RAND Journal of Economics
Vol. 54, No. 2, Summer 2023
pp. 360–384

The wrong kind of information

Aditya Kuvalekar*

João Ramos**

and

Johannes Schneider***

Agents, some with a bias, decide between undertaking a risky project and a safe alternative based on information about the project's efficiency. Only a part of that information is verifiable. Unbiased agents want to undertake only efficient projects, but biased agents want to undertake any project. If the project causes harm, a court examines the verifiable information, forms a belief about the agent's type, and decides the punishment. Tension arises between deterring inefficient projects and a chilling effect on using the unverifiable information. Improving the unverifiable information always increases overall efficiency, but improving the verifiable information may reduce efficiency.

1. Introduction

■ From politicians to doctors to civil servants, examples abound of people avoiding risky but socially efficient decisions for fear of being sued. When faced with a risky decision, individuals rely on information to evaluate their action's costs and benefits. However, if the action results in perverse consequences and litigation ensues, only part of that information is verifiable by a court. For example, policy makers decide on reforms on the basis of experts' reports, which are

* University of Essex.
** USC Marshall.
*** Universidad Carlos III de Madrid; jschneid@eco.uc3m.es.

We are indebted to the editor, Nicola Persico, and three anonymous referees for excellent comments that improved the article substantially. We thank Nageeb Ali, Rosella Argenziano, Heski Bar-Isaac, Dan Bernhardt, Dhruva Bhaskar, Antonio Cabrales, Odilon Câmara, Marco Celentani, Joyee Deb, Siddharth Hari, Chad Kendall, Nenad Kos, Elliot Lipnowski, Antoine Loeper, Anthony Marino, John Matsusaka, Moritz Meyer-Ter-Vehn, Ignacio Ortuño, Harry Pei, Jacopo Perego, Maher Said, and Nico Schutz for helpful comments and discussions. Audiences at various conferences and seminars provided helpful feedback. Aditya Kuvalekar gratefully acknowledges support fromMICIN/AEI/10.13039/501100011033 grant PGC2018-09159-B-I00. Johannes Schneider gratefully acknowledges financial support from the German Research Foundation (DFG) through CRC TR 224 (Project B03), MICIN/AEI/10.13039/501100011033, grants: PID2019-111095RB-I00, PID2020-118022GB-I00, IJC2020-042708, CEX-2021-001181-M, and Comunidad de Madrid, grants EPUC3M11 (V Pricit) and H2019/HUM-5891. Funding for APC: Universidad Carlos III de Madrid (Agreement CRUE-Madroño 2023).

verifiable by outsiders, but also on their own expertise; doctors decide the course of treatment for a patient based on test results and also their examinations. If the policy implemented or the treatment prescribed fails, these agents face the threat of punishment. Anticipating that threat, well-intentioned agents tend to overweight the verifiable information and ignore useful yet unverifiable information. An agent fears that, in litigation, the court—relying only on the verifiable portion of the information—will mistakenly decide that he acted recklessly and for personal benefit rather than in society's interest. In consequence, a well-intentioned agent suffers from a *chilling effect*: He shies away from a socially efficient action for fear of being sued.

Lawmakers, in turn, may be motivated to deter biased agents—agents whose preferences differ from the lawmakers'—from undertaking inefficient projects. However, they must also consider how the threat of punishment affects agents' use of information. They must strike a balance between deterring biased agents from taking socially inefficient actions and encouraging unbiased agents to use both the verifiable and the unverifiable information. Therefore, the optimal design of the law depends on the precision of both the verifiable and the unverifiable information. This dependency raises the questions that motivate our article: How does the precision of information—verifiable and unverifiable—affect the quality of agents' decisions? Can better information exacerbate the chilling effect? If so, can the lawmaker design the law so that the benefits of the superior information outweigh the costs of the chilling effect? Our article studies and answers these questions. We show that in equilibrium agents may make less efficient decisions if the verifiable information becomes more precise; but the efficiency of agents' decisions always improves with greater precision of the unverifiable information.

To capture the above-described environment, we employ the following simple model. There are three players: a designer of the law, a court, and an agent. The agent chooses between undertaking a risky project and taking a safe alternative. The designer wants the agent to undertake the risky project if and only if it is likely to succeed; otherwise the designer prefers the safe alternative. The agent, in turn, can be unbiased or biased toward taking the risky project. The court serves as an exogenous institution that applies the law with the goal of screening out and punishing biased agents. The designer moves first and determines the maximum punishment the court can impose on the agent. Then the agent decides whether to undertake the risky project or to take the safe alternative. To assess the likelihood of success, he relies on two pieces of information about the risky project—one verifiable and one unverifiable. If the agent undertakes the project and it fails, he is taken to court. The court examines the verifiable information and forms a belief about whether the agent is biased. Based on this belief and the limits set by the designer, it decides whether and how much to punish the agent.

To make an efficient decision, an unbiased agent uses both the verifiable and the unverifiable information. At times, the agent should undertake the project even when the verifiable information favors the safe action. A sufficiently informative unverifiable signal favoring the risky project may trump the negative verifiable information. This observation leads to the key tradeoff for the designer: On the one hand, if the threat of punishment is large in case of a failure, the unbiased agent fears that he will be convicted by the court. This induces the chilling effect: the agent ignores the unverifiable information and bases his decision primarily on the verifiable information. On the other hand, if the threat of punishment is low, it will fail to deter the biased agent from undertaking the risky project even when it is inefficient. Our main result shows that the cost of deterrence—the chilling effect—becomes larger as the verifiable information becomes more precise. In fact, it can overpower the benefits of improved information and lead to a reduction in *ex ante* efficiency (Proposition 1). In contrast, the cost of deterrence becomes smaller as the unverifiable information becomes more precise, leading to an unambiguous increase in *ex ante* efficiency (Proposition 2).

The driving force behind our main results is the difference between the two agents' considerations when they contemplate undertaking the project or taking the safe action. To illustrate the mechanism, assume that the agents get punished if the risky project is implemented, it fails, and the verifiable information favors the safe option. The larger the agents' punishment, the more the

safe alternative appeals to both agents. With more precise verifiable information, the project has a higher chance of failure—and hence punishment—when the verifiable information favors the safe alternative. In other words, the fear of punishment increases for both types with more precise verifiable information. In addition to that punishment effect, there is a second effect relevant only for the unbiased type. Because the unbiased agent wishes to undertake the project only when it is likely to succeed, when a more precise verifiable signal favors the safe alternative, the risky project is even less attractive than before. Therefore, there is a stronger chilling effect on the unbiased type: he is more afraid to undertake the project even when it is efficient if that requires going against his verifiable information. As a consequence, deterrence is now accompanied by a stronger chilling effect and, as Proposition 1 shows, efficiency decreases.

The same argument does not hold when the unverifiable information becomes more precise. In this case, deterrence becomes easier. The reason is that it is socially optimal to deter the biased agent from undertaking the project when both the verifiable and the unverifiable information favor not choosing it. More precise unverifiable information leads to a stronger punishment effect here but on the biased type only. The unbiased type would never wish to undertake the project in such situations even without punishment. At the same time, the unbiased agent is now more optimistic about the project's success whenever the unverifiable information recommends undertaking it. Thus, he is more likely to make use of the unverifiable information; the chilling effect declines. Taking both effects together, screening gets easier for the designer and, as Proposition 2 shows, efficiency increases.

Our results provide a cautionary tale in a world in which we observe constant improvements in available information, both verifiable and unverifiable. For example, doctors get better diagnostic tools; politicians get access to more specialized expert reports; civil servants are expected to use newer software to compare prices. Although marginal improvements in unverifiable information are always welfare improving, one should be careful when marginally improving the precision of verifiable information. That is, improvements in technologies such as better diagnostic tools for doctors may simply not substitute for similar improvements acquired through experience. Even worse, such improvements could backfire.

Finally, our results extend to various alternative institutional settings. In Section 4 we discuss a large range of model generalizations to show how the arguments translate to other environments. The key model ingredients driving our result are the following: (i) some but not all agents strive for efficiency; (ii) it is common knowledge that agents possess valuable information beyond what is *ex post* verifiable; (iii) agents are screened *ex post* based on outcomes and the verifiable information. All three elements are present in several real-world settings beyond the legal system. In Section 5 we discuss a variety of settings with and without formal courts to highlight the applicability of our arguments.

□ **Related literature.** At a superficial level, the main takeaway of our article—that superior verifiable information may reduce welfare—is reminiscent of several literatures. Although our results are connected to some of them, we highlight in this section the different economic forces that lead to our results. To this end, we discuss each of the literatures separately.

Exclusion of verifiable information. Federal Rules of Evidence 403 and 404 allow judges to exclude evidence with probative value. Lester et al. (2012) argue that such exclusion may increase welfare. A cost-minimizing fact finder may opt to evaluate evidence with lower statistical power, as it is less costly to do so. Bull and Watson (2019) provide a model of "robust litigation," in which litigants can choose whether to present hard, verifiable information. They show that, depending on the strength of the litigant's private signal relative to that of the hard information, the hard information can be misleading and lead to a loss in welfare.

Unlike Bull and Watson (2019), we abstract from any signaling concerns in the disclosure of the hard information and focus on a setting in which disclosure is mechanical. In this setting,

inefficiency is caused by the agent's hesitation to take an efficient action due to the chance that such an action will lead to (i) harm and (ii) punishment.

In our setting, the defendant's action is publicly observed, and the court's role is to determine whether the intent behind the action was suspect. In line with Sanchirico (2001) and Schrag and Scotchmer (1994), we are interested in how evidence shapes agents' behavior outside the courtroom (and thus how it affects society's welfare). Sanchirico (2001) and Schrag and Scotchmer (1994) study how the law can deter an agent whose preferences do not align with society's. We complement this setting by introducing an unintended side effect of deterrence: the chilling effect on an unbiased agent.[1]

The chilling effect. The chilling effect has been recognized in the literature. An early attempt to capture it formally is in Garoupa (1999). In more recent work, Kaplow (2011, 2017b, 2017a) documents the need to balance deterrence against the chilling effect in a variety of settings.

We build on this literature by taking the chilling effect as the starting point of our analysis. Allowing the punishment scheme to vary with the quality of information, we explore whether the chilling effect can be mitigated through the combination of superior information and an optimal judicial system.

Other side effects of deterrence. A small literature has considered other, orthogonal side effects of deterrence. Stigler (1970) argues that imposing a harsh punishment for minor crimes may erode societies' willingness to punish any crime and suggests intermediate punishment as a remedy. Lagunoff (2001) points out that democratic societies have strategic reasons to limit punishment because an erroneous interpretation of the law by courts may hurt the "wrong" part of the population. Pei and Strulovici (2021) show that severe punishment reduces the number of crimes that witnesses report, thereby reducing the cost of committing a crime. Intermediate punishments can deter some individuals from committing crimes, but those that commit crime are likely to commit several crimes. Unlike these researchers, we concentrate on how the quality of information affects the tradeoff between deterrence and the chilling effect.

Incorporating different types of information. Our main comparative static—increasing the precision of verifiable information can harm welfare—is reminiscent of Morris and Shin (2002) if one views verifiable (unverifiable) information as public (private) information. However, our channel differs from theirs. Coordination motives—the main driver in Morris and Shin (2002)—are entirely absent in our model. To highlight the difference between the environments, consider the setting in which the private information is very precise. Because of coordination motives, small increases in the precision of public information can harm welfare in the setting of Morris and Shin (2002). Players overweigh public information, leading to a welfare reduction if it is sufficiently noisy. An analogous result does not appear in our setting. With precise private information, agents can be screened effectively via the outcome.

