault-based rule under non-guiltiness standard is obviously independent of stigmatization cost, we have:

**Lemma 4.** $W^S(\ast)$ is strictly convex w.r.t. $\hat{\theta}$ everywhere. Moreover, for a small enough harm $\left(\frac{\theta}{1+\theta} < h \leq \frac{2}{2+\theta}\right)$ the value function is always decreasing for every $\alpha$, while for a large harm $\left(\frac{2}{2+\theta} < h < 1\right)$ it is U-shaped.

And for fault-based liability:

**Lemma 5.** $W^F(\ast)$ is strictly increasing and strictly concave in $\alpha$ and bounded from below by $W^F\mid_{p,\hat{\theta}}$.

Combining Lemma 4 and 5, we have the following result:

**Proposition 3.**

1. For a sufficiently small appraisal cost, the following hold:
   
   i) for relatively small negative externalities $\left(\frac{\theta}{1+\theta} < h \leq \frac{2}{2+\theta}\right)$ there exists an $\hat{\alpha}$ such that for $\alpha < \hat{\alpha}$ fault-based policy is welfare maximizing; the opposite is true for $\alpha > \hat{\alpha}$.

   iii) for relatively large negative externalities $\left(\frac{2}{2+\theta} < h < 1\right)$ and for a large stigmatization cost ($\theta$ large enough) there exist $\alpha^\ast$ and $\alpha^{**}$ such that for $\alpha \in (0, \alpha^\ast)$ and for $\alpha \in (\alpha^{**}, 1)$ strict liability policy is welfare maximizing; the opposite is true for $\alpha \in (\alpha^\ast, \alpha^{**})$.

2. For a sufficiently large appraisal cost, strict liability policy is always welfare maximizing.

Proposition 3 gives the main result. Here we give the intuition. If $h$ is relatively small and the appraisal costs $F$ are not too high, there exists a threshold for $\alpha$, such that strict liability performs better than fault-based regime for smaller values of $\alpha$, while the opposite is true for values above this threshold. This case is depicted in Figure 2a. The non-guiltiness standard acts as lower bound for the welfare of fault-based rule and in case of no stigmatization ($\alpha=0$), the difference to strict liability depends on the magnitude of the appraisal cost and the damage $(\mathcal{A}(F, h))$. For all $\alpha > 0$, that is, whenever there is stigmatization, the optimal policy will never be given by a fault-based policy cum non-guiltiness standard, as explained in Corollary 1.
It is true that in the case of relatively small $h$ the superiority of a strict liability regime is due to the introduction of the appraisal costs. Also the second part of the Proposition is intuitively clear; if the appraisal costs in fault-based become very large, strict liability will always be welfare maximizing, simply because the society does not bear those fact-finding costs. What is more, introducing stigmatization costs amplifies the superiority of the fault-based regime as compared to the standard literature. However, and importantly, this is not anymore true if the harm increases.

For relatively large $h$ the welfare comparison is shown in Figure 2b. From Lemma 4 we know that for relatively large $h$ the welfare of strict liability is U-shaped. This implies that there is potentially a second intersection of the two value functions. Hence, the standard superiority of fault-based liability is not lost just because of an additional social expenditure that comes from the appraisal costs.

To explain the rationale behind this claim, recall the basic trade-off created by the presence of stigmatization costs. On the one hand, the stigma affects the deterrence level, which allows the SP to save enforcement costs. On the other hand, these costs are not a transfer of resources and therefore the higher the stigmatization costs the higher the burden on the injurer. If the harm is relatively large, the SP wants to avoid infractions and therefore keep the deterrence high. For relatively small $\alpha$ the aforementioned trade-off is best solved by a fault-based rule, because the benefit of a low probability of detection outweighs the negative effect coming from the stigmatization costs. More clearly, the fault-based regime allows to better spread this non-transferable cost, because the reaction of the society differs according to the “distance” of the wrongdoer from the standard. However, if $\alpha$ increases, the pendulum swings towards strict liability. Now, this relatively high non-transferable cost should be avoided and this can be done if the deterrence increases. This depends also on a minimum level of the intensity of the stigma $\theta$, because the stigmatization costs facing the wrongdoer actually are $\theta\alpha$. Therefore, the choice of a lawmaker is to utilize this mechanism not only as an alternative for a costly enforcement, but primarily as a tool to harness illegal behavior. Strict liability is more effective in this sense because it can utilize the full potential of stigmatization. Applying a strict liability regime,
and therefore not giving information about the level of guiltiness, increases
the cost for all offenders compared to a fault-based regime tightening the
incentive to infract regardless of the personal benefit.¹¹

