THE WRONG KIND OF INFORMATION

Aditya Kuvalekar João Ramos Johannes Schneider*

November, 2021

Abstract

An agent decides whether to approve a project based on his information, some of which is verified by a court. An unbiased agent wants to implement projects that are likely to succeed; a biased agent wants to implement any project. If the project fails, the court examines the verifiable information and decides the punishment. The court seeks to deter ill-intentioned agents from implementing projects likely to fail while incentivizing the use of the unverifiable information. We show how information of different kinds affects welfare. Improving the verifiable information can reduce welfare, whereas improving the unverifiable information always increases welfare.

1 Introduction

People and organizations adapt their choices and behavior to fit the rule of law. When laws and regulations are designed to mitigate agency problems—to deter ill-intentioned agents from acting against the common interest—there is a side-effect: well-intentioned agents refrain from socially desirable actions for fear of being mistaken for ill-intentioned agents.

From politicians to doctors to civil servants, examples abound of people avoiding risky (but efficient) decisions for fear of being sued.[1] This phenomenon—the unin-

*Kuvalekar: U of Essex; Ramos: Queen Mary U and USC Marshall; Schneider: U Mannheim & U Carlos III de Madrid. We are indebted to the editor, Nicola Persico, and three anonymous referees for excellent comments that improved the paper substantially. We thank Nageeb Ali, 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 and Maher Said for helpful comments and discussions. Aditya Kuvalekar gratefully acknowledges support from the Ministerio Economia y competitividad grant PGC2018-09159-B-I00. Johannes Schneider gratefully acknowledges financial support from the German Research Foundation (DFG) through CRC TR 224 (Project B03), Agencia Estatal de Investigación (grant PID2019-111095RB-I00 and grant PID2020-118022GB-I00), Ministerio Economía y Competitividad (grant ECO2017-87769-P), and Comunidad de Madrid (grant MAD-ECON-POL-CM H2019/HUM-5891).

[1] In a recent speech the Chief Justice of India said that a celebrated transparency law has, in fact, led to "fear and paralysis" among government officials. See, https://www.thehindu.com/news/national/abuse-of-rti-has-led-to-paralysis-and-fear-among-officials-says-cji-bobde/article30320357.ece. Also, Wang (2019) documents that the establishment of higher punishments in order to combat corruption may actually undermine the ability of bureaucrats to accomplish daily tasks due to chilling effects.

tended consequences of a law or a regulation affecting activities outside its intended scope—is the *chilling effect*.

The central subject matter of our paper is the chilling effect that stems from the following pervasive feature: individuals rely on different sources of information when making a decision, and while some sources are verifiable in court, others are not. For example, doctors, civil servants and politicians often possess extremely valuable situational knowledge based on their experience, over and above verifiable information such as exams and reports. The chilling effect of legal consequences is that individuals hesitate to rely on unverifiable information if it contradicts the information that is verifiable in court.

In this paper, we explore the interaction between the chilling effect described above and the agent's information. Our main contribution is to show that it is not merely the amount of information but, rather, the *nature of the information* that has important consequences for welfare. More precisely, although there are unambiguous welfare gains as the precision of the *unverifiable information* increases, welfare may decline in the precision of the *verifiable information*.

In the real world, there are constant improvements in the available information, both verifiable and unverifiable. For example, doctors have better diagnostic tools at their disposal; politicians have access to more specialized expert reports; civil servants have extensive new training programs and specific software for comparing prices; and so on. Our results point out that such improvements may have adverse consequences for welfare.

Model: Motivated by the forces described above, we provide a natural setting with the following four features. First, some agents' preferences may be misaligned with society's. Second, agents act on their information, but only a part of it is verifiable ex-post. Third, society wishes to deter the ill-intentioned agents from making bad decisions, while encouraging agents with good intentions to rely on the (often useful) unverifiable information. Fourth, we assume justice is constrained in its need to prove liability, for example via the fundamental principle of criminal liability—*actus reus non facit reum nisi mens sit rea*—the act is not culpable unless the mind is guilty.[2]

More concretely, there are two players: a court (it) and an agent (he). The agent (unbiased or biased) decides whether to implement a risky project or to take the safe action. The agent relies on two conditionally independent binary signals about a binary state of the world to inform his decision. A key element of our model is that, of the two signals, only one is verifiable (ex-post) in the court, while the other is not. The risky project succeeds when the state is good and fails when the state is bad. The unbiased agent's preferences are aligned with society's. He values successful projects and suffers from failures. The biased agent always wants to implement a project. If the agent implements the risky project, the state of the world is publicly

[2]See Fowler v. Padget (1798).

realized. Thereafter, the court decides whether to punish the agent and what the punishment will be by observing the agent's behavior, the public outcome and also the agent's verifiable information.