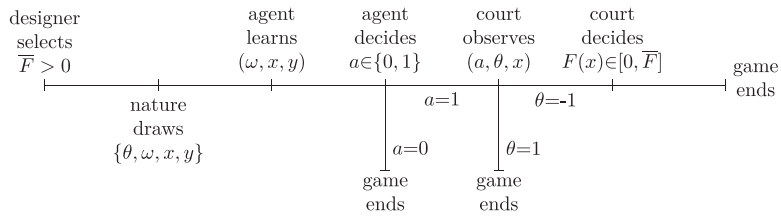
In a principal-agent setting, Blanes i Vidal and Möller (2007) study a problem in which the principal has two pieces of information, only one of which can be shared with the agent before he chooses his effort level. They show that sharing information may harm welfare. Although our setting is outwardly similar to theirs, in ours the agent (not the principal) possesses the information and the punishment is (endogenously) determined *ex ante* to optimally discipline the agent. The economic forces in our environment do not rely on an agent that is suspicious of the selection of the signal received but on an agent afraid of being punished for relying on all of his available information.

Contract theory. The closest articles in the contracting literature are Prendergast (1993) and Prat (2005). In a principal-agent setting, Prendergast (1993) focuses on how to incentivize an agent

[1] All four articles discuss their findings in light of Federal Rule 404, which concerns the exclusion or inclusion of character evidence in the trial. In our environment, all evidence of the agent's character comes from his behavior and not from observable character traits. Thus, the court in our model complies with Federal Rule 404.

FIGURE 1

TIMING OF THE GAME



The designer selects the maximum punishment, $\overline{F}$. The agent observes his type realization, ω, and the realization of the two noisy signals, (x, y), about the risky project's quality. Based on (ω, x, y), the agent decides whether to take the risky action, a=1, or the safe action, a=0.if the agent takes the risky action, the court observes the realized project quality, θ, and the realization of the verifiable signal, x. If the project fails ($\theta = -1$), the court selects a punishment $F(x) \in [0, \overline{F}]$. Then payoffs realize.

to acquire relevant information at a cost. He highlights how the agent may focus more on acquiring information about the principal's prior belief than information about the underlying state of the world.

Prat (2005) argues that the *content of information* leads to qualitatively different effects of increased precision. Whereas information about consequences is beneficial, that about actions is harmful. We view our exercise as complementary to Prat's (2005) and Prendergast's (1993). Whereas they focus on situations in which the underlying information is about different objects (state of the world versus the principal's prior belief in Prendergast (1993) and consequences versus actions in Prat (2005)), both signals in our framework provide information about the same object. We show how the *nature of the information* about the same object—the quality of the project—affects welfare.

2. Model

■ There are three players: an agent ("he"), a court ("it"), and a designer ("she"). The agent decides whether to undertake a risky project that may succeed or fail. The agent is uncertain about the quality of the project and relies on the information available to him when making a decision. If the agent undertakes the risky project, and it fails, the court examines the verifiable part of the agent's information and decides the punishment. The court applies the law, which depends on the designer's initial choice of the maximum punishment. Figure 1 summarizes the basic model structure.

□ **Project quality and information.** The project's quality is either good ($\theta = 1$) or bad ($\theta = -1$). If undertaken, a good project succeeds and a bad project fails.

The *ex ante* probability that the project is good is β. There are two imperfectly informative signals about the project's quality: the verifiable information and the unverifiable information. The verifiable information is a random variable $\boldsymbol{X}$ with realization $x \in \{-1, 1\}$. The precision of the verifiable information is given by $p_x := \mathbb{P}(\boldsymbol{X} = \theta) \in (1/2, 1)$, the probability that the verifiable signal matches the quality of the project. Analogously, the unverifiable information is a random variable $\boldsymbol{Y}$ with realization $y \in \{-1, 1\}$ and precision $p_y := \mathbb{P}(\boldsymbol{Y} = \theta) \in (1/2, 1)$. We summarize the informational environment by $S := (\beta, p_x, p_y)$. The signals $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent conditional on the project's state θ.[2]

[2] To keep the analysis simple, our baseline signal structure is very stylized. Continuous signals are discussed in Online Appendix C; in Online Appendix E we discuss state-dependent levels of precision; and in online Appendix F, we cover conditionally dependent signals.

□ **Designer.** At the beginning of the game, the designer chooses the maximum punishment, $\overline{F}$, the court can inflict on the agent. The designer receives a payoff of 1 from a successful project and a payoff of -1 from a failed project. If no project is undertaken, she receives a payoff of 0.

□ **Agent.** The agent is privately informed about his type ω. He can be unbiased ($\omega = u$) or biased ($\omega = b$). The common prior γ denotes the *ex ante* probability that $\omega = u$. The agent observes the realizations x and y of the two signals and decides whether to act ($a = 1$)—that is, undertake the project—or not ($a = 0$).

The *ex post* payoffs, u^{ω}, of an agent of type ω from his action are given by

$$u^{u}(a, \theta) = a\theta \quad \text{and } u^{b}(a, \theta) = a.$$

An unbiased agent benefits from successful projects but suffers from failed projects; a biased agent benefits whenever he acts. In addition, the court can reduce an agent's utility by punishment F.

□ **Court.** The court observes the agent's action a and the realization of the verifiable information x. It has no access to the unverifiable information. Based on the information, the court applies the law and potentially inflicts punishment $F \in [0, \overline{F}]$. We assume the following on the court's behavior: (i) The court can only punish upon harm, that is, if the risky project fails,[3] and (ii) the court is set to screen agents. That is the court receives a positive payoff F if it inflicts F on a biased agent. It suffers a loss FL if it does so on an unbiased agent. $L > 0$ is a scaling parameter.

□ **Welfare.** Let $a^{\omega}(x, y)$ be the type-ω agent's probability of acting on (x, y), and let $F(x)$ be the court's punishment strategy. We define welfare, $W(\cdot)$, to be the *ex ante* expected utility of the designer. Formally,

$$W(a^{u}(\cdot), a^{b}(\cdot), F(\cdot), \overline{F}; S, \gamma) := \mathbb{E}_{\omega,x,y,\theta}[a^{\omega}(x, y)\theta]$$

We focus on the designer-optimal perfect Bayesian equilibria to which we henceforth simply refer to as "the equilibrium".

□ **On the court's objective function.** Before moving to the analysis, we pause to discuss our assumptions about the court's behavior. Together, they capture the doctrine "actus reus non facit reum nisi mens sit rea" (the act is not culpable unless the mind is guilty). The doctrine requires that a person can be found guilty only if there has been (a) a physical element—an unlawful action—and (b) a mental element—a violation of the standards of care such as negligence or an intention to harm.

In our model, taking a risky action that fails serves as the physical element necessary for conviction. Regarding the mental element, a biased agent intrinsically exercises a lower standard of care as compared to the unbiased agent. Thus, his bias constitutes the guilty mind. Taken together, actus reus and mens rea imply the courts' objective: it wishes to convict the biased agent for undertaking the risky project that fails.

Moreover, we have chosen the concept of *subjective* mens rea in our baseline setting. The guilty mind depends on the inferred preferences of the agent; the law aims to screen agents' types. An alternative interpretation is *objective* mens rea. In that case, the guilty mind depends on the inferred unverifiable information of the agent and the time of decision making; the law aims to screen the agent's information set.

Historically, at least criminal law often relies on subjective mens rea. In tort law cases, mens rea plays a less formal role. However, informally the defendant's (perceived) type remains important in establishing liability (see, e.g., Cane, 2000, for a discussion). Although both types of

[3] This assumption is motivated by realism and does not affect our results. We show this in third subsection of Section 4 which also provides further discussion on this point.

mens rea provide similar results, subjective mens rea appears more appropriate for two reasons. It provides a sharper description of our central tradeoff, and a welfare-maximizing designer prefers it over objective mens rea. We provide further discussion and examples of the settings we have in mind in second subsection of Section 4.

3. Analysis

■ We characterize the equilibria using backward induction. We first analyze the court's best response, then that of the agent. Finally, we determine the designer's optimal choice. Putting everything together, we present our main result: increasing the precision of the unverifiable information, p_y, always improves welfare. In contrast, increasing the precision of the verifiable information, p_x ,may reduce welfare.

□ **Court's best response.** After observing the agent's action, the realization of the state, and the realization of the verifiable signal, the court decides how much to punish the agent. It takes the agent's equilibrium behavior and the maximum punishment set by the designer, $\overline{F}$, as given. By assumption the court can only convict if the project failed—that is, $a = 1$ and $\theta = -1$. Recall that the court receives a payoff of F if it convicts a biased agent and takes a loss of FL if it convicts an unbiased agent. No conviction implies 0 payoff. Thus, when deciding on the punishment, the court uses Bayes' rule to form a belief, denoted by γ_x, about the probability that the agent is unbiased. In calculating γ_x it takes into account both the verifiable information and the agent's equilibrium behavior. Subject to that belief, the court's expected payoff from conviction is

$$(1 - \gamma_x)F - \gamma_x FL.$$

The court chooses to convict the agent only if the above is weakly larger than zero—the payoff of not convicting. Because the interim payoff of convicting is decreasing on γ_x, there is a unique belief that makes the court indifferent between convicting and not: $\overline{\gamma} := 1/(1 + L)$. The court's optimal strategy is a simple cutoff strategy: it inflicts the maximum punishment, $F = \overline{F}$, if $\gamma_x < \overline{\gamma} := 1/(1 + L)$ and no punishment, $F = 0$, if $\gamma_x > \overline{\gamma}$. The court is indifferent between sentences if $\gamma_x = \overline{\gamma}$. The relevant case for our purposes is $\gamma > \overline{\gamma}$ which we assume from now on.

□ **Agent's best response.** The agent observes (x, y) and decides whether to act. If he decides not to act, he receives a payoff of 0. The payoff from acting depends on whether the project ultimately succeeds or fails and on the punishment the court inflicts if it fails. The agent uses the information (x, y) to update his prior belief via Bayes' rule. He forms a posterior belief, β_{xy}, that describes the interim probability that the project is good. Taking the court's decision as given, his interim expected payoff from acting is as follows:

$$\beta_{xy} u^{\omega}(a = 1, \theta = 1) + (1 - \beta_{xy})(u^{\omega}(a = 1, \theta = -1) - F(x))$$

The agent prefers acting only if the above is larger than 0—the payoff from not acting. Because the interim expected payoff is monotonically increasing in β_{xy}, the agent follows a cutoff strategy with type-specific cutoffs

$$\overline{\beta}^{u}(F(x)) := \frac{F(x) + 1}{F(x) + 2} \quad \text{and} \quad \overline{\beta}^{b}(F(x)) := \frac{F(x) - 1}{F(x)}. \tag{1}$$

A type-ω agent strictly prefers to act if $\beta_{xy} > \overline{\beta}^{\omega}(F(x))$, prefers to not act if $\beta_{xy} < \overline{\beta}^{\omega}(F(x))$, and is indifferent between the two if $\beta_{xy} = \overline{\beta}^{\omega}(F(x))$. Notice that $\overline{\beta}^{u}(F(x)) > \overline{\beta}^{b}(F(x))$. Therefore, whenever the unbiased agent weakly prefers acting, the biased agent strictly prefers to act.

We abuse notation slightly and denote by $a^{\omega}(x, y)$ the probability that an agent of type ω acts, taking $F(x)$ and $\overline{F}$ as given.

□ **Designer's best response.** The designer selects the maximum punishment, $\overline{F}$, with the goal of maximizing welfare. Notice that, if $F(x) = 0$, an unbiased agent would act on (x, y) whenever $\beta_{xy} > \frac{1}{2}$. Because the unbiased agent's preferences coincide with those of the society's, his actions too would coincide with the society's preferred actions when $F(x) = 0$. Therefore, we say that it is interim efficient to act on (x, y) if $\beta_{xy} > \frac{1}{2}$.