5. Concluding remarks

The paper introduces stigmatization effects into a standard model of pub-
lic enforcement of law. The injurer takes the potential negative effect of
stigmatization into account. The model specifically draws on the different
informational content of different liability regimes, strict liability or fault-
based rules. Since in the case of fault-based liability courts have to engage
in fact-finding to ascertain whether the wrongdoer adhered to the fault stan-
dard or not, the specific gain from the infraction is known. This implies that
in the latter case stigmatization can depend on the gain of the injurer,
whereas in the case of strict liability it does not. Put simply, in a strict liability
regime the effect of stigmatization is a constant, whereas in a fault-based
regime it depends on how far away the gain is from the fault standard.

The normative aim is then to draw conclusions about the welfare under
either legal regime. The relative superiority depends both on the prevalence
of stigmatization in the society and the level of the harm. The basic trade-off
comes from the benefit of stigmatization as a deterrence tool and the cost it
puts on a wrongdoer. How this trade-off is solved best depends crucially on
the interplay between the harm and the intensity of the stigma. If the harm
is relatively small fault-based rules will tend to perform better as long as the
additional appraisal costs are not too large. This is so, because now in order
to solve the aforementioned trade-off more weight is put on the cost effect
of stigmatization. This can be achieved by the more fine-grained approach,
which is offered by fault-based liability. On the other hand, if the harm is
relatively large strict liability performs better if the prevalence of stigmatiza-
tion is large enough. If a sufficiently high fraction of the population is willing
to engage in stigmatization a strict liability regime is superior because it can
entirely use the stigmatization as a tool to harness illegal behavior. In this
case the deterrence motive dominates.

Broadening the types of incentive that injurers take into account leads to
conclusions that are not in line with the previous literature. In a model of risk
neutral individuals without appraisal costs and without stigmatization the

¹¹. These results remain if the assumption of risk neutrality for the condemners is relaxed.
Suppose there is an additional population of individuals who condemn, but do not commit a
harmful act. Moreover, for simplicity, assume that the taxation is only imposed on the firm
level. In this case, an individual would suffer a disutility whenever the deterrence level
increases. Formally, the society at large faces in expectation the additional term
\( \phi U(w) + (1 - \phi)U(w - h) \). In this scenario, the marginal benefit of increasing \( p \) is
\[
\frac{\partial}{\partial p} [U(w) - U(w - h)].
\]
Hence, there is an incentive of the SP to keep the deterrence level
higher according to the degree of risk aversion. However, given the linearity in \( \phi \) the shape
of the value function is unaltered, but the threshold of \( h \) that makes strict liability potentially
better than fault-based liability is reduced.
conclusion about the superiority of fault-based liability follows from the fact that it gives the law maker an additional tool. On the one hand, fines are irrelevant because they constitute a redistributive source of wealth. On the other hand, the fault-standard can be chosen so that it equals the deterrence level. Therefore, even if for a utilitarian lawmaker fault-based and strict liability are indifferent in terms of efficiency, the choice should be in favor of the former. In our model, abstracting from appraisal costs, this is not necessarily the case, because with stigmatization costs the fault standard affects the deterrence level. The main result of this paper is about this trade-off of stigma as deterrence versus stigma as non-transferable costs and the normative conclusions of having strict liability as welfare-maximizing policy also in presence of risk neutrality.

Appendix

**Proposition 1.** For the first part, \( p^S(1 + \theta \alpha) > h \) after substitution becomes:

\[
h [2(\theta \alpha)^2 + 2(\theta \alpha) - 1] > \theta \alpha (1 + \theta \alpha) \tag{9}
\]

Expression (9) solved with equality gives the threshold \( \bar{h} = \frac{\theta \alpha (1 + \theta \alpha)}{2(\theta \alpha)^2 + 2(\theta \alpha) - 1} \).

