An Economic Analysis of Criminal Intent and Willful Blindness

Giri Parameswaran
Haverford College
gparames@haverford.edu

PRELIMINARY & INCOMPLETE

October 19, 2021

Abstract

To be guilty of a crime, an agent must not only have committed a wrong act, but must have done so with a guilty mind ('*mens rea*'). Although a crucial feature of criminal law, this requirement for *mens rea* has been largely ignored in the law and economics literature. I present a model of agent behavior and legal outcomes when there is a *mens rea* requirement. In a baseline, where punishments are welfare neutral, I show that, unlike *strict liability*, a *recklessness* standard generates efficient deterrence. However, if punishments generate social benefits (e.g. through fines), then *strict liability* may be preferred. When agents have the opportunity to acquire information about facts relevant to their conduct, a *recklessness* standard coupled with willful blindness doctrine generates both efficient deterrence and efficient information acquisition.

1 Introduction

> *'Actus non facit reum nisi mens sit rea'* — an act does not make [a person] guilty, unless the mind be guilty.

A distinguishing feature of criminal law is that criminal liability requires proof of "an evil-meaning mind with an evil-doing hand".[1] It is not enough that the defendant performed the 'guilty act' (*actus reus*); she must have done so with a 'guilty mind' (*mens rea*). As the U.S. Supreme Court put it: " the contention that an injury can amount to a crime only when inflicted by intention is no provincial or transient notion. It is ... universal and persistent in mature systems of law." By contrast, in most other areas of law, defendants may be held liable even if there was no intent to harm.

Despite being well established, the principle that crimes require intent is not without controversy. First, certain crimes, such as public welfare offenses (e.g. violations of pollution or traffic ordinances), are held to the weaker *strict liability* standard for which intent is not required. Second, many commentators (e.g. see Hamdani, 2007; Finkelstein, 2000) worry that, relative to *strict liability*, a *mens rea* requirement produces inefficient outcomes because, by shielding agents who did not intend the consequences of their actions from culpability, it may encourage 'willful blindness'.

Despite its centrality to criminal law, the economics literature has given little attention to the role of *mens rea* (though see Posner, 1985; Parker, 1993; Hamdani, 2007). In this paper, I present the first comprehensive model of the criminal law that takes seriously the role of *mens rea*. This paper seeks to answer several questions: (i) what is the effect of a *mens rea* requirement on deterrence? (ii) in what contexts (if any) is *strict liability* preferred to a *mens rea* regime? (iii) how does the *mens rea* requirement affect agents' choices to acquire information about the implications of their actions? and (iv) what role can a willful blindness doctrine play?

The standard for establishing intent varies across crimes and across jurisdictions. For concreteness, I use the framework defined in the *Model Penal Code* (Am. Law Inst.), with the understanding that this framework mirrors interpretations applied by common law courts (see Dressler, 2018). The Model Penal Code (MPC) defines five different levels of *mens rea*: *strict liability*, *negligence*, *recklessness*, *knowledge* and *purpose*. The distinction between

[1] See Morrisette v United States, 342 U.S. 246, 251 (1952).

strict liability, *recklessness* and *knowledge* turns on the agent's degree of awareness that material elements of an offence exist or will result from her conduct. *Strict liability* does not require any awareness. *Recklessness* requires that the agent was aware of a 'substantial and unjustifiable risk' of the prohibited outcome, whilst *knowledge* requires her to be 'practically certain'. *Purpose* adds to *knowledge* the additional requirement of motivation: the agent must have had the 'conscious object' of engaging in the prohibited conduct.

For most crimes, the MPC requires that the defendant's conduct be at least reckless (i.e. that the agent acted *recklessly*, *knowingly* or *purposefully*).[2] The MPC explains[3] that the recklessness standard roughly corresponds to the common law requirement of 'general intent'. *Recklessness* is also the appropriate standard for establishing that the agent acted with 'malice'[4], which is a typical statutory and common law *mens rea* requirement (see Dressler, 2018).

Where the law makes the actor *strictly liable*, the MPC requires that this standard apply only to "non-criminal offenses, subject to no severer sanction than a fine, [and] may be employed for regulatory purposes . . . because the condemnatory aspect of criminal conviction or of a correctional sentence is explicitly excluded."[5] The Supreme Court reached a similar conclusion in *Morissette v United States*, where it allowed the strict liability standard for public welfare offences but preserved the *mens rea* requirement for traditional crimes.[6]

Two features of the above framework are worth noting. First, none of the *mens rea* standards turn on the agent's awareness that their conduct was prohibited. Ignorance of the law is no excuse; it is presumed that the agent is aware of her obligations under the law. Instead, intent turns on the agent's awareness of the relevant facts. For example, a person who leaves a restaurant carrying another's umbrella is presumed to know that theft is illegal. Whether he has committed a crime or not depends on his awareness, in the given instance, that the umbrella was, in fact, the property of another. (He might have mistakenly thought that it

[2]§2.02(3) establishes *recklessness* as the default standard for culpability, unless the code provides otherwise. Additionally, *recklessness* is the explicit standard for manslaughter (§210.3), assault (§211.1), reckless endangerment (§211.2), arson (§220.1), criminal mischief (§220.3), burglary resulting in injury (§221.1), amongst others. Where an offence is graded, *recklessness* is typically the standard for the lowest grade offence, and *knowledge* or *purpose* are required for higher grades.

[3]See the explanatory note following MPC §2.02.

[4]An agent acts with 'malice' if either the agent intended to cause a harm or was reckless as to the possibility of causing foreseeable harms. See *R v Cunningham* 3 WLR 76, and *New York Times v Sullivan* 376 U.S. 254.

[5]See explanatory note following §2.05 of MPC.

[6]The Court distinguished conduct *malum prohibitum* ('wrong because it is prohibited') from conduct *malum in se* ('wrong in itself'). The former 'result[s] in no direct or immediate injury to person or property but merely create the danger or probability of it which the law seeks to minimize.' *Malum prohibitum* offences include public welfare offences, and regulatory rules necessary for public health and safety.

was his own, or that the umbrella had been abandoned by its former owner.) In section 4.4, I present an analysis to explain why it may be appropriate to treat mistakes of fact differently from mistakes of law.

Second, the MPC definitions of *negligence* and *recklessness* are distinct only in so far as the former requires an objective test ('would a *reasonable person* be aware of a substantial and unjustifiable risk') while the the latter requires a subjective one (was the actual defendant in question aware of the substantial and unjustifiable risk).[7] In both cases, the standard for what constitutes a 'substantial and unjustifiable risk' is the same. Thus, mirroring the standard for negligence, I will take the standard for *recklessness* to be conduct that, at the margin, results in a net social harm.[8] The standard for recklessness is not simply some arbitrary point along the spectrum from *strict liability* to *knowledge*; it is a specific threshold of awareness that appropriately trades-off the social benefits and costs of the act.

To answer the above questions, I construct a formal model of decision making under a *mens rea* regime. An agent must decide whether to perform some action or not. There are two states of the world; the action has a positive effect on social welfare in one state (the 'good' state) and a deleterious effect in the other (the 'bad' state). The agent has subjective beliefs about the likelihood that the state is 'bad'. (For the purposes of welfare analysis, I assume that the social planner shares the agent's beliefs.) A *mens rea* standard is a threshold belief $\pi \in [0, 1]$, such that the agent who takes the action in the bad state is guilty of a crime only if her subjective belief that the state was bad exceeded this threshold. My framework thus incorporates the range of 'awareness'-type standards, ranging from *strict liability* ($\pi = 0$) to *knowledge* ($\pi \to 1$). If found guilty, the agent incurs a penalty F.

In the baseline model, I assume that punishment is welfare neutral and that the agent has no opportunity to acquire more information about the relevant 'facts' (i.e. the state). In one extension, I allow for punishments that are either socially costly (e.g. due to costs of incarceration) or socially beneficial (e.g. due to revenues from fines). In a separate extension, I allow the agent to learn about the facts by acquiring a noisy signal of the true state. Within this extension, I consider two cases: one where the agent may remain willfully blind with impunity, and one where the court reacts to the agent's willful blindness. Throughout the analysis, I assume that the legal regime (i.e. the *mens rea* standard and the penalty) are chosen by a benevolent policy maker. Important in my analysis is the

[7] The MPC, thus, differs from an older view in the common law that associates recklessness with 'gross negligence' — conduct that is more serious than mere negligence.

[8] Authorities disagree about the level of risk necessary to render an act reckless (Charlow, 1991). My approach mirrors Williams (1953) in that it calls for "an enquiry into the degree of probability of harm and a balancing of this harm against social utility".

somewhat novel assumption (though see Bebchuk and Kaplow, 1992) that different agents have different beliefs about the likelihood that their conduct in the bad state will be detected and prosecuted. Agents are thus distinguished on two dimensions: their belief about the state and their belief about the likelihood of detection. I refer to these pair of beliefs as the agent's type.

The model generates four significant results. First, in the baseline case, I show that the unique socially optimal legal regime involves a true *'recklessness' mens rea* standard, and that such a standard is efficient in that it ensures that (almost) all types of agents make the socially efficient choice about whether to take the action or not. Furthermore, I show that, whenever agents have heterogeneous beliefs about the likelihood of detection, a *strict liability* regime will be inefficient. Thus a *recklessness* standard optimally deters whereas *strict liability* does not.

Second, when punishment itself has welfare consequences, the optimal *mens rea* regime will deviate from the efficient *recklessness* standard. When punishment is socially costly, the optimal legal regime involves a *quasi-recklessness* standard that is more demanding (in the sense of requiring a greater degree of awareness) than the efficient recklessness standard. This is consistent with the argument in Shavell (1985). By contrast, when punishment is socially beneficial, the optimal legal regime will be less demanding than the efficient recklessness standard; in some cases the optimal standard is *strict liability*. Punishment may be socially beneficial, for example if it is in the form of community service or a fine that mitigates the government's need to levy distortionary taxes. This provides a novel explanation for why *strict liability* is typically limited to crimes where the punishment is only a monetary fine.

Third, in an extension where agents have an option to obtain more information about the true state of the world, but where there is no legal consequence for willful blindness, I confirm that a *mens rea* standard generates inefficient information acquisition. However, in contrast to Hamdani (2007), I show that, unless information acquisition is cheap and the signal is very precise, *strict liability* will not be efficient either. In fact, echoing arguments in Parker (1993), —strict liablity causes some agents to inefficiently acquire information, and other agents to inefficiently not acquire information. Instead, the optimal legal regime will be a *quasi-recklessness* standard that lies between *strict liability* and the efficient *recklessness* standard. Naturally, the less demanding standard provides a greater incentive for agents to acquire information.

Finally, I consider a variant extension where courts are able to invoke the doctrine of willful

blindness.[9] Under this doctrine, courts may penalize the willfully blind by imputing to them the beliefs that they would have, had they chosen to acquire information and received 'bad' news. In this case, I show that a true *recklessness* standard is again socially efficient, and results in both (almost) efficient information acquisition and (almost) efficient deterrence. The doctrine of willful blindness remains controversial, and different courts have adopted it to different extents. The analysis in this paper provides a strong endorsement of that doctrine.

Literature Review

The role of intent in the criminal law has received little attention in the law and economics literature, with notable exceptions being Parker (1993), Hamdani (2007) and Posner (1985). Parker (1993) provides an optimal deterrence account of the role of *mens rea*. His account takes it as given that optimal penalties will be sufficiently high as to over-deter some agents from taking an action when it would be efficient to do so. A *mens rea* requirement provides relief, to agents who were sufficiently unaware that their action would create a harm, from these severe (and unanticipated) penalties. The requirement that the agent acted with intent, thus corrects a social harm that stems from imposing large penalties.

Parker (1993) and Hamdani (2007) both also make claims about the role that a *mens rea* standard plays informational acquisition. Parker (1993) argues that *strict liability* tends to cause agents to inefficiently acquire information (even when a social planner would choose not to, given the cost) due to worry about being penalized. Hamdani (2007) argues creates inefficiently weak incentives for information acquisition because it enables agents to remain willfully blind. Both forces are present in my model, and indeed, absent the court penalizing willful blindness, I show that the optimal *mens rea* standard lies somewhere between *strict liability* and the true *recklessness* standard. Hamdani (2007) further claims that *strict liability* is efficient, but this is only true in the special case where all agents share common beliefs about the likelihood of detection. None of these papers deals with optimal policy when a penalty for willful blindness is available.

Posner (1985) makes a different contribution to these, focusing instead on the 'motivational' dimension of *mens rea*. Posner argues that the efficient allocation of goods is best served by having agents engage in voluntary transactions, rather than to enable some agents to acquire

[9] An agent who is willfully blind is often held to have acted 'knowingly' —the court imputes knowledge even if it wasn't actually present.

goods by theft. The *mens rea* requirement in the case of theft, then, is intended to deter agents from coercively acquiring property when voluntary methods are available.