In the main text, we focus on environments S such that

$$\beta_{xy} \geq 1/2 \Leftrightarrow \max\{x, y\} = 1.$$

That is, we consider the cases in which it is (interim) efficient to act if and only if the agent receives at least one positive signal. We chose this case because it illustrates our main point most clearly.[4]

Basic tradeoff. The two equations in (1) highlight the main tradeoff in designing the maximum punishment level $\overline{F}$. If the expected punishment, $F(x)$, is too low, the biased agent acts even if it is inefficient to do so. If $F(x)$ is too high, the unbiased agent suffers from the *chilling effect*: the fear of being punished if the project fails results in not acting when it is efficient to act.

The optimal punishment scheme balances the deterrence of the biased agent with the encouragement of the unbiased agent.

□ **Optimal punishment scheme.** Having laid out the agent's and the court's incentives, we now proceed to solve for the optimal punishment scheme the designer will set, given the information structure.

First, notice that it is efficient to act when the verifiable information is positive, $x = 1$, regardless of the realization y. Indeed, if both agents act on $x = 1$, the court's posterior probability is equal to the prior γ. Because $\gamma > \overline{\gamma}$, the court never punishes an acting agent when $x = 1$. Conflict arises only when the verifiable information is negative, $x = -1$. Here it is efficient to act on positive unverifiable information, $y = 1$, but efficient to not act on negative unverifiable information, $y = -1$.

Given $x = -1$, the designer wishes to deter the biased agent from acting when $y = -1$ whereas incentivizing the unbiased agent to act when $y = 1$. This leads to the following two natural questions that guide our analysis. Assuming that the agent gets punished when the project fails and $x = -1$, we ask:

1. What is the minimum punishment, F^b, that prevents the biased type from acting when receiving negative unverifiable information, $y = -1$?
2. What is the maximum punishment, F^u, that will allow the unbiased type to act when receiving positive unverifiable information, $y = 1$?

Invoking the agent's best response, we obtain

$$F^u = \frac{2\beta_{-1,1} - 1}{1 - \beta_{-1,1}} \quad \text{and} \quad F^b = \frac{1}{1 - \beta_{-1,-1}}. \tag{2}$$

Whether $F^u > F^b$ or $F^b > F^u$ depends on the information structure, S. As we shall see, the equilibrium has a different structure depending on whether $F^u > F^b$ or vice versa.

To understand the difference between these two cases, it is helpful to consider them separately. To illustrate the underlying intuition, we ignore the court's incentives momentarily and assume that the court punishes with $F(-1) = \overline{F}$. Then, by the definitions of F^b and F^u, an unbiased agent strictly prefers to act on $y = 1$ if $\overline{F} < F^u$, whereas the biased agent strictly prefers to act on $y = -1$ if $\overline{F} < F^b$.

[4] Our results are not specific to this case. For a formal treatment, see Online Appendix D.

TABLE 1 Strategy profiles in the optimal equilibria

When $F^b > F^u$

(a) When $\overline{F} = 0$

(x, y)	a^u	a^b
(–1,–1)	0	1
(–1,1)	1	1

(b) When $\overline{F} = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	0	1

When $F^u > F^b$

(c) When $\overline{F} = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	η^b
(–1,1)	1	1

(d) When $\overline{F} = F^u$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	η^u	1

Case $\mathbf{F^b > F^u}$. If $\overline{F} \leq F^u$, then the unbiased agent acts on $y = 1$ whereas the biased agent acts on all signal realizations. The punishment is too low to achieve any deterrence. Therefore, welfare is constant for all $\overline{F} \in [0, F^u]$, and it is without loss to assume that the designer sets $\overline{F} = 0$; she offers the agent a *universal free pass* (Table 1(a)). If $\overline{F} \in (F^u, F^b)$, then the unbiased agent will prefer to not act on $y = 1$, yet the biased agent will continue to act on $y = -1$. However, then the designer can attain strictly higher welfare by setting $\overline{F} = F^b$: doing so deters the biased agent from acting on $y = -1$, leaving the unbiased agent's behavior unchanged as seen in Table 1(b). Welfare improves.

Hence, the optimal punishment scheme is either a universal free pass ($\overline{F} = 0$) or $\overline{F} = F^b$. The latter deters the biased agent from acting on two negative signals at the cost of fully chilling the unbiased agent's action on $y = 1$. With $\overline{F} = 0$, the court's incentives play no role, whereas with $\overline{F} = F^b$, the court expects only the biased agent to act on $x = -1$. It is optimal to set $F(-1) = \overline{F}$ given the agent's behavior.

Case $\mathbf{F^u > F^b}$. Here it is possible to partially deter the biased agent without imposing a chilling effect on the unbiased agent. By setting $\overline{F} = F^b$, the biased agent is deterred from acting on $y = -1$, whereas the unbiased agent is encouraged to act on $y = 1$. Notice, however, that in this case, the biased agent cannot be fully deterred from acting on $y = -1$ in equilibrium. The reason comes from the court's equilibrium behavior, an issue we have so far ignored. If the biased agent does not act on $y = -1$, and if both types of agents act on $y = 1$, then the court's posterior belief about the agent's type upon seeing a failure and when $x = -1$ is $\gamma > \overline{\gamma}$. The court will not convict the agent. Naturally, the biased agent could exploit the court's behavior and act on $y = -1$. Therefore, in equilibrium, the court must be indifferent about convicting the agent. To make the court indifferent, the biased agent must mix with an interior probability.[5] The equilibrium is summarized in Table 1(c). Alternatively, we can have $\overline{F} = F^u$, which fully deters the biased agent, $a^b(-1, -1) = 0$, but at the cost of a partial chilling effect, $a^u(-1, 1) < 1$, as seen in Table 1(d). Again, the reason for the mixing of the unbiased type on $(-1, 1)$ is to provide incentives to the court for conviction upon failure and $x = -1$.

As we have seen, the ranking of the critical levels F^u and F^b determines how deterrence and the chilling effect pair. For example, if $F^b > F^u$, full deterrence also implies a maximal chilling effect. If $F^u > F^b$, full deterrence is possible at a lower cost. The ranking depends on the information structure S and, in particular, on the levels of precision for the verifiable and the unverifiable information, p_x and p_y. Lemma 1 characterizes the effect of a change in p_x and p_y on the difference $F^b - F^u$. It is at the heart of our main result.

[5] In other equilibria, the biased agent acts with high probability on $y = -1$ and the court strictly prefers conviction. However, these equilibria are not designer optimal.

Lemma 1. The difference between the critical punishment levels, $F^b - F^u$, is continuous in both precision levels, p_x and p_y. The difference is *increasing* in p_x and *decreasing* in p_y.

To understand why the difference is increasing in precision p_x, first note that both F^b and F^u are decreasing in p_x. The likelihood of failing, conditional on $x = -1$, increases with p_x, and thus the agent expects—*ceteris paribus*—a higher punishment. However, the increases in both punishment and probability of failure affect the different types in different ways. Because of the misalignment of preferences between the unbiased and the biased agent, F^u falls faster than F^b. The biased agent suffers only indirectly from the higher failure rate—through the higher punishment (the punishment effect). The unbiased agent also suffers directly—through the failure itself (the outcome effect).

The reason that the difference is decreasing in p_y is more direct. Following an increase of p_y, both outcome and punishment effects encourage the unbiased agent to act on $y = 1$; thus, F^u increases. The punishment effect discourages the biased agent from acting on $y = -1$; thus, F^b decreases.

□ **Signal precision.** Suppose that the verifiable information becomes more precise; that is, p_x increases to some $p'_x > p_x$. By Lemma 1, we could have $F^b < F^u$ at p_x and $F^b > F^u$ at p'_x. Define the following *critical threshold* of information quality.

Definition 1. The precision level p^*_x is a critical threshold of information quality given (p_y, β) if the following two conditions hold:

1. The critical punishment levels are equal: $F^b(p^*_x, p_y, \beta) = F^u(p^*_x, p_y, \beta)$ (see (2)).
2. The informational environment $S = (p^*_x, p_y, \beta)$ is in the interior of environments in which it is efficient to act if and only if $x + y \geq 0$.

Figure 2 sketches these levels for a fixed β in the (p_x, p_y) plane. It is efficient to act if and only if $x + y \geq 0$ inside the shaded region. The thick black line plots the critical information quality, $p^*_x(p_y)$.

With some abuse of notation, let $W^*(p_x)$ [resp. $W^*(p_y)$] denote the (*ex ante*) equilibrium welfare corresponding for some precision level p_x [resp. p_y] in an otherwise-fixed environment (p_y, β, γ) [resp. (p_x, β, γ)].

Proposition 1. An increase in the precision of the verifiable signal can reduce the welfare in non-knife-edge cases. Formally, if p^*_x is a critical threshold, then there is an $\epsilon > 0$ such that $W^*(p_x) > W^*(p'_x)$ whenever $p^*_x - \epsilon < p_x < p^*_x < p'_x < p^*_x + \epsilon$.

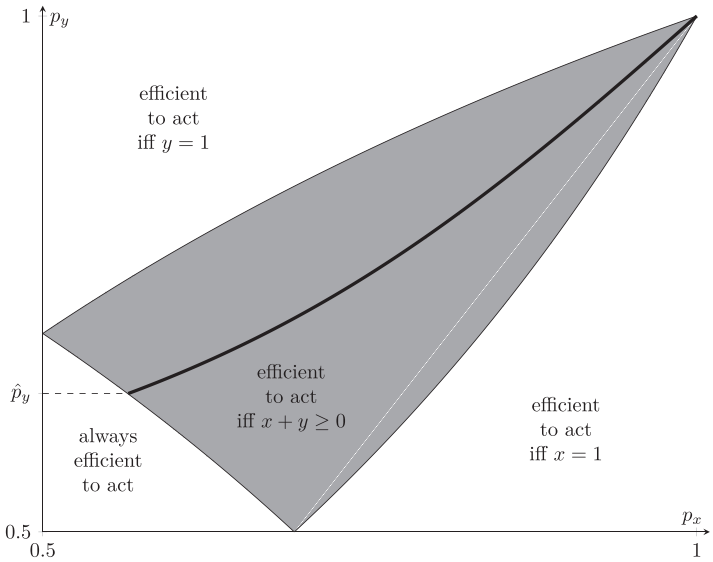
Proposition 1 is driven by the sign change of $F^b - F^u$ around p^*_x as described in Lemma 1. For example, suppose that γ, the prior probability of the agent being unbiased, is high and p_x is slightly below p^*_x. Here, because $F^b < F^u$, the equilibrium is as in Table 1(b). In particular, it is possible to have the biased agent act with probability less than one on $(-1, -1)$ whereas having the unbiased agent act with probability one on $(-1, 1)$.

Increasing p_x to slightly above p^*_x implies that $F^b > F^u$. We can no longer have the biased agent act on $(-1, -1)$ with probability less than one whereas having the unbiased agent act with a positive probability on $(-1, 1)$. Therefore, the designer is left with two options. Either she gives a universal free pass, or she achieves partial deterrence of the biased agent at the cost of a chilling effect on the unbiased agent.

Although the above discussion focuses on the negative effect of improving the verifiable information, there is also a positive effect. An increase in p_x implies that, conditional on $\theta = 1$, realization $x = 1$ occurs more often—an improvement in welfare. Yet the effect of such an

FIGURE 2

CRITICAL VALUES OF THE QUALITY OF INFORMATION



The shaded area is the parameter region (P_X, P_Y) in which it is efficient to act iff $x + y \geq 0$ (our baseline case). On the top left of the shaded region, it is efficient to act iff $y \geq 0$; and on the bottom right iff $x \geq 0$. The bottom left is the area in which even two negative signals cannot overturn the prior β and it is always efficient to act. The thick black line depicts the beliefs at which $F^b = F^u$ and (p_x^*, p_y^*). Changes in p_x represent movements parallel to the $x-$axis; Changes in p_y represent movements parallel to the $y-$axis. Welfare drops for horizontal moves crossing the black line (see Figure 3). In this example, $\beta = 9/13$

improvement is continuous in p_x, whereas the effect due to a regime change, from $F^b < F^u$ to $F^b > F^u$, is discrete. Therefore, welfare declines discretely.