The term in square brackets is strictly positive for \( \theta \alpha \in \left(\frac{\sqrt{3} - 1}{2}, 1\right) \). Then, only in this case, for every \( h > \bar{h} \) (9) is satisfied. For the second part, if \( \theta \alpha \in \left[0, \frac{\sqrt{3} - 1}{2}\right] \), then (9) is never satisfied and therefore an optimal solution always requires underdeterrence.

**Lemma 1.** Applying the implicit function theorem to (4) and using (A2), (A3) and (A4), the results are straightforward:

\[
\frac{\partial \tilde{g}}{\partial p^F} = \frac{1 + \left[\theta + \kappa (\tilde{g})\right] \alpha}{1 - pK_{\tilde{g}} (\tilde{g}) \alpha} > 0 \tag{10}
\]

\[
\frac{\partial \tilde{g}}{\partial \alpha} = \frac{pK_{\tilde{g}} (\tilde{g}) \alpha}{1 - pK_{\tilde{g}} (\tilde{g}) \alpha} \geq 0 \tag{11}
\]

\[
\frac{\partial \tilde{g}}{\partial \alpha} = \frac{p \left[\theta + \kappa (\tilde{g})\right]}{1 - pK_{\tilde{g}} (\tilde{g}) \alpha} \geq 0 \tag{12}
\]

**Lemma 2.** The first part is done by contradiction. Suppose \( \tilde{g} > \hat{g} \). Then, the given probability of detection is not a deterrent and therefore for every optimal policy the costly probability should be decreased; the result is then a consequence of Lemma 1 \( \left(\frac{\partial \tilde{g}}{\partial p^F} > 0\right) \). For the second part, if \( \tilde{g} = \hat{g} \), then from (A2) it follows that \( \hat{g} = p^F \) and therefore \( \hat{g}_{\min} = p^F \).
Remark: In what follows, to derive our main results, we make use of the specific functional form for $\hat{g} / g$. Hence, in order to simplify the reading of the following proofs and better grasp the intuitions, we report (4), (10), (11), and (12) for $\hat{g} / g = \hat{g} - g - h$. Moreover, note that since from Lemma 2 $\hat{g} \leq \tilde{g}$, the following expressions hold when the foregoing inequality applies.

\[
\hat{g} = \frac{p (1 + \hat{g} \alpha)}{1 + p \alpha} \tag{13}
\]

\[
\frac{\partial \hat{g}}{\partial p} = \frac{1 + \hat{g} \alpha}{(1 + p \alpha)^2} > 0 \tag{14}
\]

\[
\frac{\partial \hat{g}}{\partial \hat{g}} = \frac{p \alpha}{1 + p \alpha} > 0 \tag{15}
\]

\[
\frac{\partial \hat{g}}{\partial \alpha} = \frac{p (\hat{g} - p)}{(1 + p \alpha)^2} \geq 0 \tag{16}
\]

Lemma 3. In order to check sufficiency we evaluate the Hessian matrix for the problem (5):

\[
\hat{g}' = \left[ \begin{array}{c}
-2F (1 + \hat{g} \alpha) - a \hat{g}^2 + 2ah (1 + \hat{g} \alpha) + \alpha \hat{g}^2 + 3a^2 p^2 + 3ap + 2 \\
\alpha (Fp (2 + ap) - \hat{g} + h) \\
\alpha (Fp (2 + ap) - \hat{g} + h) \\
\end{array} \right] 
\]

where, at the optimum with $\hat{g}^* = h + pF$, the first minor is negative and the determinant is clearly positive:

\[
\frac{p \alpha (2 (1 + ah) (1 - F) + p \alpha (3 + 3p \alpha + p^2 \alpha^2) (1 - F^2) + h^2 \alpha^2)}{(1 + p \alpha)^4} > 0.
\]

To show the uniqueness of the solution we let w.l.o.g. $F = 0$ to simplify the algebra. Then, the f.o.c. (6) becomes:

\[
\frac{\alpha h^2 + 2h - p (2 \alpha^2 p^2 + 5ap + 4)}{2 (ap + 1)^2}
\]

In setting the numerator of the previous equation equal to zero we would find the optimal $p^F$. We use Mathematica to compute the discriminant:

\[
\text{Discriminant} = -4 (28 \alpha^2 + 27 \alpha^6 h^4 + 108 \alpha^5 h^3 + 163 \alpha^4 h^2 + 110 \alpha^3 h) < 0
\]