This paper contributes to a broader literature on criminal law and deterrence, beginning with the seminal work of Becker (1968). One strand of this literature studies the relationship between punishment and enforcement in effecting optimal deterrence (e.g see Polinsky and Shavell, 1979, 1984, 1992; Shavell, 1985, 1987; Mungan, 2019). A second strand extends the Beckerian framework to contexts where there is heterogeneity or uncertainty amongst agents about salient feature of the law, for example about the likelihood of apprehension (e.g. see Bebchuk and Kaplow, 1992, 1993; Polinsky and Shavell, 1991). A third strand investigates the incentive to acquire information about whether acts are subject to sanctions (e.g. see Kaplow, 1990*b*; Shavell, 1992). For a good survey of this literature, see Garoupa (1997). However, all of this work is fundamentally situated in a *strict liability* framework. This paper contributes to the literature by expanding the analysis to incorporate a true *mens rea* regime.

This paper also engages with a literature that compares different liability rules (especially strict liability and negligence) in the context of torts (e.g. see Shavell, 1980; Calfee and Craswell, 1984; Craswell and Calfee, 1986; Rubinfeld, 1987; Schäfer and Müller-Langer, 2009). The paper has particular connections with Landes and Posner (1981), which explores the economic implications of intent in the case of intentional torts.

The remainder of this paper is structured as follows: the model is outlined in section 2. Section 3 analyzes legal regimes from the perspective of optimal deterrence. Section 4 studies the incentives for information acquisition. Section 5 briefly explores some extensions, and section 6 concludes.

2 Model

There are two states of the world, $s \in \{0, 1\}$, which encode various relevant facts. An agent must decide whether to take an action A or not N. The agent receives a benefit $B > 0$ from taking the action, independent of the state, but may incur a penalty, as described below. The action produces a net social benefit $S_0 > 0$ in state 0, and a net social harm $S_1 < 0$ in state 1. Individual utility and social welfare are normalized to 0 when the action is not taken.

2.1 Beliefs

The agent is characterized by a pair of beliefs (p, ϕ), where $p \in [0,1]$ is the agent's belief of the likelihood that the state is $s = 1$, and $\phi \in [0,1]$ is her belief that a wrong action in state 1 will be detected.[10]

The agent forms her belief about the state based on her circumstances, so that different agent-types may reasonably hold different beliefs. Suppose that the agent's belief p is a draw from a distribution, with CDF $G(p)$ and PDF $g(p)$.[11]

The agent's belief ϕ, about her probability of being detected, is independent of her belief about the state of the world. I assume that ϕ is a draw from distribution $H(\phi)$, where H is independent of G, and has support on a convex subset of $[0,1]$. The true detection probability is $\hat{\phi}$.

2.2 The Law

The law is characterized by a pair (π, F), where $\pi \in [0,1]$ is the *mens rea* standard and $F \in [0, \overline{F}]$ is the penalty imposed if a crime is detected. According to the *mens rea* standard, an agent performing act A has committed a crime just in case the state is 1 and the agent's belief that it was so is at least π. If $\pi = 0$, the standard is *strict liability*. For $\pi > 0$, the agent is liable only if she foresaw (with sufficient probability) that her conduct was prohibited. As $\pi \to 1$, the standard requires that the agent acted *knowingly*. For simplicity, I assume that the agent perfectly observes the law, and that the agent's subjective belief p is perfectly revealed to the court at trial. In section 5, I consider two different extensions where each of these assumptions are relaxed in turn.

There is a maximum allowable penalty $\overline{F}$, that reflects notions of proportionality (that the punishment must fit the crime), and the eighth amendment's constitutional requirement.

[10] In the baseline model (Section 3), I take p as the agent's final belief. In Section 4, I consider the possibility that the agent may acquire additional information about the state.

[11] I am agnostic as to the source of variation in beliefs. One possibility is that agent-types share a common prior belief $\rho \in (0,1)$ that the state is 1. Each type receives a signal $\sigma \in [0,1]$ drawn from a state-dependent distribution $\hat{G}_s$ with density $\hat{g}_s$. If the distributions $\hat{G}_0$ and $\hat{G}_1$ satisfy the monotone likelihood ratio property so that $\frac{\hat{g}_1(\sigma)}{\hat{g}_0(\sigma)}$ is strictly increasing, then a higher signal will indicate that state 1 is more likely.) The agent's posterior belief is $p = \frac{\rho g_1(\sigma)}{\rho g_1(\sigma) + (1-\rho) g_0(\sigma)}$. By MLRP, there is a one-to-one mapping between the signal σ and the posterior belief p. The distribution over signals, then, induces a distribution over posteriors.

For technical reasons, I assume that $H\left(\frac{B}{\overline{F}} \cdot \frac{S_0 - S_1}{S_0}\right) < 1$, which amounts to requiring that the maximum allowable penalty $\overline{F}$ not be too small.

Punishment itself may generate costs or benefits. The net social cost of punishment is $\chi(F)$, and may depend on the size of the penalty F. The net social cost includes the totality of all direct and indirect costs and benefits accruing to both the agent and the rest of society. It includes, for example, the explicit and opportunity costs of incarceration, and benefits to society in the form of revenues from fines and future harms avoided (if an incarcerated agent would otherwise re-offend). If $\chi(F) = 0$, the legal mechanism is welfare 'neutral'.[12] In the analysis, I allow the net social cost of punishment to potentially be either positive (i.e. truly socially costly) or negative (i.e. socially beneficial). Where punishment is socially beneficial, the benefit is assumed to be not so large as to overwhelm the negative social harm from the action itself.

2.3 Policy-making and Welfare

The legal regime (π, F) is chosen by a socially-minded policy maker. The policy-maker commits to a legal policy prior to both the state and the agent's type being realized, and thus chooses policies to maximize *ex ante* social welfare. To establish a benchmark for welfare analysis, I compare equilibrium outcomes to those that would be chosen by a benevolent social planner who shares the agent's belief about the state p (and thus, has no informational advantage over the agent). Policy-making is efficient if it induces the agent to make the same choices in equilibrium as the social planner would.

3 Deterrence

The analysis unfolds over the next two sections. In this section, I explore the implications of *mens rea* on the optimal deterrence of criminal activity. In the next section, I explore its implications for the agent's incentive to acquire information.

[12] Punishments may be welfare neutral if, for example, the cost to the agent of paying a fine exactly matches the benefit to society from the additional revenue generated by the fine. A fine may, in fact, generate a net social benefit, if it reduces the government's need to otherwise raise revenues through distortionary taxes. (A fine that fails to deter an agent is, by definition, non-distortionary.)

3.1 Agent's Choice

Consider a legal regime (π, F). A type (p, ϕ) agent will choose action A provided that the benefit B from doing so exceeds the expected penalty.

$$B - p\phi \mathbf{1}[p \geq \pi]F > 0$$
$$p < \max\left\{\pi, \frac{B}{\phi F}\right\}$$

The agent will take the action if either she lacks *mens rea* (i.e. $p < \pi$) and thus will not be held culpable, or if her assessment of the expected penalty from taking the action is less than the benefit (i.e. $B - p\phi F > 0$, which implies that $p < \frac{B}{\phi F}$). In the latter case, what matters is not just the probability of doing wrong, but the joint probability of doing wrong and being detected. A more demanding *mens rea* requirement (i.e. one which requires greater awareness before the agent is held culpable) makes the agent more likely to take the action.

3.2 Social Welfare: The First Best

Consider a benevolent social planner who is endowed with the same belief p as the agent. The planner will take the action provided that:

$$pS_1 + (1-p)S_0 > 0$$
$$p < \frac{S_0}{S_0 - S_1} = p^\dagger$$

There is a unique threshold level of awareness $p^\dagger \in (0, 1)$ such that the social planner will take the action unless her subjective belief that the state is 1 exceeds this threshold (i.e. unless $p \geq p^\dagger$).

Note that, as is appropriate for first best analysis, the social planner's choice is independent of the features of the criminal justice system (i.e. of π, F and $\hat{\phi}$). I now ask whether the legal system is an appropriate mechanism to achieve this first best solution.

3.3 Optimal Policy when Punishment is Welfare Neutral ($\chi = 0$)

Suppose that punishment is welfare neutral (i.e. $\chi = 0$). The *ex ante* social welfare associated with a legal regime (π, F) is:

$$W = \int_0^{\pi} \left[(1-p)S_0 + pS_1\right] g(p)dp + \int_{\pi}^{1} \left[(1-p)S_0 + pS_1\right] H\left(\frac{B}{pF}\right) g(p)dp \qquad (1)$$

which is the sum of two terms. The first term is the expected social gain (or loss) accruing from agent-types who take the action because they lack *mens rea*. The second term is the expected social gain accruing from types who are sufficiently aware, but take the action anyway, because they assess the probability of detection to be low. The policy maker chooses π and F to maximize W.

3.3.1 Strict Liability

I begin by considering the special case of strict liability ($\pi = 0$). The policy maker chooses $F^{SL} \in [0, \overline{F}]$ to maximize social welfare W, taking $\pi = 0$ as given.

For a generic penalty F, define $\phi(F) = \frac{B}{p^{\dagger}F}$. By the above analysis, F efficiently deters agents whose belief about the likelihood of being detected is $\phi(F)$; such agents will take the action if and only if $p < p^{\dagger}$, i.e. when it is efficient to do so. Agents with $\phi > \phi(F)$ are over-deterred by the penalty F; they will not take the action for some $p < p^{\dagger}$, where it would be efficient to do so. Conversely, agents with $\phi < \phi(F)$ are under-deterred by the penalty; they will take the action for some $p > p^{\dagger}$, when doing so is inefficient.

Let $\underline{F} = \sup\{F \geq 0 \,|\, H(\phi(F)) = 1\}$. If $F < \underline{F}$, then all agent-types have $\phi < \phi(F)$, and so the law under-deters all agents. $\underline{F}$ is the minimum penalty that causes at least *some* agents to be over-deterred. By assumption, $H(\phi(\overline{F})) < 1$, and so $\underline{F} < \overline{F}$.

Lemma 1. *The optimal strict liability penalty F^{SL} satisfies $H(\phi(F^{SL})) \in (0, 1)$, which implies that $F^{SL} > \underline{F}$.*

When agents have heterogeneous beliefs about the likelihood of detection, it is impossible to efficiently deter all agents under *strict liability*. Lemma 1 requires that the optimal *strict liability* penalty be sufficiently large as to over-deter some agents and under-deter others. Indeed, the optimal *strict liability* penalty optimally trades-off the expected social welfare

loss from agents who are over-deterred against the loss from those who are under-deterred. See the left hand panel in Figure 1.

The requirement that $F^{SL} > \underline{F}$ is intuitive. The policy maker would never choose a penalty lower than $\underline{F}$, since doing so results in all agent-types being under-deterred. By marginally increasing F, the policy maker could decrease the incidence of under-deterrence without causing any agent to be over-deterred. Such a change would clearly improve social welfare.

In fact, the logic continues to hold for any arbitrary *mens rea* requirement that the policy-maker might implement. Let $\pi \in [0,1]$ be an arbitrary *mens rea* standard, and let $F(\pi)$ be the (conditionally) optimal penalty given π.

Lemma 2. *Fixing any arbitrary mens rea standard $\pi \in [0,1]$ (whether optimal or not), the (conditionally) optimal penalty $F(\pi)$ satisfies $F(\pi) > \underline{F}$ (i.e. $H(\phi(F(\pi))) < 1$).*

Under any *mens rea* standard, the policy maker will always choose a penalty large enough that it over-deters at least some measure of agents. If the penalty only under-deters, then there is always a social benefit to increasing the penalty further.

3.3.2 *Mens Rea*

Now, suppose the policy maker is free to choose both the *mens rea* standard π and the penalty F. By the first order conditions, the optimal *mens rea* threshold must satisfy:

$$\frac{\partial W}{\partial \pi} = \left(1 - H\left(\frac{B}{\pi F}\right)\right)\left[(1-\pi)S_0 + \pi S_1\right] g(\pi) = 0 \tag{2}$$

Since, by Lemma 2, $H\left(\frac{B}{\pi F}\right) < 1$, the social welfare maximizing policy is the one that causes $(1-\pi)S_0 + \pi S_1 = 0$, i.e. $\pi^* = \frac{S_0}{S_0 - S_1} = p^\dagger$. The *mens rea* standard that maximizes social welfare coincides with the first-best standard. Moreover, we can naturally interpret this as a *recklessness* standard, since it requires the agent to trade-off the expected social gains and losses from her actions, taking her beliefs as given.