The final ingredient of our model is the formalization of a judicial system: a designer can select the penal code ex-ante, but the court convicts the agent based on the following two legal principles:

(i) an *actus reus*—conviction is possible only in case of damages; and,
(ii) *mens rea*—the court's objective is to punish only ill-intentioned agents.[3]

If the agent implements the project and it fails, the court observes the verifiable information and decides on the agent's punishment.

The main tradeoff in setting the optimal punishment is as follows. On the one hand, the fear of punishment may cause a *chilling effect* on the unbiased agent. If the threat of punishment is too large, he will ignore useful (but unverifiable) information.

On the other hand, the designer may give a *free pass* to a biased agent—absent sufficient punishment, he implements the project even when it is inefficient to do so. To highlight the main mechanism, we focus here on the case in which any positive signal (verifiable or unverifiable) is sufficient to make implementation efficient.[4]

For instance, consider an unbiased agent who receives a negative verifiable signal suggesting that he should not implement the project and a positive unverifiable signal indicating that he should. It is efficient (from society's perspective) to implement the project in such a scenario. However, if the unverifiable information is wrong, the agent may be punished because the court observes only the verifiable information. Naturally, the agent will act only if the punishment is not too harsh.

Now, suppose that the verifiable information becomes more precise, but it still is efficient to implement the project upon observing a negative verifiable signal and a positive unverifiable signal. The improved precision of the verifiable information has a direct, obvious benefit: better information. However, this is paired with a more subtle consequence: a more precise verifiable information—in the case of a negative realization—means that the agent is now less optimistic about the project's likelihood of success. Therefore, the agent faces a greater risk of being punished. As a result, if the punishment remains unchanged, the agent may decide not to implement the project—he is "chilled away" from taking the efficient action.

There seems to be a direct remedy to the above issue; lower the punishment if the verifiable information becomes more precise. Therefore, the question is: can we adjust

[3]This notion of mens rea—convicting biased agents—is called subjective mens rea in the legal literature. Alternatively, objective mens rea means that the court wishes to convict the agent only if it is sufficiently confident that the agent's (overall) information indicating the harmful outcome was sufficiently strong. While subjective mens rea outperforms objective mens rea from a welfare perspective, our main qualitative insights continue to hold in both settings. A detailed discussion concerning these issues is in Section 4.2. The special case of strict liability, in which conviction is possible absent mens rea, corresponds to the model discussed in Section 4.1.

[4]Remaining cases are analyzed in the online appendix.

the punishment so that the benefits of improved information outweigh the costs of a stronger chilling effect?

Our main results state that, even when adapting the punishment to the informational environment optimally, the benefits of better information may not compensate for the increased chilling effect.

In particular, increasing the precision of the verifiable information can decrease welfare in non-knife-edge cases (Proposition X), while, in contrast, increasing the precision of the unverifiable information always implies welfare gains (Proposition Y).

Here, we highlight two important points. First, taken together, our main results show that the nature of information—verifiable or not—has qualitatively different effects on welfare. Second, while we choose to present a binary, stylized model, the mechanism that drives this difference (discussed below) goes beyond it.

The mechanism: The main conflict captured in our environment is that, at times, it is efficient for the unbiased agent to rely on his unverifiable information. That is, he can decide in favor of implementing the project if the unverifiable information is positive, even if the verifiable information is negative. However, the cost of structuring a punishment scheme that encourages unbiased agents to rely on unverifiable information is high. It implies that a biased agent may implement a project when both the verifiable and unverifiable information recommends against it. More precise unverifiable information helps the designer here because it makes the unbiased agent more confident about implementing the project by trusting positive unverifiable information. But, at the same time, it disincentivizes the biased agent from implementing the project when the unverifiable information is negative. Welfare increases.