Since the discriminant is negative, only one solution exists.
Corollary 1. To show the first part, observe that from (4) a non-guiltiness standard policy $\hat{g} = \hat{g}$ implies $p = \hat{g}$. Therefore, the welfare (5) becomes

$$W^F \big|_{\hat{g} = \hat{g}} = 1 + \int_p^1 (g - h)dg - \int_p^1 pFdg - \frac{p^2}{2}$$

The f.o.c. is:

$$-(p - h) + pF - F + pF - p = 0$$

which gives $p^F = \frac{h - F}{2(1 - F)}$. For the second part, to show that under stigmatization costs an optimal policy never requires a non-guiltiness standard, assume that by contradiction $\hat{g} = \hat{g}$. Then, from (4), $p^F = \hat{g}$, therefore $p = \hat{g}$. Using (8), it follows that the left-hand side of (6) becomes zero, as does the integral on the right-hand side. This implies that $0 = F(1 - p) + p$, which would lead to a negative probability. Lastly, the fact $\hat{g} = h$ comes again from (8).

Proposition 2. We want to show that $h > \hat{g}$. Using (13) and (8), we have:

$$h - \hat{g} > 0 \iff h - \frac{p^F [1 + (h + p^F F)\alpha]}{1 + p^F \alpha} > 0 \iff \frac{h - p^F - (p^F)^2 F\alpha}{1 + p^F \alpha} > 0.$$

Since the numerator of the last expression determines the sign, to show that $h - \hat{g} > 0$ we define $\Phi(p^F) = h - p^F - (p^F)^2 F\alpha$. Then, we show that $\Phi(p^F)$ is always positive in the relevant interval for $p^F$. In particular, since $p^F$ is decreasing in $\alpha$ (see proof of Lemma 5 below) its upper bound is $p^F_{\text{Max}} = p^F(\alpha = 0) = \frac{h - F}{2(1 - F)}$. Observe that $\Phi(p^F)$ is a concave parabola with a maximum in the positive orthant and for $p^F = 0$ intersects the vertical axis in $h$. Therefore, it is enough to show that $\Phi(p^F_{\text{Max}}) > 0$. Substituting and rearranging:

$$\Phi(p^F_{\text{Max}}) = \frac{-\alpha F^3 + 3\alpha F + 2\alpha F^2 h + 4F^2 h - 2F^2 - \alpha F h^2 - 6F + 2F + 2h}{2(1 - F)(1 - F) + \alpha(h - F)}$$

Since $h > F$ the denominator of the previous expression is strictly positive and therefore the numerator determines the sign. It can be written as follows:

$$2F(1 - F)(1 - h) + 2h(1 - F)^2 - \alpha F(h - F)^2 > 2F(1 - F)(1 - h) + \alpha F(1 - F)^2 > 0$$

where the second inequality comes from $1 > h > F > 0$ and $\alpha \in (0, 1)$.

Lemma 4. Applying the envelope theorem to (1), we have:

$$\frac{\partial W^S}{\partial \alpha} \bigg|_{\theta^*} = p^S \theta(h - 1 + p^S \theta \alpha)$$

[17]
which proves that the $W_s^s(*)$ function can be decreasing or increasing. To show strict concavity, differentiating (17) w.r.t. $\alpha$:

$$\frac{\partial^2 W_s^s(*)}{\partial \alpha^2} = \frac{\partial p_s}{\partial \alpha} \theta (h - 1 + p_s^0 \theta \alpha) + p_s^0 \theta^2 \alpha \frac{\partial p_s}{\partial \alpha} + (p_s^0 \theta)^2$$

$$= \frac{\partial p_s}{\partial \alpha} \theta (h - 1 + 2p_s^0 \theta \alpha) + (p_s^0 \theta)^2$$

It is easy to show that:

$$\frac{\partial p_s}{\partial \alpha} = \theta (h - 1 + 2p_s^0 \theta \alpha)$$

Using (19) in (18):

$$\frac{\partial^2 W_s^s(*)}{\partial \alpha^2} = \frac{[\theta (h - 1 + 2p_s^0 \theta \alpha)]^2}{2 - (\theta \alpha)^2} + (p_s^0 \theta)^2 > 0$$

For the second part of the Lemma in setting (17) equal to zero and using (3) we find that the minimizer is $\alpha = \frac{2}{2 + \theta}$ and that the minimum $W_s^s(*)|_{\alpha = \alpha^*}$ is always positive. For $h > \frac{2}{2 + \theta}$, $\alpha^* < 1$, and then the conclusion follows.