We want to emphasize that Proposition 1 gives a local comparative static. A sufficiently large increase of p_x increases welfare. For example, for a fixed p_y, as $p_x \to 1$, heavily punishing the agent for any failure implies that the project is implemented if, and only if, it is good. In panel (a) of Figure 3, we display welfare as a function of the precision of the verifiable information, p_x. Precisely at the critical threshold of information quality, p_x^*, we see a discontinuous decrease in it as a result of changes in the optimal punishment. For verifiable signals less informative than p_x^*, the optimal punishment can partially deter the biased agent without inducing any chilling effect. In contrast, for verifiable signals more informative than p_x^*, to deter the biased agent implies a complete chilling effect.

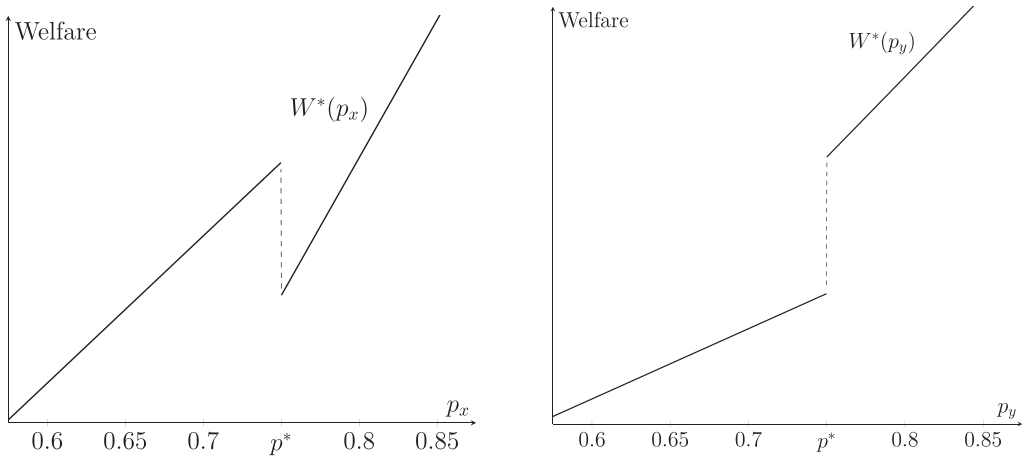
It is tempting to think that the same comparative static holds for the precision of the unverifiable signal. This naive reasoning turns out to be false.

Proposition 2. An increase in the precision of the unverifiable signal always increases welfare. That is, $W^*(p_y') \geq W^*(p_y)\ \forall p_y' > p_y$.

The main difference between p_x and p_y, and the driver of our results, lies in their effect on $F^b - F^u$ as seen in Lemma 1. The main conflict in our environment is that, at times, we want the unbiased agent to decide in favor of undertaking the project despite negative verifiable information, $x = -1$. We want him to rely on the positive unverifiable information he received. However, the associated cost is that—because of the lack of punishment—the biased agent undertakes a project even if $x = y = -1$ (that is, all the information is against it). Increasing the precision of the unverifiable information helps the designer. With increased p_y, $y = 1$ suggests a higher like-

FIGURE 3

WELFARE FOR DIFFERENT PRECISION LEVELS



(a) Welfare in precision of $\mathbf{X}$ (b) Welfare in precision of $\mathbf{Y}$

Welfare is depicted as a function of the precision of the verifiable information, $W^*(p_x)$ (left panel), and as a function of the precision of the unverifiable information, $W^*(p_y)$ (right panel). The discontinuity is at the point at which $F^u = F^b$ such that we switch from the bottom row to the top row of Table 1 (left panel) or from the top row to the bottom row (right panel). In the entire domain of information qualities pictured, acting is efficient iff $x + y \geq 0$. Also, the maximum punishment, $\overline{F}$, is chosen optimally throughout. Parameters: $\overline{\gamma} = 1/2$, $\gamma = 11/20$, $\beta = 9/13$, and $p_y = 3/4$ (left panel), $p_x = 3/4$ (right panel).

lihood of the project quality being good. Therefore, it makes the unbiased agent more confident about undertaking the project when $x = -1$ but $y = 1$. At the same time, it disincentivizes the biased agent from undertaking the project when $x = y = -1$. Thus, welfare increases.

In contrast, increasing the precision of verifiable information disincentivizes both types given negative verifiable information. It increases the threat of punishment, as it indicates a higher chance of failure and thus of punishment if the project is undertaken. In addition, and only for the unbiased type, there is a second deterring force. His payoff is connected to the success of the project directly, and the incentives to undertake the project, given $x = -1$, decline in the precision of $\boldsymbol{X}$, regardless of the punishment. Because of these two effects, increasing the precision of the verifiable signal may lead to lower welfare.

In Panel (b) of Figure 3, we display welfare as a function of the precision of the unverifiable information, p_y. As in panel (a), at the critical threshold of information quality, p_y^*, there is a jump in welfare. However, in contrast to Panel (a), the jump is upward. As the precision of the unverifiable information exceeds the critical threshold, we move from a situation of full deterrence paired with a full chilling effect to the better situation of partial deterrence with no chilling effect.

4. Extensions

■ Given the simplicity of our mechanism, it is natural to wonder about the generality of the forces that lead to the different welfare implications of improving the verifiable and unverifiable information. Do these forces hinge on the intricate details of a formal legal system? Do they rely on the specific assumptions we made about the signal structure?

In this section, we address these questions by highlighting the robustness of our main message to different model specifications. We begin with an abstract principal-agent model. Then we change the objective of the court: what if it aimed to punish the agent for acting against his better

TABLE 2 Strategy profiles in the optimal equilibria

When $F^b > F^u$

(a) When $F(-1) = 0$

(x, y)	a^u	a^b
(–1,–1)	0	1
(–1,1)	1	1

(b) When $F(-1) = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	0	1

When $F^u > F^b$

(c) When $F(-1) = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	1	1

(d) When $F(-1) = F^u$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	1	1

knowledge (objective means rea)? Next, we consider punishment for inaction: what if the court can punish the agent for not acting? We also extend the model to more than two types of agents, and we discuss a richer signal structure.

□ **A contracting model.** We temporarily leave the legal setting with its three players (the designer, the agent, and the court) and consider an abstract model with a principal and an agent, essentially combining the designer and court into a single principal. We consider two versions. First, the principal commits to a punishment rule *ex ante* (the commitment case). That is, the principal designs and commits to a punishment scheme before the agent has acted. Second, the principal decides *ex post* and without constraints on the punishment—that is, after the agent has acted (the *ex post* screening case). We show that our results continue to hold in both cases.

Commitment. There is a principal and an agent. Nature moves first and draws θ, ω, x, y according to a commonly known informational environment, (S, γ). The principal observes x. Thereafter she commits to a punishment $F : \text{supp}(\boldsymbol{X}) \to \mathbb{R}_+$. The agent observes F, his type ω, and (x, y). The agent selects $a \in \{0, 1\}$. If $a\theta = -1$, the agent gets (in addition to his gross payoff $u^\omega(a, \theta)$) punished by $F(x)$.

Given F, the problem of the agent is identical to that in the baseline case. Therefore, F^b and F^u are the same as in the baseline model, which, in turn, implies that Lemma 1 holds. Moreover, it remains without loss to consider as a candidate for an optimal fine $F \in \{F^b, 0\}$ when $F^b > F^u$ and $F \in \{F^b, F^u\}$ when $F^b < F^u$.

The only departure from the baseline model is that the principal need not be indifferent in punishing the agents. The version of Table 1 from the baseline, adapted to the commitment case, is depicted in Table 2.

On the one hand, if $F^u > F^b$, committing to $F(-1) = F^b$ guarantees that the agent takes the (interim) efficient action regardless of his type. On the other hand, if $F^b > F^u$, the payoffs are identical to those in the baseline case. The unbiased agent is too pessimistic about the state and thus is completely chilled by any fine that deters the biased agent. The principal has to decide whether she prefers to prevent the biased agent at the cost of chilling the unbiased agent or to avoid the chilling effect at the expense of no deterrence. Regardless, we lose interim efficiency.

As the environment changes, from $F^b < F^u$ to $F^b > F^u$, welfare suffers a discrete loss because of the inability to implement the interim efficient action, which outweighs the marginal gain from better information. Because Lemma 1 applies, we move—in Table 2—from the bottom row to the top row at p_x^* as p_x increases and from the top row to the bottom row at p_y^* as p_y increases.

In summary, in the commitment case, welfare always improves in the precision of the unverifiable information, whereas it may decline in the precision of the verifiable information, just like in our main results. Moreover, the driving intuition remains the same in this case.

TABLE 3 Strategy profiles in the optimal equilibria

When $F^b > F^u$

(a) When $\mathbb{E}[F] \geq F^b$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	0	0

When $F^u > F^b$

(b) $\mathbb{E}[F] = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	η^b
(–1,1)	1	1

(c) $\mathbb{E}[F] = F^u$

(x, y)	a^u	a^b
(–1,–1)	0	0
(–1,1)	η^u	1

Ex post *screening*. There are a principal and an agent. Nature moves first and draws θ, ω, x, y according to the commonly known informational environment S. Then the agent observes ω, x, y. Thereafter, the agent selects $a \in \{0, 1\}$. If $a\theta = -1$, the principal observes x and can inflict a punishment $F \in \mathbb{R}_+$ on the agent that reduces his gross payoff from acting, $u^\omega(\cdot)$, by F. The principal receives a benefit of F if she punishes a biased agent and suffers a loss LF if she punishes an unbiased agent.

As in the baseline setting, the principal's preferences determine a threshold $\overline{\gamma}$ such that the principal wants to punish if her belief γ_x, conditional on $a\theta = -1$ and realization $\boldsymbol{X} = x$, is less than $\overline{\gamma}$. Similarly, she wants to acquit if $\gamma_x > \overline{\gamma}$.

We make two observations. First, $\gamma_x < \overline{\gamma}$ cannot be an on-path belief. If it were, the principal would select $F(x) = \infty$, which, in turn, would lead to full deterrence. Second, if a free pass is not universally optimal, it cannot be an equilibrium outcome. If it were, the principal's belief would be $\gamma_{-1} < \overline{\gamma}$, which implies punishment—a contradiction.

The two observations imply that the principal either implements full deterrence or has to be indifferent in any equilibrium. If $F^b > F^u$, full deterrence is the only option, whereas when $F^u > F^b$, full deterrence cannot be optimal. Consequently, the equivalent to Table 1 for this case is Table 3.

In Table 3, η^b and η^u are such that the principal is indifferent. Being indifferent, the principal can select any punishment scheme. However, to make the agent indifferent as well, we need it to be true that the expected punishment $\mathbb{E}[F] = F^b$ or $\mathbb{E}[F] = F^u$.

The *ex post* screening case strengthens our results. In Table 3, welfare is strictly lower in the top row compared to the baseline. The reason is the following: In this environment, we lack a designer to optimally limit the punishment *ex ante*. The bottom row, however, yields the same welfare as in the baseline case. We see that the designer's ability to limit the punishment is beneficial, particularly when $F^b > F^u$, which occurs when the verifiable signal is precise.

Relationship to the baseline. The commitment case corresponds to strict liability in the legal setting. In certain situations—for example, if the realization of the verifiable information is some specific x—an action makes the agent liable "per se." That is, the court does not form an opinion about the agent's type but punishes based only on a, θ, x. Such a case is directly captured by the commitment model.

The *ex post* screening case encompasses scenarios in which the magnitude of the punishment is exogenous; for example, the agent gets fired from his job. In some of the examples we discuss in Section 5, such an exogenous punishment appears appropriate.