In contrast, if the verifiable information is more precise, then a realized negative and verifiable signal further disincentivizes both types. It increases the threat of punishment, as it indicates a higher chance of failure and, thus, of punishment if the project is implemented.

In addition, there is a second deterring force *only for the unbiased type.* His payoff is connected to the success of the project directly, and the incentives to implement it, given a negative verifiable signal, go down in the signal's precision, regardless of the punishment. Due to these two effects, increasing the precision of the verifiable signal may lead to welfare loss.

Generality of our mechanism: Given the simplicity of the mechanism discussed above, 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. That is, do these forces hinge on the intricate details of a formal legal system, as detailed in our baseline model?

From a theoretical standpoint, the court's objective of screening the agents' types is a key driver of our results. Therefore, leaving the legal application momentarily, we explore a more parsimonious model in Section 4.1: a principal (she) wishes to

screen the agent based on his action, the project outcome, and the public information. Our main results continue to hold, both when the principal commits to a punishment scheme ex ante and when she reacts to the outcome ex post.

The reason behind this result is the same as in the baseline model. More precise verifiable information puts more emphasis on the verifiable information. It makes it harder to separate the types by inferring whether they held favorable or unfavorable unverifiable information—a hurdle to screening. More precise unverifiable information puts less emphasis on the verifiable information. It makes it easier to separate the types by inferring whether they held favorable or unfavorable private information—a pathway to screening.

At first glance, it may appear that the result relies heavily on our assumption that the information is binary (favorable/unfavorable), and the precision of each signal is symmetric with regard to type I and type II errors. However, that is not the case, as we demonstrate in Appendix C. There, we generalize our model to a setting in which the public (verifiable) and private (unverifiable) signals 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 signals according to how "spread out" they are around the efficient cutoff—i.e., the posterior belief above which it is efficient to act. Loosely speaking, a more spread out signal results in the posterior distributions being more extreme relative to the efficient belief. Importantly, the spreading order is a strengthening of the Blackwell order.

We show that a more spread out public signal can decrease welfare, while a more spread out private 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 private signal but harder with a more spread out public signal. We have chosen to present our results in the baseline model with a binary signal structure since elucidating the mechanism clearly is considerably easier in this particular setting.

Finally, we would like to highlight that the aforementioned generality of the mechanism implies that its insights apply to a variety of settings, even if no formal court is present or if information is not binary. We present and discuss several applications in Section 5 that highlight the versatility of our environment.

1.1 Related Literature

At a superficial level, the main takeaway of our paper—that superior verifiable information may reduce welfare—seems reminiscent of several different strands of literature. While our results are connected to a few relevant strands, we highlight in this section the substantive differences in the economic forces that lead to our results. To this end, we discuss each of these strands of the literature 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 or not 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.

Different from Bull and Watson (2019), we abstract away 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 chances that such an action will lead to (i) harm and (ii) conviction.

In our setting, the defendant's action is publicly observed, and the court's objective 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.[5]

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, 2017a,b) 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 ramifications of the chilling effect can be mitigated through the combination of superior information and an optimally chosen 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 on 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, since an erroneous interpretation of the law by courts may hurt the "wrong" part of the population. Pei and Strulovici (2019) show that a severe punishment reduces the number of crimes that witnesses report, thereby reducing the cost of committing a crime. Intermediate

[5]All four papers discuss their findings in light of Federal Rule 404, which concerns the exclusion/inclusion of character evidence in the trial. We wish to emphasize that, in our environment, all evidence of the agent's character comes from his behavior and not from other, observable character traits. Thus, the court in our model complies with Federal Rule 404.

punishments can deter some individuals from committing crimes, but those that commit some crime are likely to commit several crimes. In contrast, we concentrate on how the quality of information affects the trade-off 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 environments, consider the setting in which the private information is very precise. Due to coordination motives, small increases in the precision of public information can harm welfare in their environment if the public information is sufficiently noisy, as players overweight that information. In contrast, an analogous result does not exist in our setting because the agent can be screened on the outcomes if the private information is very precise in our world.

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. They show that sharing information may harm welfare. While outwardly similar, in our setting, it is the agent that possesses the information, whereas 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, in contrast, on an agent afraid of being punished for relying on all of his available information.