Lemma 5. To show that the value function $W_f^f(*)$ is strictly increasing, apply the envelope theorem to (5) when $\kappa (\hat{g}, \hat{g}) = (\hat{g} - \hat{g}) - \theta$:

$$\frac{\partial W_f^f}{\partial \alpha} \bigg|_{\hat{g}, \hat{g}} = \frac{\partial \hat{g}}{\partial \alpha} (h - p_f^f + p_f^f F) - \frac{\partial F}{2} (\hat{g} - \hat{g})^2$$

Substituting (16), (8) and $\hat{g} - \hat{g} = \frac{\hat{g} - p_f}{1 + p_f \alpha}$, we obtain the claim:

$$\frac{\partial W_f^f}{\partial \alpha} \bigg|_{\hat{g}, \hat{g}} = \frac{p_f (\hat{g} - p_f)^2}{2 (1 + p_f \alpha)^2} > 0$$

To show the strict concavity, we differentiate (20) obtaining:

$$\frac{\partial^2 W_f^f(*)}{\partial \alpha^2} = \frac{1}{2} \left[ \frac{\partial p_f^f}{\partial \alpha} (h + p_f^f F - p) (1 + p_f^f \alpha)^2 + p_f^f Y \right]$$

with $Y = -2 \frac{\partial p_f^f}{\partial \alpha} (h + p_f^f F - p) (1 - F) (1 + p_f^f \alpha)^2 - 2 (h + p_f^f F - p)^2 (1 + p_f^f \alpha)$.

Therefore, for the claim it is sufficient to show that $\frac{\partial p_f^f}{\partial \alpha} < 0$ and $Y < 0$. To avoid lengthy calculations by hand, we use Mathematica to both derive the equation for $\frac{\partial p_f^f}{\partial \alpha}$ and simplify the expression for $Y$. Applying the
implicit function theorem to (6) when \( \dot{g} = h + p F F \) and avoiding the super-
script for reason of legibility, we obtain:

\[
\frac{\partial p}{\partial \alpha} = \frac{(h - (1 - F)p)(h(p\alpha - 1) + p(1 - F)(3 + p\alpha))}{-2\alpha(1 - F^2)p(\alpha^2p^2 + 3\alpha p + 3) - 4(1 - F)(1 + ah) - 2\alpha^2h^2} \tag{21}
\]

The sign of (21) is strictly negative if \( A = h(p\alpha - 1) + p(1 - F)(3 + p\alpha) > 0 \).
(Recall that \( h - (1 - F)p > 0 \) for an interior solution to (6).) Note that for
\( h = 0 \), the expression reduces to \( p(1 - F)(3 + p\alpha) > 0 \). Then it is enough to
show that \( \frac{\partial A}{\partial h} > 0 \):

\[
\frac{\partial A}{\partial h} = (p F \alpha - 1) + h\alpha \frac{\partial p F}{\partial h} + (1 - F) \left[ \frac{\partial p F}{\partial h} (3 + p F \alpha) + p F \alpha \frac{\partial p F}{\partial h} \right] \tag{22}
\]

Again, to avoid lengthy calculations by hand, we simplify \( \frac{\partial A}{\partial h} \) and we derive
the equation for \( \frac{\partial p F}{\partial h} \) using Wolfram Mathematica. Applying the implicit function
theorem to (6) when \( \dot{g} = h + p F F \) and avoiding the superscript for reason of legibility, we have that:

\[
\frac{\partial p}{\partial h} = \frac{(1 + \alpha p)(1 + h\alpha + Fp\alpha(2 + p\alpha))}{2(1 + h\alpha)(1 - F) + p\alpha(3 + 3\alpha p + p^2 \alpha^2)(1 - F^2) + h^2 \alpha^2} > 0 \tag{23}
\]