□ **Objective mens rea.** In this part, we address the question of how relevant the assumption of subjective mens rea is to our substantive results. To that end, we present an extension in which the court follows objective mens rea instead.

TABLE 4 Strategy profiles in the optimal equilibria

When $F^b > F^u$

(a) When $\overline{F} = 0$

(x, y)	a^u	a^b
(–1,–1)	0	1
(–1,1)	1	1

(b) When $\overline{F} = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	η_1
(–1,1)	0	1

When $F^u > F^b$

(c) When $\overline{F} = F^b$

(x, y)	a^u	a^b
(–1,–1)	0	η_2
(–1,1)	1	1

Our running assumption in the baseline model is that the court's objective is to infer the *agent's preferences* from the information available to it and it wants to punish only the biased agent. That is, it wants to punish an agent only if it is sufficiently convinced that the agent caused harm because his preferences are not aligned with society's. Legal philosophers call this notion subjective mens rea.

An alternative specification could be to assume that the court tries to infer the agent's *(non-verifiable) information* from what it observes: the choice made by the agent, the outcome, and the (verifiable) information. The court wants to punish the agent for acting when the available information indicated that he should have exercised restraint. Legal philosophers call this notion objective mens rea.

Although mens rea as a requirement for conviction is a doctrine from criminal law, it serves as a principle in tort cases too. That is, a person's type or intentions play an important role in courts' conviction decisions in tort cases as well. For example, standards of care such as recklessness and (gross) negligence focus on the conscious and voluntary state of mind. In particular, if the court employs the reasonable-person standard to assess the presence of negligence, then its goal is to determine whether a person with reasonable preferences would have acted in a certain way.[6]

In addition, discrimination lawsuits also use type attributes to prove intentional discrimination under Title VII of the Civil Rights Act. For example, in Wilson v. Susquehanna Township Police Department, 55 F.3rd 126 (3rd Cir. 1995),[7] the court ruled that the police chief's intent was to discriminate because it was evident (to the court) that the chief held a "strong gender bias." The court did not question the lower court's ruling that there may have been reasons to promote another person instead of the plaintiff but overruled it on the basis of the "discriminatory attitude" of the chief as "'direct evidence' of discriminatory animus."[8]

As discussed above, both subjective and objective mens rea seem to be reasonable assumptions depending on which environment is being captured. We choose to use subjective mens rea in the baseline model for two reasons. The first is an economic reason: we are interested in the welfare-maximizing equilibria. As we show later, welfare under subjective mens rea is greater than under objective mens rea (see Figure 4 for an illustration). The second reason is that, as discussed above, the courts seem to employ subjective mens rea regularly in and outside of criminal

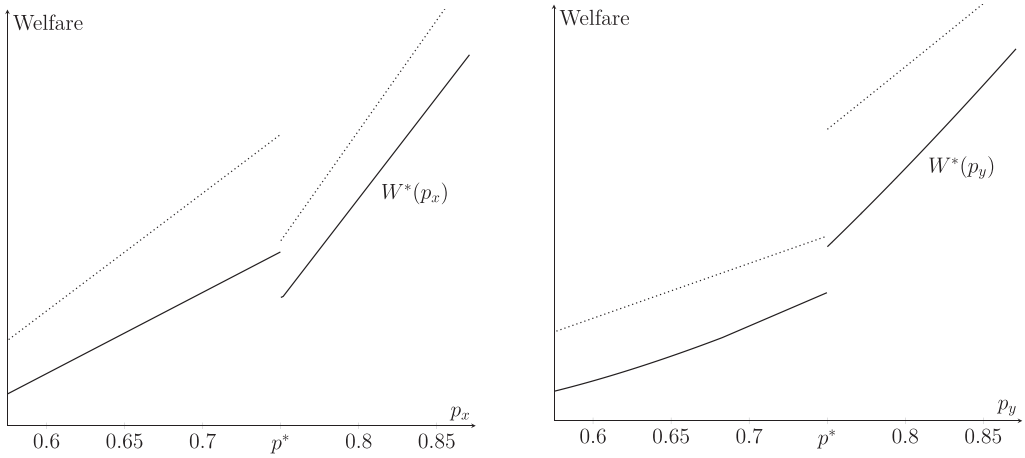
[6] The reasonable-person standard explicitly takes into account that the reasonable person is sophisticated and acts "in the shadow of the law." That is, she takes legal consequences into account when deciding whether to act.

[7] See https://m.openjurist.org/55/f3d/126/wilson-v-susquehanna-township-police-department-l.

[8] In some cases, the court even uses prior acts to determine the agent's type; see, for example, https://www.newyorker.com/magazine/2012/03/19/tax-me-if-you-can, about a case in which a citizen was acquitted because of prior proof of character. For an economic discussion on the use of character evidence in various settings, see Lester et al. (2012); Bull and Watson (2019); Sanchirico (2001). In our model, character evidence (as usually defined) is absent. Any information the court uses to determine culpability is either about the *project* or about the agent's equilibrium behavior.

FIGURE 4

WELFARE WHEN THE COURT AIMS TO CONVICT ONLY AGENTS THAT ACTED DESPITE BETTER INFORMATION



(a) Welfare changes in precision of $\mathbf{X}$

(b) Welfare changes in precision of $\mathbf{Y}$

Welfare is depicted as a function of the precision of the verifiable information, $W^*(p_x)$ (left panel), and as a function of the precision of the unverifiable information, $W^*(p_y)$(right panel). Solid lines depict the values under objective mens rea. Dotted lines are the values for the baseline case of subjective mens rea. Parameters: $\overline{\gamma} = 1/2$, $\gamma = 11/20$, $\beta = 9/13$, and $p_y = 3/4$ (left panel), $p_x = 3/4$ (right panel).

law. Having said that, we want to highlight that our main comparative statics (and the underlying intuition) hold regardless of which formulation of mens rea is used. We show this below.

Objective mens rea. Let the agent be punished if he took an action, $a = 1$, that resulted in a bad outcome, $\theta = -1$, and the court is sufficiently convinced that the agent's signal indicated that he should not have acted; that is, the agent's signal was $(-1, -1)$. That is, under objective mens rea the court punishes if

$$q := \mathbb{P}(\boldsymbol{Y} = 1 | \theta = -1, a = 1, \boldsymbol{X} = -1) \leq \overline{\gamma}.$$

Fixing all the parameters we obtain our first result.

Proposition 3. Expected welfare in equilibrium is weakly higher if the court employs subjective mens rea than if it employs objective mens rea.

The intuition underlying Proposition 3 is seen from Table 4. There are three main differences in the equilibrium behavior compared to the baseline case: (i) if $F^b > F^u$ and $F = F^b$, the biased agent acts with positive probability η_1 (as opposed to zero probability in the baseline case) on $(-1, -1)$; (ii) if $F^b < F^u$, the optimal punishment is always F^b; and (iii) the probability with which the biased agent acts on $(-1, -1)$ is η_2, which is larger than η^b, used in the baseline case. These three properties imply Proposition 3.

We now present the effects of changes in information quality for the alternative specification of the court's objective. Here, unlike in the baseline model, welfare may decline upon improving the precision of $\boldsymbol{Y}$, the unverifiable information, when the court adopts objective mens rea. If such a decline occurs, it occurs at the critical level p_y^*. Table 4 highlights the underlying reason. As p_y increases, we may move from panel (b) to panel (c). That transition implies less deterrence of the biased agent, $\eta_2 > \eta_1$, but removes the chilling effect on the unbiased agent. Thus, welfare declines in the transition only if the former effect is larger than the latter.

Other than at the critical threshold p_y^*, welfare is increasing in p_y. The following condition provides a necessary and sufficient condition to guarantee that the positive effect of increasing p_y that mitigates the chilling effect outweighs the negative effect of reduced deterrence:

$$\frac{1-\gamma}{\gamma}(\eta_2(p_y^*) - \eta_1(p_y^*)) \leq \frac{\beta p_y(1-p_x) - (1-\beta)p_x(1-p_y)}{(1-\beta)p_x p_y - \beta(1-p_x)(1-p_y)} \tag{3}$$

The following proposition provides sufficient conditions such that Propositions 1 and 2 maintain for objective mens rea. We illustrate this result in Figure 4. To state it, we need to introduce some additional notation. Let $\mathcal{Y}(\beta, p_x)$ be the set of p_y's such that it is efficient to act iff $\max\{x, y\} = 1$ when the precision of $\boldsymbol{X}$ is p_x and that of $\boldsymbol{Y}$ is p_y. Notice that $\mathcal{Y}(\beta, p_x)$ is compact, and hence we can define $\underline{p_y}(\beta, p_x) := \min \mathcal{Y}(\beta, p_x)$ and $\overline{p_y}(\beta, p_x) := \max \mathcal{Y}(\beta, p_x)$.

Proposition 4. Suppose that it is efficient to act if and only if at least one signal is positive, $\max\{x, y\} = 1$, and that the court employs objective mens rea.

1. **Precision of $\boldsymbol{X}$**: Consider an increase in the precision from p_x to $p_x' > p_x$. Proposition 1 applies. That is, welfare at p_x' may be lower than at p_x.
2. **Precision of $\boldsymbol{Y}$**: Consider an increase in the precision from $p_y \in \mathcal{Y}(\beta, p_x)$ to $p_y' \in \mathcal{Y}(\beta, p_x) > p_y$. Proposition 2 applies—welfare is unambiguously higher at p_y' than at p_y—if either of the following is true:
   (i) Condition (3) holds, or
   (ii) $p_y^* \notin \mathcal{Y}(\beta, p_x)$.

There are two economically interpretable sufficient conditions that arise from (3). First, if γ—the proportion of unbiased agents in the society—is sufficiently large, then the society prefers not to deter the relatively few biased types from acting on $(-1, -1)$. The reason is the associated cost of chilling the unbiased types. So when $F^b > F^u$, the society prefers to give a free pass. However, as p_y increases, we have $F^b < F^u$, leading to an increase in welfare, as in Proposition 2.

Another sufficient condition relates to the tolerance of the court. If $1 - \overline{\gamma} > p_y^*$—that is, the court convicts when it is sufficiently confident that the agent acted on all negative information—then an increase in p_y leads to an easier separation of the two types for the court. The reason is that the information effect is similar to the one in the baseline model. These and other conditions, and the reasoning behind them, are detailed in Appendix B.

□ **Punishment for inaction.** Throughout the article, we have assumed that the court can punish the agent only when $a = 1$ and $\theta = -1$. Our choice is motivated mainly by realism (for recent experimental evidence, see Cox et al., 2017).[9]

We now extend our model to allow the court to punish the agent for not acting. That is, suppose that the court always sees θ and x regardless of whether the agent acted. The court would ideally like to punish the agent for displaying excessive caution by not acting. However, given the lack of commitment on our court's part, this ability to punish for inaction does not remedy the issue. To see this, first recall that, regardless of the punishment scheme, for any realization of the unverifiable signal that an unbiased agent acts on with strictly positive probability, a biased agent finds it optimal to act. Therefore, for any realization, inaction only increases the likelihood that the agent is unbiased, and the court's posterior over the agent's being unbiased must be weakly higher than its prior. Because the prior is higher than the conviction cutoff—that is, $\gamma > \overline{\gamma}$—by assumption, the court chooses to not punish the agent for inaction, even if allowed to do so.

[9] Scholars debate whether not punishing inaction stems from a cognitive bias or from rational behavior (for an overview, see Woollard, 2015).