Under the efficient *recklessness* standard, the law does not punish any agent who should efficiently take the action, and so there is no possibility of over-deterrence. And this is true regardless of the size of the penalty. However, the law may still under-deter agents who are sufficiently aware, but believe that they will be detected with low probability. Since the

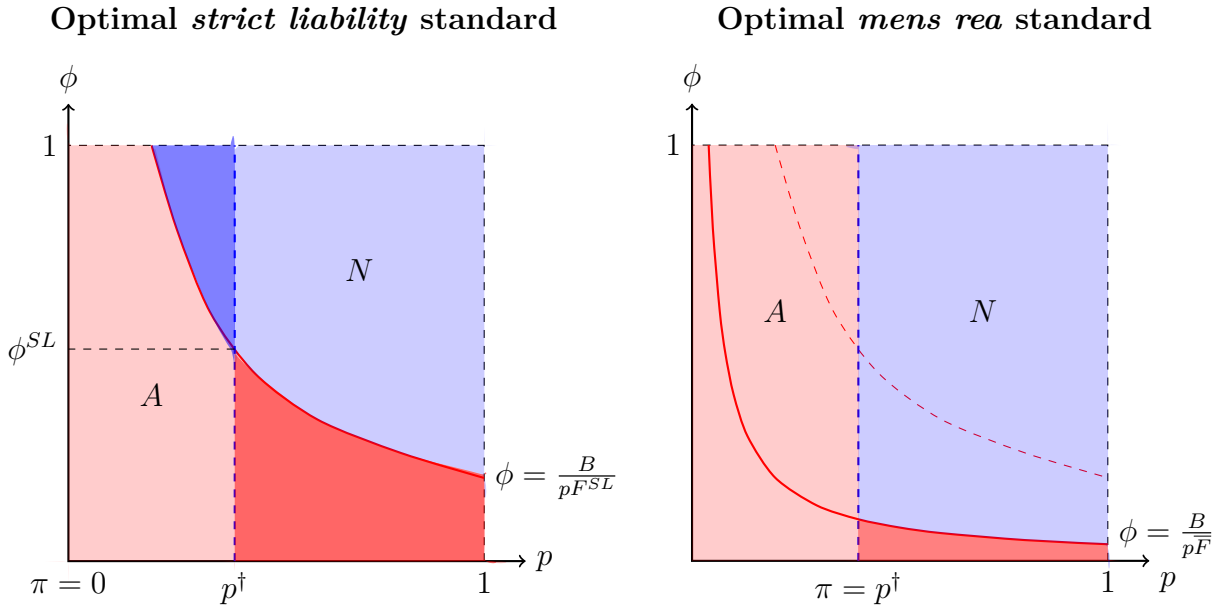

Figure 1: Optimal policy-making under *strict liability*, and with a *mens rea* standard. Agent-types take the action in the red regions, and do not in the blue regions. The darker shaded regions indicate decisions contrary to what the social planner would choose in the first best.

measure of agents who are under-deterred decreases with the size of the penalty, the policy maker will impose the maximum penalty. See the right panel of Figure 1. This implies the following result:

Proposition 1. *Suppose punishment is welfare neutral ($\chi = 0$). The unique optimal policy is characterized by a recklessness mens rea standard $\pi^* = p^\dagger$, and a maximal penalty $F^* = \overline{F}$.*

Proposition 1 establishes one of the main results of this paper. Optimal deterrence recommends a legal regime with *mens rea* standard $\pi^* = p^\dagger$, where agents are only held liable if they are sufficiently aware that their conduct will produce a social harm. Since I associate † with a *recklessness* standard (see footnote **??**), this implies that *reckelessness* is optimal. Even if the reader prefers an alternative interpretation of *recklessness*, it remains normatively true that the optimal *mens rea* standard should be located at $p^\dagger$, and moreover, if *recklessness* remains the default *mens rea* standard, that associating *recklessness* with $p^\dagger$ would improve welfare.

Note well that the mechanism highlighted by Proposition 1 is distinct from previous explanations. Shavell (1985), for example, justifies a *mens rea* standard on the basis that it is socially costly to punish agents who are undetterable because they are unaware that their actions will create social harms. By contrast, in this baseline model, punishment is welfare

neutral. A *recklessness* standard is nevertheless optimal as it enables the policy-maker to optimally deter (or not) the largest possible fraction of agents.

Several points are worth noting. First, though the maximal penalty result in this paper echoes the famous result in Becker (1968), the mechanism is quite different. Becker (1968) consider's a *strict liability* framework in which all agents have homogeneous (and correct) beliefs about the likelihood of detection ϕ, which is under the control of the policy maker. Optimal policy-making requires that the expected penalty ϕF be set at some optimal level. Amongst the many policy pairs that achieve this, the policy maker will favor the one that maximizes the penalty in order to minimize detection costs.[13] The mechanism in this paper is different. Each agent's belief ϕ is fixed, and for simplicity, outside of the policy maker's influence. Hence, increasing F has the pure effect of increasing each agent-type's assessment of the expected penalty. The policy maker sets a maximal policy not to manage detection costs (holding the expected penalty fixed), but to deter agents who are optimistic about the probability of evading detection, and who thus tend to be under-deterred.

Second, the logic of Proposition 1 inverts the insight in Parker (1993). That paper assumes that penalties will be severe (though it provides little justification for why this should be the case), and argues that a generous *mens rea* standard then becomes necessary to provide relief to unsuspecting agents who did not expect that they were doing wrong. The logic in this paper is exactly the opposite. Because the *mens rea* standard provides relief to unsuspecting agents, it becomes appropriate to set large penalties to effectively deter agents who should have known better.

Third, and related to the previous point, although the optimal policy includes a maximal penalty, the *recklessness mens rea* standard would remain optimal even if the policy-maker chose a smaller penalty. To see this, suppose the policy maker implemented a penalty $F = \underline{F} + \varepsilon$ for some small $\varepsilon > 0$. (Recall, the policy maker would never choose $F \leq \underline{F}$.) Such a penalty can hardly be described as excessive: it will likely be 'too low', and thus under-deter most agent-types. Nevertheless, as long as the penalty over-deters a positive measure of agents, the *mens rea* relief is socially desirable. The optimality of the *recklessness* standard is *robust*; it does not depend on the policy-maker also implementing the maximum penalty. This robustness becomes evident in the two extensions explored in Section 5: one extension analyzes optimal policy making when the agent imperfectly observes the *mens rea*

[13] Bebchuk and Kaplow (1992) show that, when agents have heterogeneous beliefs about the detection probability, a maximal penalty is typically not optimal; it will over-deter agents with higher than average beliefs about the probability of detection. My characterization of the optimal strict liability penalty is analogous to the characterization in Bebchuk and Kaplow (1992).

standard; the other analyzes optimal policy making when the agent's belief p is imperfectly revealed to the court. In both cases, the optimal penalty is no longer maximal. However, the optimal *mens rea* threshold continues to be the efficient *recklessness* standard (or an analogue of it).

Fourth, similar to standard models of crime (e.g. see Garoupa, 1997; Polinsky and Shavell, 1984; Shavell, 1985), crime occurs with positive probability under the optimal *mens rea* policy. However, unlike those approaches, with an efficient *recklessness* standard, the agent's action is only a crime if the state is 1 and the taking the action was socially inefficient. There is no 'efficient crime' (see Cooter and Ulen, 2011) in this analysis, as efficient actions are not criminalized. Crime along the equilibrium path occurs precisely in those cases where an agent is optimistic about their chances of evading detection. Criminals are those who think they can get away with it! The same is not true under *strict liability*, where agents are punished *ex post*, even if their action was optimal *ex ante*. Nevertheless, as long as punishment is welfare neutral, this *ex post* punishment does not affect welfare. In the next subsection, I describe how these results would be modified when punishment itself has welfare consequences.

By inspection of Figure 1, it is clear that the optimal *mens rea* policy is more efficient than the optimal *strict liability* policy, in that it under-deters fewer agents and does not over-deter at all. The efficient *mens rea* policy is not truly efficient, because it still under-deters some agents; but this under-deterrence is purely a feature of the constraint that penalties not be too large. The measure of agents who are under-deterred is decreasing in the size of the penalty, and can be made arbitrarily small by making the penalty sufficiently large.

I formalize this insight as follows: I say that a legal regime (π, F) is *ε-efficient* if, for every $\varepsilon > 0$, there exists $F(\varepsilon) > 0$, such that the measure of agent-types making inefficient choices is bounded above by ε whenever $\overline{F} > F(\varepsilon)$.

Proposition 2. *The optimal mens rea policy $(\pi^*, F^*) = (p^\dagger, \overline{F})$ is ε-efficient. By contrast, the optimal strict liability policy $(\pi, F) = (0, F^{SL})$ is not ε-efficient.*

Proposition 2 and Figure 1 make evident the importance of the two sources of heterogeneity in this model. Heterogeneity in agents' beliefs about the likelihood of the bad state is salient to social welfare; the role of the law is to provide incentives for the agent to only take the action when it is efficient to do so. By contrast, heterogeneity in beliefs about the likelihood of detection does not affect social welfare, but nevertheless affects the agent's decision about whether the take the action or not. The optimal *mens rea* standard is *ε-efficient* because it targets the dimension of beliefs that is salient to welfare. By contrast, the penalty alone

partitions the type-space according to the interaction of the two beliefs; it necessarily misclassifies agents based on their detection belief, even though this is inconsequential to welfare.

If H were a degenerate distribution, then the optimal *recklessness* and *strict liability* policies would both efficiently deter agents. This is akin to the equivalence of *negligence* and *strict liability* in the simplest models of torts. However, similar to that literature, we see that the equivalence breaks down as we complicate the model. Unlike the typical findings in that literature (e.g. see Shavell, 1980), it is the *recklessness* standard, and not *strict liability*, that is robust to adding heterogeneity or uncertainty to the model.[14]

3.4 Optimal Policy when Punishment Implicates Welfare

Return to the question of optimal deterrence. Now suppose that the legal mechanism itself imposes costs on society, and so the benefits of incentivizing socially efficient behavior by agents must be weighed against the costs of implementing that mechanism (See Shavell, 1985). Consider a legal regime (π, F) and suppose the social cost of punishment is $\chi(F) \gtrless 0$. The *ex ante* social welfare becomes:

$$W = \int_0^{\pi} \left[(1-p)S_0 + pS_1\right] g(p)dp + \int_{\pi}^{1} \left[(1-p)S_0 + p\left(S_1 - \hat{\phi}\chi(F)\right)\right] H\left(\frac{B}{pF}\right) g(p)dp$$

where $\hat{\phi}$ is the true (exogeneous) probability of detection. This modified expression for social welfare is identical to equation (1), except that it includes a social cost of punishment just in case the state is 1, the agent takes the action, the agent's action was detected, and the agent had sufficient *mens rea*. The social planner chooses both π and F to maximize W. The first order condition with respect to π is:

$$\frac{\partial W}{\partial \pi} = \left(1 - H\left(\frac{B}{\pi F}\right)\right)\left[(1-\pi)S_0 + \pi S_1\right]g(\pi) + H\left(\frac{B}{\pi F}\right)\hat{\phi}\chi(F)g(\pi) = 0$$

which is again analogous to the first order condition in the baseline model, except that it includes a correction for the social cost of punishment. It is easily verified that the efficient *recklessness* standard $p^{\dagger}$ is no longer social welfare maximizing. If $\pi = p^{\dagger}$, then $\frac{\partial W}{\partial \pi} > 0$ when

[14] Shavell (1980) shows that *strict liability* is efficient when actions are taken unilaterally. With bilateral conduct, he additionally shows that a negligence standard (with a defence of contributory negligence) becomes efficient. Rubinfeld (1987) establishes the robustness of comparative negligence in the bilateral setting.

$\chi(F) > 0$ and $\frac{\partial W}{\partial \pi} < 0$ when $\chi(F) < 0$. Assuming the second order conditions are satisfied, we have:

Lemma 3. *The optimal mens rea threshold π^* depends on the penalty F and its social cost $\chi(F)$. Furthermore:*

- *If $\chi > 0$, then $\pi^* > p^\dagger$.*
- *If $\chi < 0$, then $\pi^* < p^\dagger$.*

When punishment is costly, at the margin, the social planner will want to punish fewer agents. In particular, it is no longer socially optimal to punish agents whose conduct is just on the threshold of being socially inefficient. To see this, note that marginally increasing the *mens rea* threshold has two effects: it improves social welfare by decreasing the number of agents who are penalized, and decreases welfare by incentivizing some agents to inefficiently take the action. At the efficient *recklessness* standard $\pi = p^\dagger$, the first effect dominates the second, making the net benefit from increasing π positive. Thus, when punishment is socially costly, the optimal *mens rea* threshold will be more demanding than the efficient recklessness standard $\pi^* > p^\dagger$. Naturally, the opposite would be true if punishment were socially beneficial.