Contract Theory. The closest papers in the contracting literature are Prendergast (1993) and Prat (2005). In a principal-agent setting, Prendergast (1993) focuses on how to incentivize an agent to acquire relevant information at a cost. He highlights how the agent may focus more on acquiring information about the principal's prior belief about the underlying payoff-relevant state, rather than state itself.

Prat (2005) argues that the *content of information* leads to qualitatively different effects of increased precision. While information about consequences is beneficial, that about actions is harmful. We view our exercise as complementary to Prat (2005) and Prendergast (1993). While their focus was on situations in which the underlying information is about different objects (principal's prior in Prendergast (1993) and consequences vs. actions in Prat (2005)), both signals in our framework inform 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

An agent ("he") decides whether to undertake a risky project, which may succeed or fail. The agent is uncertain about the quality of the project, denoted by θ, and

must rely on the information available to him to make a decision. The court ("it"), having observed a failure, decides whether and how to sentence the agent. The court is bound by the penal code chosen by a committed designer (i.e., society) ex ante.

Project Quality and Information. The project's realized quality is either good ($\theta = 1$) or bad ($\theta = -1$). If undertaken, a project of bad quality always fails, and a project of good quality always succeeds.[6]

The informational environment of the project's quality, $S = (\alpha, p_x, p_y)$, consists of three elements. First, all players share a common prior, α, which denotes the ex ante probability that the project is good—i.e., $\theta = 1$.

The second element, the *precision* p_x of the *verifiable information*, denotes the ex ante probability with which the realization, $x \in \{-1, 1\}$, of a binary, *public* signal, $\boldsymbol{X}$, coincides with the true state θ—i.e., $p_x := \mathbb{P}(\boldsymbol{X} = \theta)$. With complementary probability, $1 - p_x = \mathbb{P}(\boldsymbol{X} = -\theta)$, the signal is misleading.

The third element, the *precision* p_y of the *unverifiable information*, denotes the ex ante probability with which the realization, $y \in \{-1, 1\}$, of a binary, *private* signal, $\boldsymbol{Y}$, coincides with the true state θ, $p_y := \mathbb{P}(\boldsymbol{Y} = \theta)$.

Signals are informative about the state but are noisy—i.e., $1 > p_x, p_y > 1/2$—and independent conditional on the state.[7]

Our assumptions that the two signals are (i) binary, (ii) conditionally independent, and (iii) symmetric simplify the model without sacrificing substance. All three can be relaxed without altering our results or the underlying intuition. We address the model under a general signal structure in Appendix C and models relaxing assumptions (ii) and (iii) separately in Appendices E.1 and E.2.

The Agent. The agent can be of two types, unbiased ($\omega = u$) or biased ($\omega = b$). The probability that the agent is unbiased is denoted by γ.

Each agent observes $\{\omega, x, y\}$, his type and the realization of both signals, and decides whether to undertake the project (henceforth act) ($a = 1$) or not ($a = 0$). The agents' utility is:

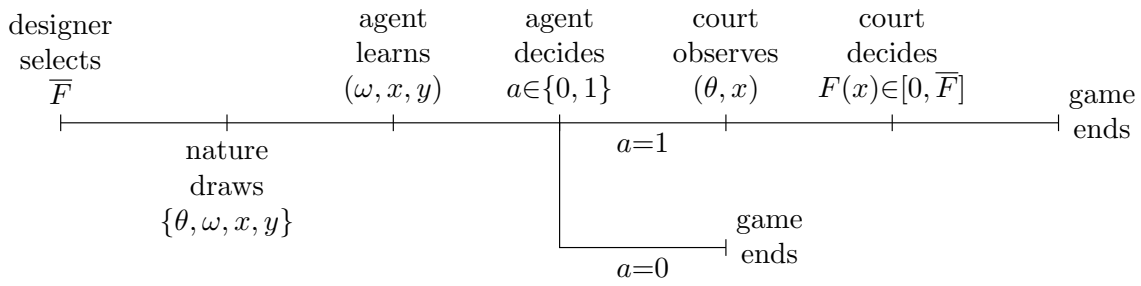
$$u^u(a; \theta) = a\theta, \quad \text{and } u^b(a, \theta) = a.$$

Unbiased agents benefit from successful projects but suffer from failed projects; biased agents benefit whenever they act. Besides that, an agent's payoff is affected by the court's decision about punishing the agent. If the agent is punished by the court to

[6] It will become clear that the assumption that project quality is a perfect predictor of the outcome is without further loss. Aggregate uncertainty shared by the agent and the court can be mapped into the verifiable and noisy signal x, the basis of the court's decision making.