□ **More than two types of agents.** We now extend our model to capture a setting in which the agent's preferences can have various degrees of misalignment with the designer's preferences. Specifically, suppose that there is a finite set of types, $\{1, 2, \ldots, K\}$. The utility of a type k agent is given by $u^k(a, \theta) := a[\lambda^k\theta + (1 - \lambda^k)]$, where $\lambda^k \in [0, 1]$. Suppose that $0 = \lambda^1 \leq \lambda^2 \ldots \lambda^K = 1$. Notice that type 1 is the biased agent in our main model, whereas type K is the unbiased agent. Let μ_k denote the *ex ante* probability that the agent is of type k. Even in this environment, the essential problem we face is the same: we want to define the maximum punishment that incentivizes agents to act on $(-1, 1)$ whereas disincentivizing agents from acting on $(-1, -1)$. Given any F, the following fact is immediate:

$$a^k(-1, \cdot) > 0 \Rightarrow a^m(-1, \cdot) = 1 \quad \forall m < k.$$

Therefore, we can define K^1 to be the highest type that acts on $(-1, 1)$ and K^{-1} to be the highest type that acts on $(-1, -1)$. Notice that $\lambda^{K^1} \geq \lambda^{K^{-1}}$.

This model delivers the same results as Propositions 1 and 2. To see why, recall that the main driver of those results is Lemma 1, which established that $F^b - F^u$ is increasing in p_x and decreasing in p_y. Similarly to F^b and F^u, we can define F_1^k to be the largest fine that allows type k to act on $(-1, 1)$ and F_{-1}^k to be the minimum fine required to deter type k from acting on $(-1, -1)$. Then, straightforward algebra (analogous to the expression of $F^b - F^u$) yields

$$F_{-1}^{K^{-1}} - F_1^{K^1} = -2(\lambda^{K^1} - \lambda^{K^{-1}}) + \frac{\beta}{1-\beta}\frac{1-p_x}{p_x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right].$$

Therefore, $F_{-1}^{K^{-1}} - F_1^{K^1}$ is increasing in p_x and decreasing in p_y, as in Lemma 1. As a consequence, both Propositions 1 and 2 continue to hold for the same reason as in our main model. If p_x increases, we can go from $F_{-1}^{K^{-1}} - F_1^{K^1} > 0$ to $F_{-1}^{K^{-1}} - F_1^{K^1} < 0$. If the likelihood of type K^{-1} is sufficiently large, then this can result in more inefficiencies, exactly as in our model with two types. Similarly, the effect of increasing p_y is also identical to that in our main model.

□ **General signal structures.** At first glance, it may appear that our result relies heavily on the assumption that the information is binary and that the precision of each signal is symmetric regarding type I and type II errors. However, that is not the case. We have chosen to present our results in the baseline model with a binary signal structure because it makes it considerably easier to elucidate the mechanism clearly. In Online Appendix C we formally demonstrate how our model extends. We generalize our model to a setting in which the verifiable and unverifiable information can come from a continuum. In such an environment, there could be several measures of precision. We propose an order—the spreading order—by comparing information according to how spread out it is around the efficient cutoff (that is, the posterior belief above which it is efficient to act). Loosely speaking, more spread-out information causes the posterior distributions to be more extreme relative to the efficient belief. Importantly, the spreading order is a strengthening of the Blackwell order.[10]

We show that a more spread-out verifiable signal can decrease welfare, whereas a more spread-out unverifiable signal is always welfare increasing. Importantly, the mechanics are identical to those in the binary world: screening the critical types becomes easier with a more spread-out unverifiable signal but harder with a more spread-out verifiable signal.

5. Applications

■ Before concluding, we wish to consider some settings—within and outside the formal legal system—to which our model applies.

[10] Although not exactly the same, the rotation order defined in Johnson and Myatt (2006) shares most features with our spreading order.

First, consider a set of bureaucrats of which some are corrupt, others work with society's interests in mind. Each bureaucrat (the agent) decides whether to approve the expenditure on a certain project, which may be overpriced. Approving expenses means taking a risk because failed, overpriced projects may lead to corruption charges against the bureaucrat. The bureaucrat relies on verifiable information (e.g., reports) and unverifiable information (e.g., expertise) to inform himself whether the project is overpriced and to decide whether to approve it. The punishment for overpricing depends on the verifiable but not the unverifiable information. If the bureaucrat is found guilty of corruption, he is sentenced by the court.

Second, consider a doctor deciding on the method for delivering a baby. The doctor relies on some verifiable information (e.g., examinations indicating the fetus's position in the womb) and on some unverifiable information (e.g., his expertise and tacit knowledge about the fetus). Although a C-section is the best method in some cases, choosing an unnecessary C-section risks dire consequences for the mother and the baby. Different doctors value compensation and their own time differently, and C-sections pay substantially higher and are scheduled. If a C-section leads to complications, and if no verifiable evidence supports the doctor's choice, he may face legal and administrative consequences. In such a case, if the examiner (a court or hospital administration) concludes that the doctor's interests are not aligned with the patient's, it may wish to punish the doctor by, for example, temporarily revoking his privileges. It applies the reasonable-person standard: to infer the doctor's underlying preferences it compares his behavior to that of a (hypothetical) unbiased doctor.

As a third example, consider a president (the agent) deciding on a foreign policy issue—for instance, whether to impose sanctions on a country in response to its invasion of a neutral country. This is a risky decision, as the president's electoral chances might be compromised following negative outcomes. The safe option is to follow standard diplomatic procedures. Different politicians may value the welfare of their constituency differently, especially when weighing it against the wishes of particular special interest groups. The decision is made based on top-secret information (unverifiable information) and news reports (verifiable information). Voters might hold the president accountable and might want to reelect him only if he values their interests above those of special interests. However, voters have access only to the outcome and theverifiable information.

Finally, consider a CEO of a firm deciding whether to acquire a smaller firm. The acquisition is risky and may affect the acquiring firm's value and stock price; and the CEO's compensation package with his current and future employers may depend on its outcome. When the CEO decides whether to proceed with the acquisition, he relies on verifiable information about the firm to be acquired but also on unverifiable information about the synergy between the firms and about the general outlook of the market. Different CEOs might weigh long-run and short-run outcomes differently. Reviewing a failed acquisition, major shareholders may want to fire a CEO that is interested only in short-term outcomes, whereas they may wish to retain one that is interested in the long-run development of the firm.

6. Conclusion

■ This article highlights that it is not merely the amount of information but its nature that has important welfare consequences. We consider a setting in which an agent decides under uncertainty and may be held liable in court if the decision causes harm. We focus on information of two different natures: that which is verifiable in court and that which is not. We show that increasing the information available to the agent has different consequences depending on its nature. Although increasing the precision of unverifiable information always increases welfare, increasing the precision of verifiable information may reduce it.

Our findings extend to a variety of settings well beyond legal systems. Whether we consider politicians seeking reelection, CEOs wanting to extend their contracts, or bureaucrats with career concerns, our results apply whenever the principal's *ex post* evaluation of a risky decision is based

only on part of the information available to the agent. The principal has to balance the chilling effect against the desire to deter, and changes in the information structure influence her ability to do so. Our findings are robust to a variety of changes in the assumptions. Although details in the timeline or the principal's choice set may differ, the main result remains. The welfare effects of a change in the precision of the information depend on the nature of that information.

The main driver of our result—the tension between deterrence and the chilling effect—has been extensively documented in the legal and management literatures as well as in the popular press.[11] Our results show that whether the chilling effect is pronounced enough to outweigh the overall gains from more information is an empirical question. Thus, a natural direction for future research is to empirically quantify the impact of the chilling effect and its interaction with the provision of information of different natures.

Appendix A: Main results: Proofs

□ **Notation and cases.**

Cases.. The posterior belief, β_{xy}, depends on the informational environment. We ignore the trivial cases in which signals are irrelevant, either because $\beta_{xy} \leq 1/2\ \forall (x,y) \in \{-1,1\}^2$ or because $\beta_{xy} \geq 1/2\ \forall (x,y) \in \{-1,1\}^2$. What remains are parameter values for which we are in exactly one of the following cases.

1. Efficient to act $\Leftrightarrow x = 1$;
2. Efficient to act $\Leftrightarrow y = 1$;
3. Efficient to act $\Leftrightarrow x + y \geq 0$;
4. Efficient to act $\Leftrightarrow x + y = 2$.

Case 1 implies that $\boldsymbol{X}$ is more informative than $\boldsymbol{Y}$. Case 2 implies the reverse. Moreover, a positive realization of the more informative signal is necessary and sufficient to make the project efficient in these cases. Cases 3 and 4 impose no clear ranking between the two types of information. Case 3 implies that β is high and that a necessary and sufficient condition for efficiency is that *one* of the signal realizations is positive. Finally, case 4 implies that β is low and that a necessary and sufficient condition for efficiency is that *both* signal realizations are positive.

Notation. Let $q^u := a^u(-1,1)$ and $q^b := a^b(-1,-1)$ be the agent's best responses to $\overline{F}$ and $F(x)$. Notice that if $F(-1) < (>) F^b$, then $q^b = 1(0)$, and if $F(-1) < (>) F^u$, then $q^u = 1(0)$. Let η^u and η^b be defined by,

$$\frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)(1-p_y+p_y\eta^b)} = \overline{\gamma} \tag{A1}$$

$$\frac{\gamma\eta^u}{\gamma\eta^u+(1-\gamma)} = \overline{\gamma} \tag{A2}$$

If $q^u = 1$ then $q^b = \eta^b \Rightarrow \gamma_{-1} = \overline{\gamma}$, making the court indifferent between any sentence. If $q^b = 0$ and $a^b(-1,1) = 1$, then $q^u = \eta^u \Rightarrow \gamma_{-1} = \overline{\gamma}$.

Finally, let $\overline{W}(\overline{F})$ be the welfare in a designer-optimal perfect Bayesian equilibrium conditional on holding the designer choice fixed at $\overline{F}$.

Lemma 2. Define $\overline{F}^* := \arg\max_{\overline{F}} \overline{W}(\overline{F})$. Then, $\overline{F}^* \in \{0, F^u, F^b\}$

Proof.

Claim 1. $q^u = 1 \Rightarrow q^b \in \{\eta^b, 1\}$ wlog.

Proof. $q^b = 0 \Rightarrow \gamma_{-1} = \gamma > \overline{\gamma}$. Therefore, $F(-1) = 0$. Therefore, the biased agent would deviate to play $q^b = 1$. Therefore, $q^b > 0$. Also, $q^b = 1 \Rightarrow \gamma_{-1} = \frac{\gamma(1-p_y)}{\gamma(1-p_y)+1-p_y} < \overline{\gamma}$. Therefore, $F(-1) = \overline{F}$. Notice that $\overline{F} < F^b \Rightarrow q^b = 1$. If $\overline{F} \geq F(-1) > F^b \Rightarrow q^b = 0 \Rightarrow \gamma_{-1} = \gamma > \overline{\gamma}$. This would imply that $F(-1) = 0$, a contradiction. Therefore, if $\overline{F} > F^b$, the biased type would mix to have $\gamma_{-1} = \overline{\gamma}$—that is, $q^b = \eta^b$, so that $F(-1) = F^b$. In the case when $F = F^b$, $q^b \in [\eta^b, 1]$. In this case, $q^b = \eta^b$ is the designer-optimal equilibrium. □

Claim 2. $F^u > F^b$ and $q^b > 0 \Rightarrow q^u = 1$.

[11] See, for instance, Hylton (2019), Chalfin and McCrary (2017), Bernstein (2014), and Bibby (1966)

Proof. $q^b > 0 \Rightarrow F(-1) \leq F^b < F^u \Rightarrow q^u = 1$. □

Claim 3. If $F^u > F^b$, $\overline{F}^* \in \{F^b, F^u\}$.

Proof. First, notice that $F(-1) < F^b \Rightarrow q^u = q^b = 1$. Instead, with $F(-1) = F^b \Rightarrow q^u = 1$, $q^b = \eta^b$, giving us a strict improvement in efficiency.