The maximal penalty result also no longer necessarily holds in this setting. Again, for concreteness, take the case of positive punishment costs. Since $\pi^* > p^\dagger$, only agents who inefficiently take the action are punished in equilibrium. Increasing the penalty F has two effects. On the one hand, it deters more agents, and thus improves social welfare. On the other hand, it increases social punishment costs (assuming $\chi'(F) > 0$) which reduces welfare. The overall effect, depends on which effect dominates. If social costs rise more quickly than agents are deterred, then a less than maximal punishment may be optimal. This is consistent with the analysis in Kaplow (1990*a*).

The social costs of punishment cause the optimal *mens rea* standard to deviate from the efficient *recklessness* standard ($\pi^* \neq p^\dagger$). If the optimal *mens rea* standard nevertheless remains interior ($\pi^* \in (0, 1)$), I say the law is characterized by a *quasi-recklessness* standard. *Quasi-reckless* standards are similar to the true *recklessness* standard in so far as the agent is only culpable if she had sufficient awareness of the likelihood of the bad state. But *quasi-reckless* standards need not optimally trade-off the social costs and benefits of taking the action.

Lemma 4. *Suppose punishment is beneficial ($\chi(F) > 0$). The optimal mens rea standard π^* may either be:*

- *A strict liability standard $\pi^* = 0$.*
- *A quasi-recklessness standard $\pi^* \in (0, \pi^\dagger)$*

When punishment is socially beneficially, a *strict liability* regime will always be a candidate optimizer. Marginally increasing the *mens rea* standard from $\pi = 0$ has no effect on deterrence (since agents with beliefs in the neighbourhood of $p = 0$ will take the action under any legal regime) but forgoes the social benefits from punishment. There may also be candidate solutions that are interior (i.e. for which $\pi > 0$). The latter regimes achieve better welfare outcomes in terms of optimal deterrence at the cost of forgoing social benefits from punishment. *Strict liability* will be preferred when the social benefits from punishment are large relative to the social costs of imperfect deterrence.

3.4.1 Justifying Strict Liability

In the analysis so far, I have allowed for the social cost of punishment to take positive or negative values, regardless of the size of the penalty F. But, as previously noted, the sign of χ may itself depend on the size and nature of the penalty F. If F is sufficiently severe as to include incarceration, then punishment may be socially costly ($\chi > 0$) on net. By contrast, if F is small enough that it only involves a fine, the penalty may be socially beneficial. (Shavell (1985) similarly assumes that incarceration involves larger social costs than a mere monetary fine.) Formally, let $\tilde{F} < \overline{F}$ denote the most severe penalty that can be sustained using fines alone. (If the social planner seeks to implement a more severe penalty, the penalty must include non-monetary sancations as well.)

Recall that *strict liability* will be socially optimal provided that the social benefit from imposing a fine (e.g. in the form of lower deadweight losses from taxation) is relatively large, and the welfare losses from inefficient deterrence is relatively small. There are two cases to consider. First, if the optimal *strict liability* penalty can be implemented using a fine alone (i.e $F^{SL} < \tilde{F}$), then a straightforward comparison between the welfare gains and losses will determine which policy is preferred. Second, if the optimal *strict liability* penalty is more severe, then it is infeasible, and to sustain *strict liability*, the policy maker must allow further welfare distortions by choosing a less severe penalty than optimal. As this distortion becomes

larger, the welfare costs from inefficient deterrence will likely overwhelm the welfare benefits from punishment, *ceteris paribus*. Thus, *strict liability* will tend to be preferred when the optimal penalty is naturally relatively low, and the benefits from punishment are relatively high.

Note that the argument here for the strict liability standard is quite different from what is commonly argued. The standard argument is that, if the strict liability standard is to be used at all, it must be limited to minor offenses that require only a small fine, to ensure that agents do not face large penalties for actions for which they are not culpable (in the sense of lacking *mens rea*). My argument starts at a different place. It is precisely because our concern to not over-deter the conduct (whose social benefits in state 0 are larger relative to the social costs in state 1) that we find a low penalty appropriate. But a low penalty allows for the penalty to take the form of a fine, and the levying of such fines can be socially beneficially. This makes strict liability preferable.

4 Ignorance

In this section, I explore the implications of *mens rea* on the agent's incentives to acquire information or, potentially, to remain (willfully) ignorant. I also briefly consider the parallels between my model which is based on uncertainty about facts and one that contemplates uncertainty about the law.

Consider an agent with belief p that the state is 1. The agent can observe a noisy signal $\sigma \in \{0, 1\}$ of the true state. The signal is correct with probability $\gamma \in [0.5, 1]$, so that $\Pr[\sigma = s] = \gamma$. If $\gamma = 0.5$, then the signal is uninformative, whilst if $\gamma = 1$, the signal perfectly reveals the state. For simplicity, in the main analysis, I assume that information acquisition is costless, though this is understood to mean that costs are positive but negligible. Appendix A extends the analysis to the case where agents must pay a cost $C > 0$ to acquire information. As shown in the Appendix, making information acquisition costly is qualitatively similar to making a costless signal less precise. More formally, for a given signal precision γ, there is a less precise signal $\gamma' < \gamma$ such that the agent's decision to acquire costly information with precision γ approximately coincides with the agent's decision to costlessly acquire information with precision γ'. To this end, the assumption that information acquisition is costless should not be seen to be too demanding. My approach is distinct from Kaplow (1990*b*), who focuses on learning about the law, and analyzes the special case of a costly but perfectly informative signal.

Suppose the agent chooses to acquire information. Let $q_0(p,\gamma)$ and $q_1(p,\gamma)$ denote her posterior beliefs that the state is 1, conditional upon receiving signal 0 and 1, respectively. By Bayes Rule, we have:

$$q_0(p,\gamma) = \frac{p(1-\gamma)}{p(1-\gamma)+(1-p)\gamma}$$
$$q_1(p,\gamma) = \frac{p\gamma}{p\gamma+(1-p)(1-\gamma)}$$

Clearly, $q_0(p,\gamma) < p < q_1(p,\gamma)$ for any $\gamma > \frac{1}{2}$. To avoid confusion, I refer to p as the agent's 'prior belief' and q as the agent's 'posterior belief'.

Converting these expressions into log probabilistic odds gives:

$$\ln\left(\frac{q_0}{1-q_0}\right) = \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{\gamma}{1-\gamma}\right)$$
$$\ln\left(\frac{q_1}{1-q_1}\right) = \ln\left(\frac{p}{1-p}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)$$

Information acquisition causes the agent's beliefs, when expressed as log-odds, to either increase (if the signal is $\sigma = 1$) or decrease (if the signal is $\sigma = 0$) by a constant amount $\Delta(\gamma) = \ln\left(\frac{\gamma}{1-\gamma}\right)$. $\Delta(\gamma)$ is positive since $\gamma \geq 0.5$ and it is increasing with γ. Moreover, if $\gamma = 0.5$, there is no shift (since the signal is uninformative), and $\Delta(\gamma) \to \infty$ as $\gamma \to 1$. See Figure 2.

After making her information acquisition decision, and updating beliefs if appropriate, the agent's beliefs are final. Her decision to take the action or not is characterized by the analysis in the previous section. In all the analysis that follows, I assume that the agents optimally choose whether to take the action or not, given their final belief about the state.

4.1 Efficient Information Acquisition

I begin by asking when it is socially efficient to acquire information — i.e. when would a social planner choose to do so? Since information acquisition costs are negligible (but positive), the social planner will acquire information whenever doing so causes *ex ante* social welfare to strictly increase. If, after acquiring information, the planner would not change her action after some signal, then *ex ante* expected social welfare will be the same whether the planner acquires information or not. (This is a straightforward consequence of the law of iterated expectations.) By contrast, if the planner's choice after acquiring information is

signal contingent, then *ex ante* expected social welfare will be strictly higher with information acquisition than without.

Now, the planner's decision to take the action or not depends on whether her final belief is above or below the efficient threshold $p^\dagger$. Thus, the planner will optimally acquire information if either (i) $q_0 < p < p^\dagger < q_1$, or (ii) $q_0 < p^\dagger < p < q_1$. In both cases, the informed planner's decision to take the action is signal-contingent. By contrast, the uninformed planner would take the action in the first case, and not take the action in the latter case.

The social value of this information, ΔW, is the expected gain in social welfare that results from the planner making a more informed decision. It is easily verified that $\Delta W > 0$ in the regions where the planner acquires information.

$$\Delta W = \begin{cases} -p\gamma S_1 - (1-p)(1-\gamma)S_0 & \text{if } p < p^\dagger < q_1(p) \\ p(1-\gamma)S_1 + (1-p)\gamma S_0 & \text{if } q_0(p) < p^\dagger < p \\ 0 & \text{if } q_1(p) < p^\dagger \text{ or } q_0(p) > p^\dagger \end{cases}$$

The planner will acquire information provided that $q_0(p,\gamma) < p^\dagger < q_1(p,\gamma)$. Given the relationship between prior and posterior beliefs implied by Bayes Rule, and writing beliefs in log-odds form, this implies information acquisition when:

$$\ln\left(\frac{S_0}{-S_1}\right) - \ln\left(\frac{\gamma}{1-\gamma}\right) < \ln\left(\frac{p}{1-p}\right) < \ln\left(\frac{S_0}{-S_1}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)$$

where $\ln\left(\frac{S_0}{-S_1}\right) = \ln\left(\frac{p^\dagger}{1-p^\dagger}\right)$ is the log-odds at the efficient threshold. There is an interval of prior beliefs p for which the social planner will acquire information. Written in log-odds form, this interval is centered at $p^\dagger$, and has width that is increasing in the precision of the information source. See Figures 2 and 3.

4.2 Equilibrium Information Acquisition

I now turn to actual information acquisition by agents under the different legal regimes. I begin by considering outcomes under *strict liability*, and then investigate the effect of adding a *mens rea* requirement.

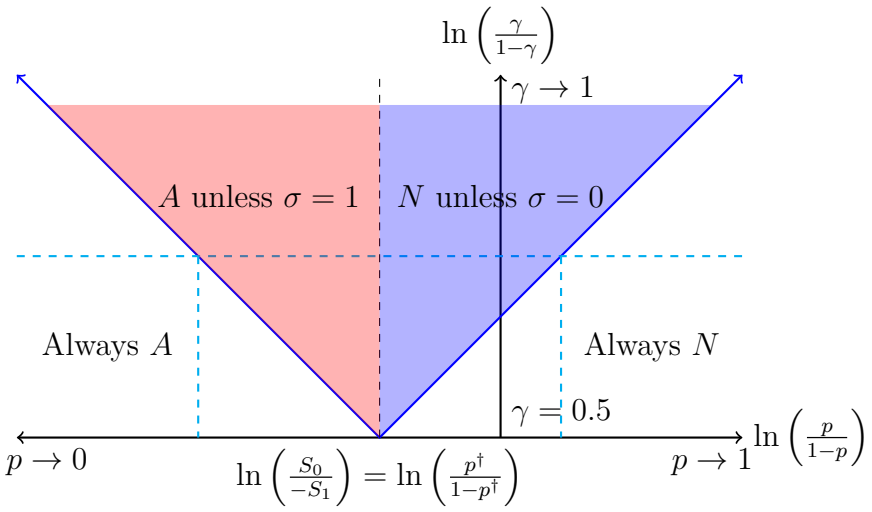

Figure 2: Response to Information. The prior odds are on the horizontal axis. The size of the informational shift is on the vertical axis. The posterior odds are the sum (or difference) of these.

4.2.1 Strict Liability

Suppose the law is characterized by *strict liability* ($\pi = 0$) and let F be an arbitrary penalty. Consider an arbitrary agent with type (p, ϕ). If $\phi < \frac{B}{F}$, then the agent is undeterrable — the agent will take the action regardless of their beliefs about the state. Such agents clearly do not have an incentive to acquire information.

Suppose $\phi > \frac{B}{F}$, so that $\phi F - B > 0$. Following the same logic as in the previous sub-section, the agent will acquire information provided that $p \in \left(q_0\left(\frac{B}{\phi F}\right), q_1\left(\frac{B}{\phi F}\right)\right)$. Equivalently, in log-odds form is:

$$\ln\left(\frac{B}{\phi F - B}\right) - \ln\left(\frac{\gamma}{1-\gamma}\right) < \ln\left(\frac{p}{1-p}\right) < \ln\left(\frac{B}{\phi F - B}\right) + \ln\left(\frac{\gamma}{1-\gamma}\right)$$

Hence, there is an interval of prior beliefs over which each agent will acquire information. Moreover this interval (when expressed over log-odds) has the same width as the social planner's interval. It is easily verified that $\frac{B}{\phi F - B} = \frac{S_0}{-S_1} = \frac{p^\dagger}{1-p^\dagger}$ whenever $\phi = \phi(F)$. Hence, the interval will be centered at $p^\dagger$ if $\phi = \phi(F)$, but not otherwise. Consider two agents: 1 and 2 with $\phi_1 < \phi(F) < \phi_2$. Because agent 1 is more cautious than the social planner, and tends to be over-deterred, there will be a region of priors over which agent 1 investigates before taking the action whereas the social planner would take the action without investigating, and region over which the social planner would investigate but for which agent 1 would not

investigate and does not take the action. The opposite dynamic is true for agent 2. We see this in the left hand panel of Figure 3.