[7] Henceforth, $\boldsymbol{X}$ and $\boldsymbol{Y}$, denote the random variables while x and y denote their respective realizations. For example, when we say $x + y = 0$, we mean the event $\{\boldsymbol{X} + \boldsymbol{Y} = 0\}$, or (x, y) denotes an ordered pair of observed signal realizations x and y. For example, we will denote $\mathbb{P}(\boldsymbol{X} = x, \boldsymbol{Y} = y|\theta)$ by $\mathbb{P}(x, y|\theta)$, etc..

Figure 1: Timing of the Game.



The designer selects the maximum punishment, $\overline{F}$. The agent observes his type realization, ω, and the ralization 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 th verifiable signal, x. If the project fails, $\theta = -1$, the court selects a punishment, $F(x) \in [0, \overline{F}]$. Then, payoffs are realized.

F, his gross utility is reduced by F. An agent's strategy, $a^\omega : \{-1, 1\}^2 \to [0, 1]$, is the probability that type ω acts on the information (x, y)—i.e., when $\boldsymbol{X} = x, \boldsymbol{Y} = y$.

The Court. After the agent makes his decision, and the outcome of the project is revealed, the court decides whether to punish the agent to $F \in [0, \overline{F}]$, where $F = 0$ means acquittal, while $\overline{F}$ is the highest possible punishment allowed by society. The *penal code*, the highest possible punishment $\overline{F}$, is chosen ex-ante to minimize inefficiencies in the agent's decision (see below).[8]

The court observes the agent's action, the project's outcome if implemented, and thus its quality θ. In addition, it observes the realization of the verifiable information x. If no project is undertaken, the quality remains unobserved.

Based on its information, $\{\theta, x\}$, the court decides on the sentence. We assume that courts follow basic legal principles—in particular, the "actus reus non facit reum nisi mens sit rea" (the act is not culpable unless the mind is guilty) doctrine.

On *actus reus* and *mens rea* Two assumptions follow from the above legal principle. The first assumption means that the court is constrained by '*actus reus.*' It can punish the agent only if there are damages. Damages occur if, and only if, the agent undertakes the project, $a = 1$, *and* the project fails, $\theta = -1$. This highlights that not acting is a 'safe' option for the agent. In reality, the reactions to a harmful action are, indeed, often more drastic than those to a harmful inaction, a phenomenon called the 'omission bias' (Baron et al., 1994). We emphasize that our results do not depend on this assumption. In Section 4.3, we extend the model to allow the court to also punish an agent for inaction and show that our results are unchanged.

[8] In reality, conviction may imply a stigma on the agent that is independent of the actual amount of the sentence. Since F reduces the agent's utility linearly, a more realistic description of the available sentences may, thus, be $F \in \{0\} \cup [\underline{F}, \overline{F}]$. To simplify notation, we choose to abstract from stigmata in the formal analysis. However, stigmata are straightforward to include. Our main results continue to hold.

According to the second assumption, the court aims to convict only agents with a '*mens rea*' (culprit mind). We focus on the concept of "subjective mens rea" in our model. That is, the court needs to be sufficiently convinced that the outcome is bad because the agent's preferences are not sufficiently aligned with society's. The court infers the agent's preferences from the information available to it.

An alternative interpretation would be "objective mens rea." In that case, the court aims to infer the agent's information set to decide the punishment. Historically, subjective mens rea has often been used in criminal law as a requirement for conviction, but even in tort law cases in which "mens rea" has less of a formal role, the defendant's (perceived) type plays a large role in establishing liability (see, e.g., Cane, 2000, for a discussion).[9]

Formally, the court receives a benefit equal to the sentence, F, if it convicts a biased agent, and a loss, FL, with parameter $L > 0$, if it convicts an unbiased agent with the sentence F. The court's (expected) utility from convicting an agent that is unbiased with probability $\tilde{\gamma}$ is

$$v(F) = F(1 - \tilde{\gamma}) - \tilde{\gamma} F L.$$

We interpret this tradeoff as a requirement of proving failure to adhere to the standards of liability: the court has to be adequately convinced that the agent is sufficiently careless; otherwise, it prefers to acquit the agent. We denote the court's strategy by $F : \{-1, 1\} \to [0, \overline{F}]$, meaning that the court convicts with sentence $F(x)$ when it sees $(a = 1, \theta = -1)$ and realization x.