Using (23) to simplify (22), we have \( \frac{\partial A}{\partial h} = \frac{N_1}{D_1} \) with:

\[
N_1 = 3\alpha(1 - F^2)p + \alpha(2 - F)h + \alpha^4 Fhp^3 + \alpha^3 Fhp^2 + 5\alpha^2(1 - F)hp
\]
\[+ \alpha^4(1 - F)p^4 + 3\alpha^4(1 - F)p^4 + 2\alpha^3(1 - F)p^3 + 11\alpha^3(1 - F)p^3 + 2\alpha^2(1 - F)p^2
\]
\[+ 13\alpha^2(1 - F)p^2 + \alpha(1 - F)p + 2\alpha^3 h^2 p + 2\alpha^3 hp^3 + 3\alpha^2 hp + 1 - F > 0
\]

and

\[
D_1 = \alpha(1 - F^2)p(\alpha^2 p^2 + 3\alpha p + 3) + 2(1 - F)(ah + 1) + \alpha^2 h^2 > 0
\]

Using (21) to simplify \( Y = -\frac{N_2}{D_2} \), with

\[
N_2 = h - (1 - F)p^2(1 + \alpha p)^2
\]
\[(h^2 \alpha + (F - 1)h(p\alpha - 1) + p\left( (1 - F)(F + 1)(2\alpha^2 p^2 + 4\alpha p + 3) - 2 \right)) > 0
\]
and

\[ D_2 = \alpha \left(1 - F^2\right) p \left(\alpha^2 p^2 + 3\alpha p + 3\right) + 2 \left(1 - F\right) (\alpha h + 1) + \alpha^2 h^2 > 0 \]

For the last part, observe that the welfare for a policy \(\tilde{g} = p^F\) is constant in \(\alpha\).

**Proposition 3.** Showing that for \(F\) small enough the two value functions have at least one intersection is an obvious consequence of the shapes. Then, we need to show that the two value functions have two intersections for \(h\) large enough. To show the second one we can proceed w.l.o.g. assuming \(F = 0\); this is so because the value function under fault-based liability is always decreasing in \(F\) (see footnote 9). As final step, to show the existence of at least one intersection point when \(F = 0\), we only need to show that for \(\alpha = 1\) and for \(h\) large enough the value function of strict liability is greater than the value function of fault-based liability. The welfare under strict liability can be written as:

\[ W^S(*) = 1 + \left(\frac{1}{2} - h\right) - \frac{\left(p^S\right)^2 \left(1 + \theta \alpha\right)^2}{2} - \frac{\left(p^S\right)^2}{2} + h p^S \left(1 + \theta \alpha\right) - p^S \left[\theta \alpha - \theta \alpha \left(1 + \theta \alpha\right)\right] \]

using the first-order condition (2) and rearranging:

\[ W^S(*) = 1 + \left(\frac{1}{2} - h\right) + \left(p^S\right)^2 - \frac{\left(p^S\right)^2 \left(\theta \alpha\right)^2}{2} \]

Note that the probability \(p^S = 1\) if \(h, \alpha, \theta \to 1\). This implies that the welfare \(W^S(*)\) in the limit reaches its maximum value 1. The welfare under fault-based liability can be written as:

\[ W^F(*) = 1 + \left(\frac{1}{2} - h\right) - \frac{p^F p^F + h^2}{2} \left(\frac{1}{1 + p^F}\right) - \frac{\left(p^F\right)^2}{2} \]

Using some simplifications, and avoiding the superscript for reason of legibility, one can show that for \(\alpha = 1\), \(\lim_{h \to 1} W^F(*) = \frac{1 + 4p - p^3 - 2p^2}{2p + 2}\). Therefore, by continuity, to show that when \(h\) and \(\theta\) are large enough, for some \(\alpha\), \(W^S(*) > W^F(*)\) is equivalent to showing

\[ \frac{1 + 4p - p^3 - 2p^2}{2p + 2} < 1 \Leftrightarrow \left(2p^2 - p + 1\right)p + \left(1 - p^3\right) > 0, \]

which is always true because the probability of detection under fault-based liability is increasing in \(h\) (see proof of Lemma 5) and always bounded to 1.

**References**


REP 129 (2) mars-avril 2019


BOWLES S., POLANIA-REYES S. [2012], Economic incentives and social preferences: Substitutes or complements?, *Journal of Economic Literature*, 50(2), 368-425.


DANA J., CAIN D. M., DAWES R. M. [2006], What you don’t know won’t hurt me: Costly (but quiet) exit in dictator games, *Organizational Behavior and Human Decision Processes*, 100(2), 193-201.


*REP* 129 (2) mars-avril 2019