Lemma 5. *A strict liability rule will result in efficient information acquisition by agents who believe the probability of detection is $\phi(F)$. All other agents will inefficiently acquire information, either by acquiring it when they shouldn't, or not acquiring at all (in the case of $\phi < \frac{B}{F}$).*

Lemma 5 stands in direct contract to Hamdani (2007) which claims, drawing on a result in Shavell (1992), that strict liability results in efficient information acquisition. The claim is true for an agent with $\phi = \phi(F)$ — i.e. for whom the penalty F efficiently deters the actions. But the claim is not true for agents more generally. A penalty that inefficiently deters will also create inefficient incentives for information acquisition, as well.

4.2.2 With a *Mens Rea* Standard

I now consider equilibrium information acquisition choices under a true *mens rea* regime. Let (π, F) denote the legal regime. Recall that the law is 'pre-determined' in the sense that the policy-maker chooses π and F prior to the agents' information acquisition decision, and commits to implementing this regime, *ex post*. At this stage, I ignore the possibility that agents are treated differentially based on whether they chose to acquire information or not; all that matters is the agent's final level of awareness p, which may simply be their prior belief. In the following subsection, I consider the possibility that the court may penalize agents who are *willfully blind*. For concreteness, I focus on the baseline model with zero punishment costs, though this should be understood as the limit of a model with positive punishment costs that are made arbitrarily small. (This allows for a simple comparison to the baseline where the optimal *mens rea* threshold was $\pi = p^\dagger$.)

Before characterizing the acquisition decision, I briefly note the importance of the commitment assumption. Notice that, after the acquisition decision, the policy-maker's decision is identical to the one studied in the previous section. Hence, absent commitment, the policy-maker has an optimal strategy, *ex post*, to set a *mens rea* threshold of $\pi = p^\dagger$ and a maximal penalty $F = \overline{F}$. Since this is forseeable, the agent will make her information acquisition decision assuming this is the case. However, although these policies are optimal *ex post*, they may not be optimal *ex ante*, in so far as they may not induce optimal information acquisition.

Consider some arbitrary legal policy (π, F). We can divide the agent's type space into 3 regions according to the nature of their information acquisition decision. If $p < \pi$, then the agent is already immune to prosecution based on her prior beliefs, and so gains nothing (and potentially invites harm) by acquiring information. Agents will not acquire information in this region. Agents with prior beliefs $p \in (\pi, q_1(\pi))$ are not protected by the *mens rea* standard, but will be if they receive the signal $\sigma = 0$ (which causes their posterior belief q_0 to fall below the *mens rea* threshold). These agents will definitely acquire information. Finally, for agents with $p > q_1(\pi)$, even after receiving a signal $\sigma = 0$, their posterior belief will still be above the *mens rea* threshold. These agents effectively face a strict liability regime, and so their information acquisition choice will coincide with the equilibrium choice under strict liability.

Lemma 6. *Fix a level of signal precision $\gamma \in (0.5, 1)$. Under a mens rea regime (π, F), a type (p, ϕ) agent will acquire information provided that either:*

- $p \in (\pi, q_1(\pi))$*, or*
- $p \in \left(\max\left\{\pi, q_0\left(\frac{B}{\phi F}\right)\right\}, \max\left\{\pi, q_1\left(\frac{B}{\phi F}\right)\right\}\right)$

Information acqusition under a *mens rea* standard exhibits an important asymmetry relative to the first best. Under the first best, agents with prior beliefs close enough to $p^\dagger$ (both above and below) will acquire information and change their behavior just in case the signal pushes their posterior across the threshold. Under a *mens rea* standard, only agents whose prior lies above (and sufficiently close to) the threshold will investigate. Agents with prior beliefs close to but below the threshold have no incentive to acquire information, since they are already protected by the *mens rea* standard. Hence, there is less information acquisition, and it is 'asymmetric'.

Figure 3 provides a comparison of the information acquisition decision under *strict liability* versus a true *mens rea* standard, for some fixed signal precision $\gamma > \frac{1}{2}$. The left hand panel represents the acquisition decision under *strict liability*, whilst the right hand panel represents the decision when the *mens rea* standard is $\pi \in \left(q_0(p^\dagger), p^\dagger\right)$. (As I establish below, the optimal *mens rea* threshold will be in this interval, though the qualitative features of the right hand panel would hold for any π.) Of course, *strict liability* is simply a special case of *mens rea* with $\pi = 0$, which is consistent with the Lemma 6, since $q_1(0) = 0$.

A comparison of the left and right hand panels is instructive. First, consistent with the argument in Hamdani (2007), a *mens rea* standard decreases the incentive to acquire information for agents with prior beliefs $p < \pi$. This is a detriment to efficiency in the region

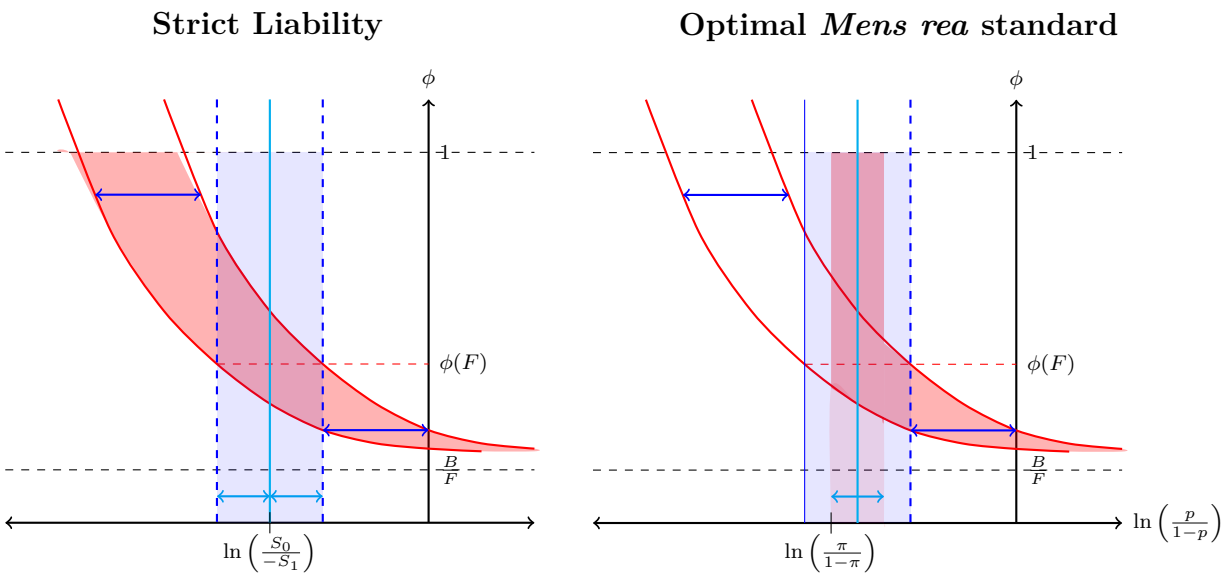

Figure 3: Acquisition of information under different legal regimes. The left hand panel shows behavior under *strict liability* whilst the right hand panel shows the behavior under a genuine *mens rea* regime, with $\pi \in (q_0(p^\dagger), p^\dagger)$. The penalty F is the same in both panels. The dark shaded red region indicates agent types who acquire information and make signal-contingent choices about whether to take the action or not. The light red region indicates agents who acquire information, but choose to take the action regardless of the signal they receive. The blue shaded region indicates the types who should ideally acquire information and make a signal-contingent choice.

$p \in (q_0(p^\dagger), \pi)$, where investigation is socially efficient and some agents would acquire information under *strict liability* but none do under the *mens rea* standard. (Note that this efficiency loss can be made arbitrarily small by taking $\pi \to q_0(p^\dagger)$.) However, it improves efficiency in the region $p < q_0(p^\dagger)$, where information acquisition is inefficient. This is the inefficient information acquisition by cautious agents that Parker (1993) notes in advocating for a *mens rea* standard.

Second, and contrary to Hamdani (2007), a *mens rea* standard increases the incentive to acquire information for agents with prior beliefs $p \in (\pi, q_1(\pi))$. For well chosen π this improves efficiency, since it means there is more investigation in regions of the type-space where the social planner would ideally acquire information. However, there is also a detriment to efficiency in this region that stems from agents with low ϕ (i.e. the light pink region) who will take the action no matter what, but nevertheless acquire information opportunistically in the hope that a good signal ($\sigma = 0$) will make them immune from prosecution. The actions chosen by these agents is the same as it would be under *strict liability*, so outcomes are no worse in that regard, though the information acquisition itself is socially wasteful. (Note that the efficiency loss from opportunistic information acquisition can be made arbitrarily small by taking $F \to \infty$. This has the effect of shifting the curves down, and thus compressing the region where opportunistic behavior occurs.)

Given the above two points, two features are clear. First, a *mens rea* standard (that doesn't treat the willfully blind differently), does not induce efficient information acquisition. Second, for an appropriately chosen legal regime (π, F), a *mens rea* standard can generate more efficient information acquisition than a *strict liability* regime ($\pi = 0$).

4.2.3 Optimal Policy

I now characterize the optimal legal regime (π, F).

Proposition 3. *Fix some signal precision $\gamma \in (\frac{1}{2}, 1)$. The optimal mens rea threshold π^* lies below the socially efficient recklessness threshold $p^\dagger$. Moreover, $\pi^* \in (q_0(p^\dagger), p^\dagger)$.*

The basic intuition for this result is as follows. Information acquisition is most socially beneficial when the prior belief p is very close to the socially efficient threshold $p^\dagger$. Since a *mens rea* regime causes agents with prior beliefs slightly above π to acquire information, the *mens rea* threshold is set slightly below $p^\dagger$ so that the set of agents who investigate straddles $p^\dagger$.

Additionally, while the optimal penalty will quite possibly be maximal (as in the baseline) this result is no longer guaranteed. To understand why, consider the implications of increasing the penalty (which causes both curves in Figure 3) to shift down. For agents with $p > q_1(p^\dagger)$ this improves efficiency by increasing the measure of agents who do not take the action (i.e. those above the higher of the two curves) and decreasing the measure who do (i.e. those below the lower of the two curves). For agents with $p \in (\pi, q_1(\pi))$, increasing F also improves efficiency by decreasing the measure of agents who only investigate opportunistically and are not responsive to the signal (i.e. the light red region). For agents with $p \in (q_1(\pi), q_1(p^\dagger))$ the efficiency implication of increasing F is ambiguous, in so far as the measure of agents in this region who acquire information and behave in a signal-contingent way may change. Whether this results in an increase or decrease in social welfare depends on the shape of the distribution function $H(\phi)$. Thus, for the most part, increasing the penalty is socially beneficial, however, if the final effect is socially costly and sufficiently large, it may cause overall welfare to decrease.

Finally, I note that the amount by which the optimal *mens rea* standard π^* diverges from the socially efficient threshold $p^\dagger$ is increasing in the signal precision γ. Indeed, as the signal becomes perfectly informative, then the optimal *mens rea* standard approaches a strict liability rule.

Corrolary 1. *As $\gamma \to 1$, the optimal mens rea threshold converges to strict liability (i.e. $\pi \to 0$).*

Corollary 1 breathes some life back into the arguments (presented in Hamdani (2007)) that optimal information acquisition requires a strict liability standard. However, one should approach this result with caution. In particular, the result is true in so far as information is both very precise and relatively cheap to acquire. As noted above, even if information is very precise, if it also costly to acquire, then it will be seemingly less valuable to the agent.

4.3 Willful Blindness

Suppose, now, that the court adopts a willful blindness doctrine. Under this doctrine, if an agent believes that there is a high chance that the fact exists, and purposefully fails to investigate in order to confirm the fact, then the agent is held to have known the fact.[15] The doctrine is controversial, and there is disagreement over how it should be applied.

[15]See Global-Tech Appliances Inc. v SEB S.A., 563 U.S. 754,760 (2011).

For the purposes of this model, I operationalize the doctrine in a slightly different way, and then demonstrate the connection to the doctrine as described above. The law is now characterized by a triple: (π_w, π, F), where π^w is the *mens rea* threshold that is applied to agents who did not acquire information, and π is the threshold applied to agents who did. In both cases, if found guilty, the agent faces the same penalty F.