If $F(-1) \in (F^b, F^u)$, then $q^u = 1$ and $q^b = 0$. Then, $\gamma_{-1} = \gamma > \overline{\gamma} \Rightarrow F(-1) = 0$, a contradiction. Therefore, $F(-1) \notin (F^b, F^u)$ in equilibrium.

If $F(-1) > F^u$ then $q^u = q^b = 0$. Instead, $F(-1) = F^u$ provides a strict efficiency improvement by having $q^u \in [0, \eta^u]$, $q^b = 0$. The optimal choice is to have $q^u = \eta^u$. $q^u \leq \eta^u$ because, otherwise, $\gamma_{-1} > \overline{\gamma}$, and, therefore, $F(-1) = 0$, a contradiction. □

Claim 4. If $F^u < F^b$, $\overline{F}^* \in \{0, F^b\}$.

Proof. Here, whenever $q^u > 0$, $q^b = 1$. Therefore, either $q^u = q^b = 1$, achieved by $\overline{F} = 0$, or $q^u = q^b = 0$, achieved by $\overline{F} = F^b$. Which of the two is optimal depends on whether $\overline{W}(0) > \overline{W}(F^b)$ or vice-versa. It is easy to check that,

$$\begin{aligned}\overline{W}(0) - \overline{W}(F^b) = \gamma[\beta(1-p_x)p_y - (1-\beta)p_x(1-p_y)] \\ + (1-\gamma)[\beta(1-p_x)(1-p_y) - (1-\beta)p_x p_y].\end{aligned}$$

Therefore, if γ is sufficiently high, $\overline{F} = 0$; otherwise, $\overline{F} = F^b$. □

Together, the claims imply that $\overline{F}^* \in \{0, F^b, F^u\}$. □

□ **Proof of Proposition 1 and Lemma 1.** Now we are equipped to present our main comparative static. To this end, let $W^*(\cdot,\cdot,\cdot,\overline{F};\cdot,\cdot) := \overline{W}(\overline{F}^*)$ denote the equilibrium welfare given S. Let $\Delta(p_x, p_y) := F^b - F^u$.

Proof of Lemma 1.

$$\begin{aligned}F^b =& \frac{1}{1-\beta_{-1,-1}} = 1 - \frac{\beta}{1-\beta}\frac{1-p_y}{p_y} + \frac{\beta}{1-\beta}\frac{1-p_y}{p_y}\frac{1}{p_x} \\ F^u =& -2 + \frac{1}{1-\beta_{-1,1}} = -2 + \frac{-\beta}{1-\beta}\frac{p_y}{1-p_y} + \frac{\beta}{1-\beta}\frac{p_y}{1-p_y}\frac{1}{p_x} \\ \Rightarrow \Delta(p_x, p_y) =& 2 + \frac{\beta}{1-\beta}\frac{1-p_x}{p_x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right]\end{aligned}$$

The above is increasing in p_x and decreasing in p_y. □

Proof of Proposition 1. Fix some (p_y, β). At p_x^*, $F^b(p_x^*) = F^u(p_x^*)$. Suppose that $p_1 < p_x^* < p_2$. Therefore, $F^b(p_1) < F^u(p_1)$ *and* $F^b(p_2) > F^u(p_2)$ by Lemma 1.

Case 1: $\overline{F}^*(p_2) = 0$.[12]

Hence, $q^b(p_2) = q^u(p_2) = 1$, By Claim 3, $\overline{F}^*(p_1) \in \{F^u(p_1), F^b(p_1)\}$. Suppose that $F(-1) = F^b(p_1)$. Therefore, $q^b(p_1) = \eta^b$ and $q^u(p_1) = 1$. Notice that (A1) features no dependence on p_1 and η^b is strictly less than 1.

Let $W_1 := \overline{W}(F^b(p_1))$ and $W_2 := \overline{W}(0)$.

$$\begin{aligned}W_i =& \beta\big[p_i + (1-p_i)[p_y + (1-\gamma)(1-p_y)q^b(p_i)]\big] \\ & - (1-\beta)\big[(1-p_i) + p_i[(1-p_y) + (1-\gamma)p_y q^b(p_i)]\big]\end{aligned}$$

Therefore,

$$\begin{aligned}W_1 - W_2 =& (p_1 - p_2)[\beta(1-p_y) + (1-\beta)p_y] \\ & + (1-\gamma)\big[\eta^b\big[\beta(1-p_1)(1-p_y) - (1-\beta)p_1 p_y\big] \\ & - \big[\beta(1-p_2)(1-p_y) - (1-a)p_2 p_y\big]\big]\end{aligned}$$

Suppose that for a small $\delta > 0$, $p_1 = p_2 - \delta$. Then,

$$W_1 - W_2 = (1-\gamma)(1-\eta^b)[(1-\beta)p_y p_1 - \beta(1-p_y)(1-p_1)] + o(\delta).$$

[12] $\overline{F}^*(p)$ denotes $\overline{F}^*$ in the environment with $p_x = p$ *ceteris paribus*.

Because it is inefficient to act on $(-1,-1)$, $\beta_{-1,-1} = \frac{\beta(1-p_y)(1-p_1)}{\beta(1-p_y)(1-p_1)+(1-\beta)p_y p_1} < \frac{1}{2}$. Equivalently, $(1-\beta)p_y p_1 > \beta(1-p_y)(1-p_1)$. Therefore, $W_1 > W_2$. Lastly, if $\overline{F}^*(p_1) = F^u(p_1)$, then $W_1 > W_2$ for small enough $\delta > 0$.

Case 2: $\overline{F}^*(p_2) = F^b(p_2)$.

Therefore, $q^b(p_2) = q^u(p_2) = 0$. Setting $F(-1) = F^u(p_1)$, we have $q^b(p_1) = 0$ *and* $q^u(p_1) = \eta^u > 0$. Because the only change is that the unbiased type acts on $(-1, 1)$ with probability η^u, the extent of the chilling effect is reduced. Therefore, as before, $W^*(p_1) > W^*(p_2)$ as $\delta \to 0$. □

□ **Proof of Proposition 2.** *Proof.* We prove the proposition here only for the *interior* of our case. We do so by looking at two types of arguments. Applying these arguments in various combinations is, in fact, sufficient to prove all other cases and the transition from one case to another. We do that in Online Appendix D.3.

The court can observe x, the realization of $\boldsymbol{X}$. Thus, we can look at the cases separately and provide an argument for each.

Argument 1 ($x = 1$). As long as we remain inside our case, the court provides a free pass ($F(x = 1) = 0$) on realization $x = 1$ for any level of p_y. In addition, both types act whenever they see $x = 1$ and ignore signal p_y entirely. Thus, any improvement on p_y conditional on a realization $x = 1$ does not affect the welfare.

Argument 2 ($x = -1$). Compare two environments with p_y, p_y' such that $p_y' > p_y$. First, assume that $\overline{F}^* = 0$ for both levels. Increasing precision does not change $a^b(\cdot)$, but projects implemented by the unbiased agent fail less often. Second, assume that $\overline{F}^* = F^u$ for both levels. Then, no agent acts when it is inefficient to act (yet there is a moderate chilling effect: see Table 1). Because precision increases, the signal on $(-1, 1)$ is stronger and welfare improves. Third, assume that $\overline{F}^* = F^b$ for both levels. Because η^b decreases in p_y, the biased agent's actions on y improve from an efficiency perspective, whereas the unbiased agent's decisions can only improve by Lemma 1. Welfare increases. What remains is to show that welfare improves as we move from $\overline{F} = 0$ to $\overline{F} = F^\omega$. A change from $\overline{F} = F^0$ to $\overline{F} = F^b$ occurs either if $F^b > F^u$ or if $F^b = F^u$. In the former case, both equilibria are available, and the switch occurs because $\overline{W}(F^b)$ overtakes $\overline{W}(0)$, an improvement in welfare. In the latter case, welfare improves because the only behavioral change is that the biased agent selects the inefficient action less often. Finally, a change from $\overline{F} = 0$ to $\overline{F} = F^u$ can occur only at $F^b = F^u$, and, by construction, $\overline{F} = F^u$ dominates $\overline{F} = F^b$. The proof is complete.

□

Appendix B: Objective mens rea: Characterization and proofs

Equilibrium characterization.. The court is indifferent if $q = \overline{\gamma}$. If $\overline{F} = F^b > F^u$, the designer-optimal equilibrium implies that $a^u(-1, 1) = 0$, $a^b(-1, 1) = 1$ and $a^b(-1, -1) = \eta_1$ with

$$\eta_1 = \min\left\{\frac{(1-p_y)}{p_y}\frac{(1-\overline{\gamma})}{\overline{\gamma}}, 1\right\}.$$

If $\overline{F} = F^b < F^u$, the designer-optimal equilibrium implies that $a^u(-1, 1) = 1$, $a^b(-1, 1) = 1$ and $a^b(-1, -1) = \eta_2$ with

$$\eta_2 = \min\left\{\frac{(1-p_y)}{p_y}\frac{(1-\overline{\gamma})}{\overline{\gamma}}\frac{1}{1-\gamma}, 1\right\}.$$

Recall, that $F^b - F^u$ does not depend on the court's choice, it is still given by

$$\Delta(p_x, p_y) = 2 + \frac{\beta}{1-\beta}\frac{1-p_x}{p_x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right]$$

If $F^b > F^u$, any punishment below F^b implies that the biased agent is never deterred from acting. If, in addition, $\overline{F} > F^u$, the unbiased agent is deterred from acting on $(-1, 1)$, which is clearly worse. Thus, a designer-optimal equilibrium exists for either $\overline{F} = 0$ or $\overline{F} = F^b$. The court's indifference condition implies η_1.

If $F^b < F^u$, a punishment above F^b does not improve upon F^b, as it would lead to actions only on $(-1, 1)$, which, in turn, implies that the court does not punish. Conditional on not facing punishment, the biased type has an incentive to deviate and act on both $(-1, 1)$ and $(-1, -1)$, which, in turn, implies that not punishing is suboptimal. No punishment yields a better outcome than the designer-optimal equilibrium under $\overline{F} = F^b$. Thus, it is sufficient to consider $\overline{F} = F^b$ only if $F^b < F^u$. The court's indifference condition implies η_2.

Proof of Proposition 3.. The level of F^b is unaffected by the court's objective, and so is the ranking F^b vs F^u. It suffices to show that welfare is lower for $\overline{F} = F^b$. For $\overline{F} = 0$, welfare is, by construction, identical, and $\overline{F} = 0$ is selected only if it improves upon $\overline{F} = F^b$. Similarly, $\overline{F} = F^u$ is selected only if it improves on $\overline{F} = F^b$ in the baseline case and never under the objective mens rea. Thus if equilibria conditional on $\overline{F} = F^b$ are welfare-inferior for one court objective, the designer-optimal equilibrium is welfare-inferior under that objective.

To see that result, observe that action profiles are identical, apart from the biased agent's decision on $(-1, -1)$. If $F^b < F^u$ she chooses $\eta_1 > 0$ for the court's objective assumed in this section (punishing for acting on wrong information)

and 0 under the court's objective in the baseline model.[13] Because acting is inefficient for the information $(-1, -1)$, the alternative objective is welfare-inferior. If $F^b > F^u$, the agent chooses

$$\eta_2 = \max\{\frac{(1-p_y)}{p_y}\frac{(1-\overline{\gamma})}{\overline{\gamma}}\frac{1}{1-\gamma}, 1\} > \frac{1-p_y}{p_y}\frac{\gamma-\overline{\gamma}}{\overline{\gamma}}\frac{1}{1-\gamma} = \eta^b.$$

Again, the alternative objective is welfare-inferior.

Proof of Proposition 4.. The first part follows by using the parameters that are used for the figures. Alternatively, one can use a constructive version similar to that of the proof of Propositions 1. We omit it, as it provides no further insight. We discuss the second part below.

When is welfare unambiguously increasing in the precision of p_y?