Proposition 4. *Fix some signal precision $\gamma \in (\frac{1}{2}, 1)$. The optimal legal regime with a willful blindness doctrine is characterized by: $\pi_w^* = q_0(p^\dagger)$, $\pi^* = p^\dagger$ and $F^* = \overline{F}$.*

As Proposition 4 shows, when the doctrine of willful blindness is permitted, the baseline results from Proposition 1 are re-established. The law is characterized by a *mens rea* standard located at the optimal threshold for recklessness $p^\dagger$. For the same reasons as in the previous subsection, this ensures that all agents with prior beliefs $p \in (p^\dagger, q_1(p^\dagger))$ acquire information. Additionally, a more exacting standard $\pi_w^* = q_0(p^\dagger) < p^\dagger$ is applied to agents who choose not to acquire information. (Notice that, as the doctrine requires, $\pi_w^* > 0$, so the agent must have a high enough belief that the fact exists before being required to acquire information.) Again, using the same logic as above, agents with prior beliefs $p < q_0(p^\dagger)$ will not acquire information, but agents with $p \in (q_0(p^\dagger), p^\dagger)$ will. Thus, all agents with $p \in (q_0(p^\dagger), q_1(p^\dagger))$ will acquire information, which is precisely the requirement for efficiency. Additionally, since the informed *mens rea* threshold is at the optimal recklessness threshold, a maximal penalty is optimal, for the same reasons as in the previous sections.

Proposition 5. *The optimal legal regime with a willful blindness doctrine is ε-efficient. Formally, for every $\varepsilon > 0$, there exists $F(\varepsilon) > B$ such that in the optimal legal regime $(\pi_w, \pi, F) = (q_0(p^\dagger), p^\dagger, \overline{F})$, the measure of agents making an inefficient choice is less than ε whenever $\overline{F} > F(\varepsilon)$.*

Lemma **??** demonstrates the efficiency of the willful blindess doctrine. Combined with the optimal *recklessness* standard, the law's ability to achieve both efficient deterrence and efficient information acquisition is restricted only by limits on the maximum allowable penalty. As discussed in the previous section, such limits may fail to efficiently deter the most optimistic agents (for whom ϕ is very low). Additionally, where there is an incentive to acquire information, limits on maximal allowable penalties may create incentives for inefficient information acquisition by agents with low ϕ. However, as was discussed in the previous subsection, these incentives disappear as F is made sufficiently large.

Under this conception of the willful blindness doctrine, rather than hold the agent to have beliefs different to the ones they do, the court simply makes the recklessness standard less

demanding. But this has the same effect as to have the court apply a single *mens rea* standard $\pi = p^\dagger$ and to assign to agents who do not acquire information the beliefs that they *would have had* had they acquired information and received the bad signal ($\sigma = 1$). Since $\pi_w = q_0(p^\dagger)$, then applying this procedure, agents with $p < \pi_w$ who do not acquire information will be treated as though their belief was $q_1(p) < p^\dagger$, and so not liable. The opposite is true agents with $p > \pi_w$ who do not acquire information. Such agents will be assigned beliefs $q_1(p) > p^\dagger$ and thus held liable if detected. But this is precisely how the willful blindness doctrine works in practice.

Finally, note that some courts have taken the stronger view that agents who are willfully blind are held to have had 'knowledge' of the facts (i.e. they are assigned a belief $q \to 1$). But this would precisely be the belief that would be assigned if the signal technology were very precisely ($\gamma \to 1$). Hence, if courts tend to apply the willful blindness doctrine only in cases where the agent had a clear opportunity to access credible information, applying this stronger criterion may not be inappropriate.

4.4 Ignorance of Law

Throughout the analysis, I have assumed that the agent understands their responsibilities under the law (since ignorance of the law is no excuse), and that any uncertainty is about the likelihood of circumstances existing that would make their conduct a violation of a known law. Nevertheless, the framework could admit the alternative interpretation, where the consequences or attendant circumstances are known, but the agent is unaware of whether that conduct is lawful ($s = 0$) or not ($s = 1$). The agent may have some prior belief p about the legality of her conduct. If so, there will be some threshold p below which the agent would be sufficiently unaware of the illegality of her conduct that the penalty would not be a deterrent. In such cases, we could make the argument that costly punishment is socially undesirable.

Can this model justify the legal position that ignorance/awareness of fact may be an excuse, but ignorance of law is not? One possibility is to note that laws can be void for vagueness if a layperson cannot easily glean what his responsibilities are under the law. Restricting attention, then, to non-vague laws, we precisely have the case where, when acquiring information about the law, $\gamma \to 1$. Any agent who inquires will, with a high amount of certainty, be able to discern whether their conduct is legal or not. This may justify the use of a strict liability standard when it comes to awareness of law. By contrast, the information/signal technology

may be much less precise when it comes to learning about fact. (E.g. the shopkeeper facing a young looking customer has little ability to discover the buyer's true age beyond checking an identity document, which has a chance of being fake.) This may motivate the optimality of a higher *mens rea* standard for mistakes of fact. A similar point is made in Kaplow (1990*b*).

5 Extensions

In this section, I present two extensions, both of which demonstrate the robustness of the optimal *recklessness* result to perturbations of the model. The first extension contemplates a scenario where the agent only imperfectly observes the *mens rea* threshold π chosen by the policy maker. The second extension considers a scenario where the court imperfectly observes the agent's subjective belief p. In both cases, I show that (an analogue to) the efficient *recklessness* standard continues to be the optimal *mens rea* policy. Additionally, I show that the maximal penalty result is generically not robust.

5.1 Uncertainty about π

Take the baseline model in Section 3.3, in which punishment is welfare neutral and there are no further opportunities to acquire information. Consider a variant model in which the agent imperfectly observes the *mens rea* standard; the agent behaves as though the *mens rea* threshold is $\hat{\pi}$, when in fact it is π. For simplicity, suppose $\hat{\pi} = \pi + z$, where z is a draw from a continuous distribution on $[-\varepsilon, \varepsilon]$ with density $f_\varepsilon(z)$, that is independent of all other variables.

The agent's behavior has the same characterization as in Section 3.1, replacing π with $\hat{\pi}$. *Ex ante* social welfare becomes:

$$W = \int_{-\varepsilon}^{\varepsilon} \left[\int_0^{\pi+z} [(1-p)S_0 + pS_1] g(p) dp + \int_{\pi+z}^1 [(1-p)S_0 + pS_1] H\left(\frac{B}{pF}\right) g(p) dp \right] f(z) dz$$

The policy maker chooses π and F to maximize *ex ante* social welfare, understanding that the agent will actually respond to a perceived threshold $\hat{\pi} = \pi + z$.

Two features of the optimal policy in this perturbed environment are worth noting. First,

the optimal *mens rea* policy is the solution to:

$$\int_{-\varepsilon}^{\varepsilon} [(1-(\pi+z))S_0 + (\pi+z)S_1] \left(1 - H(\frac{B}{(\pi+z)F}\right) g(\pi+z)f(z)dz = 0$$

This condition is analogous to equation (2) above. In the unperturbed model, the optimal *mens rea* threshold had the property that social welfare was insensitive to a marginal increase in π. In the perturbed model, the condition is weakened so that *expected* social welfare is insensitive to the true *mens rea* threshold π, where the expectaton is taken with respect to the distribution of perceived thresholds $f(z)$ and the measure of agents of agents whose conduct is sensitive to the (perceived) threshold. The basic insight is unchanged. The optimal *mens rea* threshold will be *recklessness*-like in that it implements the socially optimal policy, *on average*. Indeed, we have the following result:

Lemma 7. *There exists $\tilde{z}(\pi) \in (-\varepsilon, \varepsilon)$ such that the optimal mens rea threshold π^* satisfies:*

$$\pi^* = p^\dagger - \tilde{z}(\pi^*)$$

Moreover, for all $\pi \in (0,1)$, $\tilde{z}(\pi) \to 0$ as $\varepsilon \to 0$, so $\pi^* \to p^\dagger$ as $\varepsilon \to 0$.

Hence, the optimal *mens rea* threshold will be a perturbed *recklessness* standard, with the size and direction of the perturbation depending on the interaction of the distributions f, g and h. Moreover, as the agent's perception of the *mens rea* standard becomes more accurate, the optimal perturbed standard converges to the efficient *recklessness* standard. The optimality of the *recklessness* standard is robust to agents imperfectly observing the *mens rea* standard.

Second, the optimal penalty will typically no longer be maximal. The intuition is analogous to the case of *strict liability*. Because some agents mis-perceive the true *mens rea* standard, there will be some agents (with $\hat{\pi} < p < p^\dagger$) who should optimally take the action, but will be deterred from doing so if the penalty is too high. The optimal penalty must appropriately trade-off the cost of over-deterring these types against under-detering types with $p > p^\dagger$. Hence, whilst the optimality of the *recklessness* standard is robust to making the *mens rea* threshold imprecisely observed, the maximal penalty result is not robust. With some noise, the optimal penalty will likely be a moderate one.

5.2 Courts Imperfectly Observe the Agent's Belief p

In this extension, I relax the assumption that the agent's subjective belief p is perfectly revealed to the court at trial. Consider the following variant of the baseline model in Section 3.3: If an agent takes the action and is detected in state 1, the court observes their subjective belief to be $\hat{p}$. With probability $r \in [0, 1)$, the court observes the true belief ($\hat{p} = p$), while with probability $1 - r$, it observes some other belief, which is an independent draw from a distribution $D(\hat{p}; p)$, with density $d(\hat{p}; p)$, which may depend on p. This extension is distinct from the previous one in that there is asymmetric information. The the agent knows their belief p and makes their choice accordingly, while the court only observes a noisy signal $\hat{p}$.

«Analysis to follow»

6 Conclusion

In this paper, I have presented the first comprehensive framework that understand the role that a *mens rea* requirement plays in generating optimal deterrence and optimal information acquisition. The analysis reveals several insights about the nature of optimal legal policy.

In the baseline model with no information acquisition and welfare neutral punishment, I showed that the optimal *mens rea* policy coincides with a true *recklessness* standard. This gives weight to the norm in criminal law of requiring that an agent's conduct be at least reckless to be held liable. When punishment itself has welfare consequences, I show that the optimal *mens rea* policy will typically deviate from the efficient *recklessness* standard. Of particular note, when punishments generate net social benefits, then the optimal policy might in fact be *strict liability*. This provides a novel justification of the *strict liability* standard, as a way to generate positive social benefits from the undeterrable otherwise harmful conduct of agents.

I also study the role that a *mens rea* regime may have in incentivizing optimal information acquisition by agents about facts relevant to their conduct. Absent a willful blindness doctrine, I show that the optimal *mens rea* standard will be less demanding than the efficient *recklessness* standard, and approaches *strict liability* as the information sources becomes very precise. However, this policy still produces inefficient outcomes.

By contrast, with a willful blindness doctrine, the efficient *recklessness* standard is efficient, in that it induces (almost) all agents to optimally acquire information acquisition and to

take the socially desirable choice. Thus, my analysis provides a strong endorsement of the willful blindness doctrine.

References

Bebchuk, Lucian Arye and Louis Kaplow. 1992. "Optimal sanctions when individuals are imperfectly informed about the probability of apprehension." *The Journal of Legal Studies* 21(2):365–370.

Bebchuk, Lucian Arye and Louis Kaplow. 1993. "Optimal sanctions and differences in individuals' likelihood of avoiding detection." *International Review of Law and Economics* 13(2):217–224.

Becker, Gary S. 1968. Crime and punishment: An economic approach. In *The economic dimensions of crime*. Springer pp. 13–68.

Calfee, John E and Richard Craswell. 1984. "Some effects of uncertainty on compliance with legal standards." *Va. L. Rev.* 70:965.

Charlow, Robin. 1991. "Wilful ignorance and criminal culpability." *Tex. L. Rev.* 70:1351.

Cooter, Robert and Thomas Ulen. 2011. *Law and economics*. 6th ed. Addison Wesley.

Craswell, Richard and John E Calfee. 1986. "Deterrence and uncertain legal standards." *JL Econ. & Org.* 2:279.

Dressler, Joshua. 2018. *Understanding criminal law*. Vol. 8th edn. Lexis Pub.

Finkelstein, Claire. 2000. "The Inefficiency of Mens Rea." *Calif. L. Rev.* 88:895.

Garoupa, Nuno. 1997. "The theory of optimal law enforcement." *Journal of economic surveys* 11(3):267–295.

Hamdani, Assaf. 2007. "Mens rea and the cost of ignorance." *Va. L. Rev.* 93:415.

Kaplow, Louis. 1990*a*. "A note on the optimal use of nonmonetary sanctions." *Journal of Public Economics* 42:245–247.

Kaplow, Louis. 1990*b*. "Optimal deterrence, uninformed individuals, and acquiring information about whether acts are subject to sanctions." *JL Econ & Org.* 6:93.

Landes, William M and Richard A Posner. 1981. "An economic theory of intentional torts." *International Review of Law and Economics* 1(2):127–154.

Mungan, Murat C. 2019. "Positive Sanctions versus Imprisonment." *George Mason Law & Economics Research Paper* (19-03).

Parker, Jeffrey S. 1993. "The economics of mens rea." *Virginia Law Review* pp. 741–811.

Polinsky, A Mitchell and Steven Shavell. 1979. "The optimal tradeoff between the probability and magnitude of fines." *The American Economic Review* 69(5):880–891.

Polinsky, A Mitchell and Steven Shavell. 1984. "The optimal use of fines and imprisonment." *Journal of Public Economics* 24(1):89–99.

Polinsky, A Mitchell and Steven Shavell. 1991. "A note on optimal fines when wealth varies among individuals." *The American Economic Review* 81(3):618–621.

Polinsky, A Mitchell and Steven Shavell. 1992. "Enforcement costs and the optimal magnitude and probability of fines." *The Journal of Law and Economics* 35(1):133–148.

Posner, Richard A. 1985. "An economic theory of the criminal law." *Colum. L. Rev.* 85:1193.

Rubinfeld, Daniel L. 1987. "The efficiency of comparative negligence." *The Journal of Legal Studies* 16(2):375–394.

Schäfer, Hans-Bernd and Frank Müller-Langer. 2009. Strict liability versus negligence. In *Encyclopedia of Law and Economics*. Edward Elgar Publishing Limited.

Shavell, Steven. 1980. "Strict liability versus negligence." *The Journal of Legal Studies* 9(1):1–25.

Shavell, Steven. 1985. "Criminal law and the optimal use of nonmonetary sanctions as a deterrent." *Columbia Law Review* 85(6):1232–1262.

Shavell, Steven. 1987. "The optimal use of nonmonetary sanctions as a deterrent." *The American Economic Review* pp. 584–592.

Shavell, Steven. 1992. "Liability and the incentive to obtain information about risk." *The Journal of Legal Studies* 21(2):259–270.

Williams, G.L. 1953. *Criminal Law: The General Part*. Stevens.
URL: *https://books.google.com/books?id=x9Q6AQAAIAAJ*

Appendices

A Costly Information Acquisition

Suppose information acquisition now costs $C > 0$. How does this affect the decision to acquire information? Start with the social planner's decision. As I established in the main section, the social planner will acquire information if either $p < p^\dagger < q_1$ or $q_0 < p^\dagger < p$, and the expected social welfare gain ΔW from doing is given by (4.1). Now, the planner will acquire if $\Delta W > C$.

Take the case of $q_0 < p^\dagger < p$. The social planner will acquire if:

$$\begin{aligned} p(1-\gamma)S_1 + (1-p)\gamma S_0 &> C \\ p &< \frac{\gamma S_0 - C}{\gamma S_0 - (1-\gamma)S_1} \\ \frac{p}{1-p} &< \frac{S_0}{-S_1} \cdot \frac{\gamma - \frac{C}{S_0}}{1 - \left(\gamma - \frac{C}{-S_1}\right)} \end{aligned}$$

Similarly, taking the case of $p < p^\dagger < q_1$, the social planner will acquire information provided that:

$$\frac{p}{1-p} > \frac{S_0}{-S_1} \cdot \frac{1 - \left(\gamma - \frac{C}{S_0}\right)}{\gamma - \frac{C}{-S_1}}$$

Two features are worth noting. First, it should be clear that, setting $C = 0$ (and taking logs) we recover the same conditions as in the main analysis. Second, if $S_0 = -S_1$, then for any γ, the condition for acquiring information when costs are $C > 0$ exactly coincides with the condition for acquiring information with a lower signal precision $\gamma' = \gamma - \frac{C}{S_0}$. In this special case, costly information acquisition with a given signal precision γ induces the same behavior as costless information acquisition and a lower signal precision $\gamma' < \gamma$.

What if $S_0 \neq -S_1$? We use the same insight to approximate the social planner's behavior. Let:

$$\gamma' = \gamma - \frac{1}{2}\left[\frac{C}{S_0} + \frac{C}{-S_1}\right] = \gamma - \frac{C}{\overline{S}}$$

where $\overline{S} = \frac{2S_0(-S_1)}{S_0 - S_1}$ is the harmonic mean of S_0 and $-S_1$. When information acquisition is costly, agent behavior is approximated by behavior when information acquisition is costless and with an *effective* signal precision γ' which is lower than the true signal precision γ by

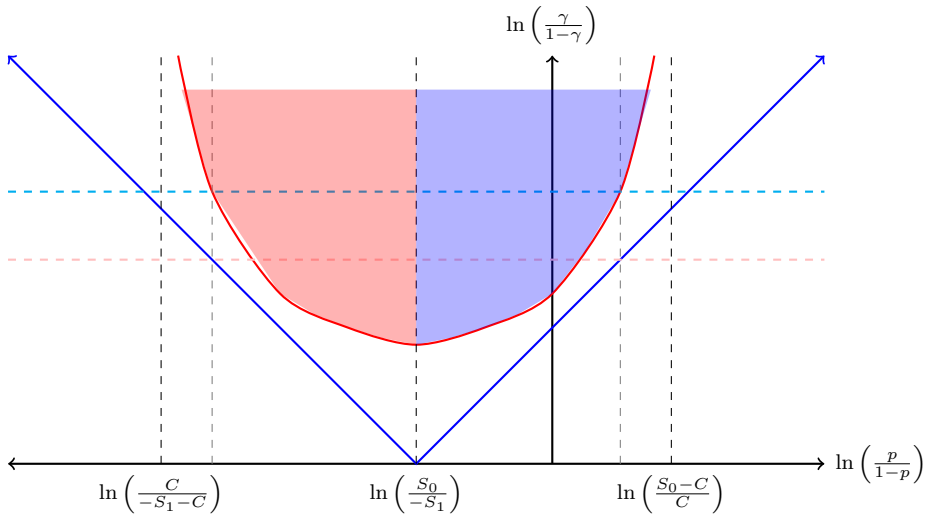

Figure 4: Costly Acquisition of Information. There is a fixed cost $C > 0$ of acquiring information. The social planner will acquire information in the shaded region. The dark blue lines enclose the region where information would be acquired if doing so were costless. Naturally, when it is costly, the planner acquires information over a smaller region. The dashed blue line denotes some generic precision level γ, and the dashed pink line denotes some lower precision level γ'. The planner's decision to acquire costly information at signal precision γ is approximately the same as her decision to acquire costless information at signal precision γ'.

the ratio of information costs to the average (absolute) social gains/losses that result from a change in the agent's behavior. Note that this average $\overline{S}$ is a harmonic mean rather than an arithmetic mean. The larger is $\overline{S}$ relative to C, the more valuable information is, and so the closer is the *effective* signal precision to the true precision. We see this in Figure 4.

The analogous analysis follows for the agent's optimal decision making, given a penalty F and beliefs about detection ϕ. The agent will acquire information provided that:

$$\frac{B}{\phi F - B} \cdot \frac{1 - \left(\gamma - \frac{C}{B}\right)}{\gamma - \frac{C}{\phi F - B}} < \frac{p}{1-p} < \frac{B}{\phi F - B} \cdot \frac{\gamma - \frac{C}{B}}{1 - \left(\gamma - \frac{C}{\phi F - B}\right)}$$

We can approximate the agent's information acquisition behavior by:

$$\frac{B}{\phi F - B} \cdot \frac{1 - \gamma'(\phi)}{\gamma'(\phi)} < \frac{p}{1-p} < \frac{B}{\phi F - B} \cdot \frac{\gamma'(\phi)}{1 - \gamma'(\phi)}$$

where $\gamma'(\phi) = \gamma - \frac{C\phi F}{2B(\phi F - B)} = \gamma - \frac{C}{2B\left(1 - \frac{B}{\phi F}\right)}$. It is as if information is free but less precise. We note, importantly that the amount by which the signal precision is downwardly distorted depends on ϕ. The larger is ϕ, the smaller is this distortion. Hence, for a given cost C and

given true signal precision γ, agents with a low ϕ will acquire information over a smaller range of beliefs p than agents with a high beliefs ϕ.

B Proofs

***Proof of Lemmas 1** and 2.* Take any $\pi \in [0,1)$. Social welfare can be written:

$$W = \int_0^1 \left\{ \int_0^{\max\{\pi, \frac{B}{\phi F}, 1\}} [(1-p)S_0 + pS_1] g(p) dp \right\} h(\phi) d\phi$$

and so a marginal increase in the penalty causes social welfare to change by:

$$\frac{\partial W}{\partial F} = -\int_0^{\max\{\frac{B}{\pi F}, 1\}} \frac{B}{\phi F^2} \left[\left(1 - \frac{B}{\phi F}\right) S_0 + \frac{B}{\phi F} S_1 \right] g\left(\frac{B}{\phi F}\right) h(\phi) d\phi$$

Suppose $F \leq \underline{F}$. Then, $\frac{B}{\phi F} > p^\dagger$ for all ϕ s.t. $h(\phi) > 0$. Hence, $\left[\left(1 - \frac{B}{\phi F}\right) S_0 + \frac{B}{\phi F} S_1 \right] \leq 0$ for all ϕ s.t. $h(\phi) > 0$. But this implies that the integrand is everywhere positive, and so $\frac{\partial W}{\partial F} > 0$. Social welfare would be improved by increasing F. Since this is true whenever $F \leq \underline{F}$, it must be that $F > \underline{F}$ at the optimum. □

***Proof of Proposition 1**.* Recall, $p^\dagger = \frac{S_0}{S_0 - S_1}$ and that expected social welfare $(1-p)S_0 + pS_1$ is positive for all $p < p^\dagger$ and negative for all $p > p^\dagger$.

First, we show that the optimal F^* satisfies $F > \frac{B}{p^\dagger}$. Suppose not, i.e. $F \leq \frac{B}{p^\dagger}$. Recall, by (**??**) that:

$$\frac{\partial W}{\partial F} = -\frac{1}{F} \int_\pi^1 [(1-p)S_0 + pS_1] \left(\frac{B}{pF}\right) h\left(\frac{B}{pF}\right) g(p) dp$$

Notice that the integrand can be positive only if $h(p) > 0$ for some $p < p^\dagger$. But since $p^\dagger < \frac{B}{F}$, $H(\frac{B}{pF}) = 1$ (and hence $h(\frac{B}{pF}) = 0$) for all $p < p^\dagger$. Hence, the integrand that defines $\frac{\partial W}{\partial F}$ is either zero or negative, and so $\frac{\partial W}{\partial F} > 0$. Social welfare is strictly increasing in F whenever $F \leq \frac{B}{p^\dagger}$. Hence $F^* > \frac{B}{p^\dagger}$.

Next, recall:

$$\frac{\partial W}{\partial \pi} = \left(1 - H\left(\frac{B}{\pi F}\right)\right) [(1-\pi)S_0 + \pi S_1]$$

Since optimal $F^* > \frac{B}{p^\dagger}$, we know that W is constant in π for $\pi \in [0, \frac{B}{F^*})$, it is increasing in π for $\pi \in (\frac{B}{F^*}, p^\dagger)$ and decreasing in π for $\pi > p^\dagger$. Clearly it is optimal to choose $\pi = p^\dagger$.

Finally, since $\pi = p^\dagger$, using the same argument as above, it must be that $\frac{\partial W}{\partial F} > 0$, and so W is strictly increasing in F. It is optimal to set $F = \overline{F}$. □

Proof of Proposition 2. Consider the optimal *mens rea* policy $(p^\dagger, \overline{F})$. All agents with $p < p^\dagger$ efficiently take the action. For agents with $p > p^\dagger$, those with $\phi < \frac{B}{p\overline{F}}$ inefficiently take the action, while the remainder efficiently abstain. Thus, the measure of agents who make the wrong choice is $\mu(p^\dagger, \overline{F}) = \int_{p^\dagger}^1 H\left(\frac{B}{p\overline{F}}\right) g(p)dp$.

Clearly $\mu(p^\dagger, \overline{F}) \leq H\left(\frac{B}{p^\dagger \overline{F}}\right)(1 - G(p^\dagger))$. For any $\varepsilon > 0$, let $F(\varepsilon)$ be defined by:

$$F(\varepsilon) = \frac{B}{p^\dagger H^{-1}\left(\frac{\varepsilon}{1-G(p^\dagger)}\right)}$$

It follows, by construction, that $\mu(p^\dagger, \overline{F}) < \varepsilon$ whenever $\overline{F} > F(\varepsilon)$. Hence the optimal *mens rea* policy $(p^\dagger, \overline{F})$ is ε-efficient.

Next, consider the optimal *strict liability* policy $(0, F^{SL})$. Define $\tilde{F} = \inf\{F > 0 \,|\, H(\phi(F)) = 0\}$ if the infimum exists, and $\tilde{F} = \infty$ otherwise. (It follows that $H(\phi(F) \in (0,1)$ iff $F \in (\underline{F}, \tilde{F})$.) Let $\mu_O(F) = \int_0^{p^\dagger}\left(1 - H\left(\frac{B}{pF}\right)\right) g(p)dp$ denote the measure of agents who are over-deterred for a given penalty F. Similarly, let $\mu_U(F) = \int_{p^\dagger}^1 H\left(\frac{B}{pF}\right) g(p)dp$ denote the measure of agents who are under-deterred. Clearly, $\mu(F) = \mu_O(F) + \mu_U(F)$.