First, consider $p_y < p_y' < p_x^*$ such that $p_y, p_y' \in \mathcal{Y}(\beta, p_x)$. Here, the equilibria from the top row of Table 4 are available. It is easy to check that welfare is continuous and increasing in p_y for each of these equilibria. Therefore, $W^*(\beta, p_x, \cdot, \gamma)$, which selects the maximum of the welfare generated by the two equilibria, is also continuously increasing on $[\underline{p_y}(\beta, p_x), p_y^*)$.

Using a similar argument $W^*(\beta, p_x, \cdot, \gamma)$ is continuously increasing on $(p_y^*, \overline{p_y}(\beta, p_x]$. Finally, a switch from p_y to a p_y' such that $p_y' > p_y^* > p_y$ that entails switching from $\overline{F} = 0$ to $\overline{F} = F^b$ is also welfare improving as it only increases deterrence without having a chilling effect. Therefore, the only case we need to consider is the case in which $\overline{F} = F^b$ on both sides of p_y^*, and precision increases from $p_y < p_y^*$ to $p_y' > p_y^*$. In all other cases, welfare increases in p_y.

A necessary and sufficient condition for the designer to prefer $\overline{F} = F^b$ over the free pass when $p_y < p_y^*$ is $W(p_y)$ is higher under $\overline{F} = F^b$. That is the case when

$$\begin{aligned}\beta\big(p_y(1-p_x) + (1-\gamma)(1-p_x)(1-p_y)\big) - (1-\beta)\big(p_x(1-p_y) + (1-\gamma)p_xp_y\big) >\\ \beta\big(p_y(1-p_x)(1-\gamma) + (1-\gamma)(1-p_x)(1-p_y)\eta_1\big) - (1-\beta)\big(p_x(1-p_y)(1-\gamma) + (1-\gamma)p_xp_y\eta_1\big)\end{aligned}$$

which can be simplified to

$$\frac{1-\gamma}{\gamma}(1-\eta_1) > \underbrace{\frac{\beta p_y(1-p_x) - (1-\beta)p_x(1-p_y)}{(1-\beta)p_xp_y - \beta(1-p_x)(1-p_y)}}_{:=\hat{\Delta}(p_y)} > 0 \tag{B1}$$

where the last inequality follows because—by assumption—it is efficient to act when any signal is positive.

Next consider the case in which $\overline{F} = F^b$ and define

$$\begin{aligned}f_1(p_y) :=& \beta\big[p_x + (1-p_x)(1-\gamma)[p_y + (1-p_y)\eta_1]\big]\\ &- (1-\beta)\big[(1-p_x) + p_x(1-\gamma)[1-p_y+p_y\eta_1]\big]\\ f_2(p_y) :=& \beta\big[p_x + (1-p_x)[p_y + (1-p_y)(1-\gamma)\eta_2]\big]\\ &- (1-\beta)\big[(1-p_x) + p_x[(1-p_y) + p_y(1-\gamma)\eta_2].\big]\end{aligned}$$

Notice that $W^*(p) = f_1(p)$ if $p < p_y^*$ and $W^*(p) = f_2(p)$ if $p_y' \geq p_y^*$. Both $f_1(\cdot)$ and $f_2(\cdot)$ are increasing in p_y. Thus, if $f_2(p_y^*) \geq f_1(p_y^*)$, welfare is increasing in p_y also around p_y^*. Otherwise, it is not.

Formally,

$$\begin{aligned}f_2(p_y^*) - f_1(p_y^*) =& \beta(1-p_x)p_y^*\gamma + \beta(1-p_x)(1-p_y^*)(1-\gamma)(\eta_2-\eta_1)\\ &- (1-\beta)p_x(1-p_y^*)\gamma - (1-\beta)p_xp_y^*(1-\gamma)(\eta_2-\eta_1)\end{aligned}$$

or equivalently

$$\begin{aligned}f_2(p_y^*) - f_1(p_y^*) =& \gamma\underbrace{[\beta(1-p_x)p_y^* - (1-\beta)p_x(1-p_y^*)]}_{>0}\\ &- (1-\gamma)(\eta_2-\eta_1)\underbrace{[(1-\beta)p_xp_y^* - \beta(1-p_x)(1-p_y^*)]}_{>0}\end{aligned}$$

[13] For convenience, we call the court's objective in the baseline case as the "baseline object" and the court's objective in this section as the "alternative objective".

The signs of the two quantities above follow from the fact that it is efficient to act on $(-1, 1)$ and inefficient to act on $(-1, -1)$. Thus, welfare increases around p_y^* if and only if

$$\frac{(1-\gamma)}{\gamma}(\eta_2 - \eta_1) \leq \underbrace{\frac{\beta(1-p_x)p_y^* - (1-\beta)p_x(1-p_y^*)}{(1-\beta)p_x p_y^* - \beta(1-p_x)(1-p_y^*)}}_{=\widehat{\Delta}(p_y^*)}. \tag{B2}$$

Notice that if condition (B2) is violated for p_y^* it is also optimal to implement $\overline{F} = F^b$ for p_y^* because $1 - \eta_1 \geq \eta_2 - \eta_1$ and hence a violation of (B2) implies (B1). Thus, a necessary and sufficient condition for Proposition 2 to hold is that

$$(\eta_2 - \eta_1)\frac{(1-\gamma)}{\gamma} \leq \widehat{\Delta}(p_y^*).$$

Observe that $\widehat{\Delta}(p_y^*)$ is independent of the courts threshold belief $\overline{\gamma}$. Moreover,

$$\eta_2 - \eta_1 = \begin{cases} 0 & \text{if } \overline{\gamma} \leq 1 - p_y \\ 1 - \frac{1-p_y}{p_y}\frac{1-\overline{\gamma}}{\overline{\gamma}} & \text{if } 1 - p_y < \overline{\gamma} < \frac{1-p_y}{1-p_y\gamma} \\ \frac{\gamma}{1-\gamma}\frac{1-p_y}{p_y}\frac{1-\overline{\gamma}}{\overline{\gamma}} & \text{if } \overline{\gamma} \geq \frac{1-p_y}{1-p_y\gamma}. \end{cases}$$

Notice further that $\eta_2 - \eta_1$ is increasing in $\overline{\gamma}$ if and only if $\overline{\gamma} \in [1 - p_y, \frac{1-p_y}{1-p_y\gamma}]$ and therefore its maximum at $\overline{\gamma} = \frac{1-p_y}{1-p_y\gamma}$ where $\eta_2 - \eta_1 = \gamma$ which implies that $\eta_1 - \eta_2 \in [0, \gamma]$.

Thus, independent of $\overline{\gamma}$, (B2) holds if

$$(1-\gamma) \leq \frac{\beta(1-p_x)p_y^* - (1-\beta)p_x(1-p_y^*)}{(1-\beta)p_x p_y^* - \beta(1-p_x)(1-p_y^*)}$$

which can be simplified to

$$\frac{1-\beta}{\beta}\frac{p_x}{1-p_x} \leq \frac{1-\gamma(1-p_y^*)}{1-p_y^*\gamma}. \tag{B3}$$

Because $p_y > 1/2$ the right-hand side of the above larger than 1 which in turn implies that if $\beta > p_x$, then Proposition 2 holds for any $(\gamma, \overline{\gamma})$.

Moreover, $p_y > 1/2$ implies that the right-hand side of condition (B3) is increasing in γ with limit $p_y/(1-p_y)$ as $\gamma \to 1$

$$\frac{1-\beta}{\beta}\frac{p_x}{1-p_x} < \frac{p_y}{1-p_y}$$

which holds because we are in the case in which a any positive signal makes it efficient to act. Thus, for any $(\beta, p_x, p_y, \overline{\gamma})$ such that we are in our case of interest, there exists a $\hat{\gamma} < 1$ such that if the likelihood that the agent is unbiased $\gamma > \hat{\gamma}$, Proposition 2 holds.

Finally, even if (B3) fails, there always is a threshold $\underline{\gamma}^* > 1 - p_y^*$ such that Proposition 2 holds if $\overline{\gamma} < \underline{\gamma}^*$. The reason is that for $\overline{\gamma}$ low enough $\eta_2 - \eta_1 = 0$.

References

Bernstein, E. "The Transparency Trap." *Harvard Business Review*, Vol. 92 (2014), pp. 58–66.

Bibby, J.F. "Committee Characteristics and Legislative Oversight of Administration." *Midwest Journal of Political Science*, Vol. 10 (1966), pp. 78–98.

Bull, J. and Watson, J. "Statistical Evidence and the Problem of Robust Litigation." *The RAND Journal of Economics*, Vol. 50 (2019), pp. 974–1003.

Cane, P. "Mens Rea in Tort Law." *Oxford Journal of Legal Studies*, Vol. 20 (2000), pp. 533–556.

Chalfin, A. and McCrary, J. "Criminal Deterrence: A Review of the Literature." *Journal of Economic Literature*, Vol. 55 (2017), pp. 5–48.

Cox, J.C., Servátka, M., and Vadovič, R. "Status Quo Effects in Fairness Games: Reciprocal Responses to Acts of Commission Versus Acts of Omission." *Experimental Economics*, Vol. 20 (2017), pp. 1–18.

Garoupa, N. "The Economics of Political Dishonesty and Defamation." *International Review of Law and Economics*, Vol. 19 (1999), pp. 167–180.

Hylton, K.N. "Economic Theory of Criminal Law." *Oxford Research Encyclopedia of Economics and Finance*. doi: 10.1093/acrefore/9780190625979.013.344.

Johnson, J.P. and Myatt, D.P. "On the Simple Economics of Advertising, Marketing, and Product Design." *American Economic Review*, Vol. 96 (2006), pp. 756–784.

Kaplow, L. "On the Optimal Burden of Proof." *Journal of Political Economy*, Vol. 119 (2011), pp. 1104–1140.
Kaplow, L. "Optimal Design of Private Litigation." *Journal of Public Economics*, Vol. 155 (2017a), pp. 64–73.
Kaplow, L. "Optimal Multistage Adjudication." *The Journal of Law, Economics, and Organization*, Vol. 33 (2017b), pp. 613–652.
Lagunoff, R. "A Theory of Constitutional Standards and Civil Liberty." *The Review of Economic Studies*, Vol. 68 (2001), pp. 109–132. doi: 10.1111/1467-937X.00162.
Lester, B., Persico, N., and Visschers, L. "Information Acquisition and the Exclusion of Evidence in Trials." *The Journal of Law, Economics, & Organization*, Vol. 28 (2012), pp. 163–182.
Morris, S. and Shin, H.S. "Social Value of Public Information." *American Economic Review*, Vol. 92 (2002), pp. 1521–1534.
Pei, H. and Strulovici, B. "Crime Entanglement, Deterrence, and Witness Credibility." *mimeo*.
Prat, A. "The Wrong Kind of Transparency." *American Economic Review*, Vol. 95 (2005), pp. 862–877.
Prendergast, C. "A Theory of"Yes Men"." *The American Economic Review*, pp. 757–770.
Sanchirico, C.W. "Character Evidence and the Object of Trial." *Columbia Law Review*, Vol. 101 (2001), pp. 1227–1311.
Schrag, J. and Scotchmer, S. "Crime and Prejudice: The Use of Character Evidence in Criminal Trials." *Journal of Law, Economics, & Organization*, pp. 319–342.
Stigler, G.J. "The Optimum Enforcement of Laws." *Journal of Political Economy*, Vol. 78 (1970), pp. 526–536.
Blanes i Vidal, J. and Möller, M. "When Should Leaders Share Information with their Subordinates?" *Journal of Economics & Management Strategy*, Vol. 16 (2007), pp. 251–283.
Woollard, F. *Doing and Allowing Harm*. 1. Oxford: Oxford University Press, 2015.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Posterior beliefs with $\boldsymbol{Y}$ and $\boldsymbol{Y}'$