Notice that $\mu_O(\underline{F}) = 0$, $\mu_O(F)$ is strictly increasing in F for $F > \underline{F}$, and $\mu_O(F) \to G(p^\dagger)$ as $F \to \infty$. Hence, by the intermediate value theorem, for any $\varepsilon < G(p^\dagger)$, there exists $F_O(\varepsilon)$ s.t. $\mu_O(F) < \varepsilon$ iff $F < F_O(\varepsilon)$. Similarly notice that $\mu_U(0) = 1 - G(p^\dagger)$, $\mu_U(F)$ is strictly decreasing in F on $(0, \tilde{F})$, and $\mu_U(F) \to 0$ as $F \to \tilde{F}$. Hence, there exists $F_U(\varepsilon)$ s.t. $\mu_U(F) < \varepsilon$ iff $F > F_U(\varepsilon)$. Moreover, $F_O(\varepsilon) \to \underline{F}$ and $F_U(\varepsilon) \to \tilde{F}$ as $\varepsilon \to 0$. Then, since $\underline{F} < \tilde{F}$, for ε small enough, $F_0(\varepsilon) < F_U(\varepsilon)$.

Take this case of ε sufficiently small, s.t. $F_0(\varepsilon) < F_U(\varepsilon)$. Then, for any penalty F, either $F < F_U(\varepsilon)$ or $F > F_O(\varepsilon)$, and so either $\mu_O(F) > \varepsilon$ or $\mu_U(F) > \varepsilon$. Hence $\mu(F) = \mu_O(F) + \mu_U(F) \geq \max\{\mu_O(F), \mu_U(F)\} > \varepsilon$ for all F. If so, there is no policy $(0, F)$ that can be ϵ-efficient. □

Proof of Lemmas 3 and 4. Suppose $H\left(\frac{B}{\pi F}\right) < 1$. Then:

$$\frac{\partial W}{\partial \pi} = \left[1 - H\left(\frac{B}{\pi F}\right)\right] g(\pi) \cdot \left((1-\pi)S_0 + \pi S_1 + \frac{H\left(\frac{B}{\pi F}\right)}{1 - H\left(\frac{B}{\pi F}\right)}\hat{\phi}\chi(F)\right)$$

Define:

$$V(\pi, F) = (1-\pi)S_0 + \pi S_1 + \frac{H\left(\frac{B}{\pi F}\right)}{1 - H\left(\frac{B}{\pi F}\right)}\hat{\phi}\chi(F)$$

Note that $sign(V) = sign\left(\frac{\partial W}{\partial \pi}\right)$. Additionally:

$$\frac{\partial V}{\partial \pi} = S_1 - S_0 - \frac{h\left(\frac{B}{\pi F}\right)}{\left[1 - H\left(\frac{B}{\pi F}\right)\right]^2}\left(\frac{B}{\pi^2 F}\right)\hat{\phi}\chi(F)$$

Take some arbitrary F. Suppose $\chi(F) > 0$. Then $\frac{\partial V}{\partial \pi} < 0$, and so the optimal π (which we denote by $\pi(F)$) is characterized by $V(\pi(F), F) = 0$. Since $V(p^\dagger, F) = \frac{H\left(\frac{B}{p^\dagger F}\right)}{1 - H\left(\frac{B}{p^\dagger F}\right)}\hat{\phi}\chi(F) > 0$ for every F, and $\frac{\partial V}{\partial \pi} < 0$, it must be that $\pi(F) > p^\dagger$ for all F. Since $\pi^* = \pi(F^*)$, $\pi^* > p^\dagger$.

Now, suppose instead that $\chi(F) < 0$. Then, clearly W is decreasing for all $\pi < \frac{B}{F}$ and for all $\pi \geq p^\dagger$. Depending on the size of $\chi(F)$ and curvature of H, W may be increasing in π over a sub-interval of $(\frac{B}{F}, p^\dagger)$ (though $V < 0$ in neighbourhoods of $\frac{B}{F}$ and $p^\dagger$). If W is decreasing for all π, then clearly $\pi^* = 0$. If W increases over some interval, then V has an even number of roots in $(\frac{B}{F}, p^\dagger)$. The even numbered roots (ordered from lowest to highest) select local maxima. Let $\pi(F)$ denote the best amongst these local maxima. Necessarily, $V(\pi(F), F) = 0$ and $\pi(F) < p^\dagger$. There are two candidate solutions: $\pi = 0$ and $\pi = \pi(F)$. $\square$

Proof of Proposition 3. Fix some signal precision $\gamma \in (0.5, 1)$, and consider a legal policy (π, F). By Lemma 6, we know that the agent will acquire information if $p \in (\pi, q_1(\pi))$ or if $p \in \left(\max\left\{\pi, q_0\left(\frac{B}{\phi F}\right)\right\}, \max\left\{\pi, q_1\left(\frac{B}{\phi F}\right)\right\}\right)$. Define $\underline{\phi}(p) = \frac{B}{p\gamma F}[(1-p)(1-\gamma) + p\gamma]$ and $\overline{\phi}(p) = \frac{B}{p(1-\gamma)F}[(1-p)\gamma + p(1-\gamma)]$. It follows that, if $p > \pi$ and a type (p, ϕ) agent investigates, then $\phi \in (\underline{\phi}, \overline{\phi})$.

Given a legal regime (π, F), social welfare is:

$$W = \int_0^{\pi} [(1-p)S_0 + pS_1]g(p)dp + \int_{\pi}^{1} [(1-p)S_0 + pS_1]H(\underline{\phi}(p))g(p)dp +$$
$$+ \int_{\pi}^{q_1(\pi)} [(1-p)\gamma S_0 + p(1-\gamma)S_1] \left[1 - H(\underline{\phi}(p)\right] g(p)dp +$$
$$+ \int_{q_1(\pi)}^{1} [(1-p)\gamma S_0 + p(1-\gamma)S_1] \left[H(\overline{\phi}(p) - H(\underline{\phi}(p)\right] g(p)dp$$

The first order condition w.r.t. π gives:

$$\frac{\partial W}{\partial \pi} = \left[1 - H(\underline{\phi}(\pi))\right] [(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1]g(\pi) +$$
$$\left[1 - H(\overline{\phi}(q_1(\pi)))\right] [(1-q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1]g(q_1(\pi)) \cdot \frac{\partial q_1(\pi)}{\partial \pi}$$

I can verify that $\overline{\phi}(q_1(\pi)) > \underline{\phi}(\pi)$. To see this, note by Bayes' rule that, for any p, $q_1(p)[p\gamma + (1-p)(1-\gamma)] = p\gamma$. Then:

$$\frac{\overline{\phi}(q_1(\pi))}{\underline{\phi}(\pi)} = \frac{p\gamma}{q_1(p)(1-\gamma)} \cdot \frac{q_1(p)(1-\gamma) + (1-q_1(p))\gamma}{p\gamma + (1-p)(1-\gamma)} = q_1(p) + \frac{\gamma}{1-\gamma}(1-q_1(p)) > 1$$

since $\gamma > \frac{1}{2}$. It is also easily verified that $\frac{\partial q_1(\pi)}{\partial \pi} > 0$.

I verify, below that $H(\underline{\phi}(\pi)) < 1$ at the optimum. This leaves two possibilities: either $H(\overline{\phi}(q_1(\pi))) = 1$ or $H(\overline{\phi}(q_1(\pi))) < 1$. In the former case, the first order condition simplifies to $(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1 = 0$, which implies that $\pi^* = q_0(p^\dagger)$.

Next, suppose $H(\overline{\phi}(q_1(\pi))) < 1$. Then, the first order condition can be re-written:

$$[(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1] + [(1-q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1] \cdot \underbrace{\left\{ \frac{1 - H(\overline{\phi}(q_1(\pi)))}{1 - H(\underline{\phi}(\pi))} \cdot \frac{g(q_1(\pi))}{g(\pi)} \right\}}_{>0} = 0$$

Hence, to satisfy the first order condition, the expressions $(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1$ and $(1-q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1$ must have opposite sign. But we can verify that $(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1 < 0$ if $(1-q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1 < 0$, and $(1-q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1 > 0$ if $(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1 > 0$. Hence, we need $(1-\pi)(1-\gamma)S_0 + \pi\gamma S_1 < 0$ and

$(1 - q_1(\pi))\gamma S_0 + q_1(\pi)(1-\gamma)S_1 > 0$. But this implies that:

$$\frac{S_0}{-S_1} \cdot \frac{1-\gamma}{\gamma} < \frac{\pi}{1-\pi} < \frac{S_0}{S_1}$$

which, in turn, implies that $\pi \in (q_0(p^\dagger), p^\dagger)$. This proves the main claim.

I must also characterize the optimal penalty. Taking the first order condition w.r.t. F gives:

$$\begin{aligned}\frac{\partial W}{\partial F} = &\int_\pi^1 [(1-p)S_0 + pS_1]h(\underline{\phi}(p))\frac{\partial \underline{\phi}(p)}{\partial F}g(p)dp + \\ &- \int_\pi^{q_1(\pi)} [(1-p)\gamma S_0 + p(1-\gamma)S_1]h(\underline{\phi}(p))\frac{\partial \underline{\phi}(p)}{\partial F}g(p)dp + \\ &+ \int_{q_1(\pi)}^1 [(1-p)\gamma S_0 + p(1-\gamma)S_1]\left[h(\overline{\phi}(p))\frac{\partial \overline{\phi}(p)}{\partial F} - h(\underline{\phi}(p))\frac{\partial \underline{\phi}(p)}{\partial F}\right]g(p)dp\end{aligned}$$

which we can re-write as:

$$\begin{aligned}\frac{\partial W}{\partial F} = &\int_\pi^1 [(1-p)(1-\gamma)S_0 + p_1]h(\underline{\phi}(p))\frac{\partial \underline{\phi}(p)}{\partial F}g(p)dp + \\ &+ \int_{q_1(\pi)}^1 [(1-p)\gamma S_0 + p(1-\gamma)S_1]h(\overline{\phi}(p))\frac{\partial \overline{\phi}(p)}{\partial F}g(p)dp\end{aligned}$$

Now, since $\pi \in (q_0(p^\dagger), p^\dagger)$, it follows that $(1-p)(1-\gamma)S_0 + p\gamma S_1 < 0$ for all $p \geq \pi$. Then since $\frac{\partial \underline{\phi}(p)}{\partial F} < 0$, it follows that the first term is strictly positive. However, the second term is ambiguous. To see this, note again that $\frac{\partial \overline{\phi}(p)}{\partial F} < 0$, but $(1-p)_0 + p(1-\gamma)S_1 > 0$ for $p \in (q_1(\pi), q_1(p^\dagger))$ and $(1-p)_0 + p(1-\gamma)S_1 < 0$ for $p > q_1(p^\dagger)$. Hence, the sign of $\frac{\partial W}{\partial F}$ is probably positive under most parameterizations, but we cannot rule out that it may be negative over some range.

Finally, I show that $H(\underline{\phi}(\pi)) < 1$ in equilibrium. Suppose not. Then, the equilibrium F must be sufficiently small to ensure that $\underline{\phi}(\pi) = \frac{B}{p\gamma F}[(1-\pi)(1-\gamma) + \pi\gamma]$ is above the support of H. But, then F is interior, and so must satisfy the first order conditions. Recall, all components of $\frac{\partial W}{\partial F}$ are necessarily positive except:

$$\int_{q_1(\pi)}^{q} {}_1(p^\dagger)[(1-p)\gamma S_0 + p(1-\gamma)S_1]h(\overline{\phi}(p))\frac{\partial \overline{\phi}(p)}{\partial F}g(p)dp$$

This integral must be negative if the first order conditions are satisfied. But this requires that $h(\overline{\phi}(p)) > 0$ for some subset of $(q_1(\pi), q_1(p^\dagger)$. Then, since ϕ is decreasing in p, it must

be that $H(\overline{\phi}(q_1(p^\dagger)) < 1$. Finally, since $\phi(p^\dagger) < \phi(q_1(p^\dagger)$, and H is weakly increasing, we have that $H(\underline{\phi}(p^\dagger) < 1$. (Almost there, but not quite.)

□

Proof of Corollary 1. Follows immediately from the proof of Proposition 3. As $\gamma \to 1$, for any $\pi \in (0,1)$, $q_1(\pi) \to 1$ and $q_0(\pi) \to 0$. The first order condition then converges to:

$$\frac{\partial W}{\partial \pi} \to \left[1 - H(\underline{\phi}(\pi))\right] \pi S_1 g(\pi) = 0$$

which implies that $\pi^* \to 0$. □

Proof of Proposition 4. Analogous to the proof of Proposition 2. □