Welfare and Solution Concept Our main objective is to understand how the different natures of the information affect decision making. We are, thus, interested in how the precision of the verifiable and unverifiable signals affects the ex-ante probability of implementing good and bad projects.

To that end, we define "welfare," denoted by $W(\cdot)$, as the ex ante expected efficiency gains to society (relative to the status quo).[10] The welfare measure coincides with the payoff of an unbiased agent.

More formally, we define

$$W(a^u, a^b, F, \overline{F}, S, \gamma) := \sum_{\theta \in \{-1,1\}} \mathbb{P}(\theta) \sum_{(x,y) \in \{-1,1\}^2} \mathbb{P}(x, y|\theta) \left[\gamma a^u(x, y) + (1 - \gamma) a^b(x, y) \right] \theta.$$

[9] We provide further discussion in Section 4.2. In presenting the baseline model, we focus on subjective mens rea for two reasons. First, it allows for a sharper description of the central tradeoff. Second, we show that, applying subjective mens rea leads to higher welfare. Thus, if a designer could choose between the appropriate concepts, she would select subjective mens rea.

[10] Notice that our notion of welfare does not take into account the payoff to the court or to the agent. This captures the idea that, in a large society, the fine amounts recovered are inconsequential, and it is only the payoff from the projects that matters.

The timing of the game is sketched in Figure 1: First, the designer chooses the maximum allowed sentence, $\overline{F}$. Then, the state of the world, (θ, ω, x, y), realizes. The agent observes (ω, x, y) and decides on action a. If the agent implements a project, the court observes (θ, x) and decides whether or not to convict the agent.[11] If it convicts, it also determines a sentence, $F(x) \leq \overline{F}$. Finally, given the focus on society's ability to set up the punishment scheme, we concentrate on the perfect Bayesian equilibrium that is optimal—i.e., closest to the (ex-post) efficient frontier.

3 Analysis

The analysis is done by backward induction. We first characterize the court's best response—given that the agent chose to implement the project, the project's realization, and the verifiable information. Next, we characterize the agent's best response—given the penal code, his type, and the information available to him. In Section 3.1, we characterize the optimal penal code, balancing the tradeoff of giving a free pass to the biased agent with the chilling effect that any threat of punishment may generate on the unbiased agent. Finally, Section 3.2 contains our main results.

Before proceeding, we want to emphasize that, given the misalignment of preferences between an unbiased and a biased agent, for a given realization of the signals, if the unbiased agent acts on some information, then the biased agent also acts on that information. Thus, if it is ex-ante unlikely that the agent is unbiased—formally, $\gamma < \gamma^* \equiv \frac{1}{1+L}$—the problem is trivial. Irrespective of the information, the court convicts all agents if the project is implemented and fails. In the remainder of the paper, we focus on the interesting case in which the ex-ante belief about the agent being unbiased is sufficiently high, $\gamma > \gamma^*$.

Court's Best Response. The court's information set is (a, θ, x). First, due to the actus reus constraint, the court may convict only if $a = 1$, and $\theta = -1$. Let $\tilde{\gamma}(x) := \mathbb{P}(\omega{=}b|\boldsymbol{X} = x, a{=}1, \theta{=}-1)$ be the probability that the court attaches to the agent being the unbiased type when the information set is $(a{=}1, \theta{=}-1, \boldsymbol{X} = x)$. The court wants to maximize $v(F) = F\left(1 - \tilde{\gamma}(x) - \tilde{\gamma}(x)L\right)$, and its best response is straightforward. It sentences $F = \overline{F}$ if $\tilde{\gamma}(x) < \gamma^* := \frac{1}{1+L}$ and acquits if $\tilde{\gamma}(x) > \gamma^*$. If $\tilde{\gamma}(x) = \gamma^*$, it is indifferent between any sentences.

Agent's Best Response. Assume that the court sentences $F(x)$ for a given realization (x, y) if the project fails. Further, let $\beta_{xy} := \mathbb{P}(\theta = 1|\boldsymbol{X} = x, \boldsymbol{Y} = y)$. Let $a^\omega(x, y; \overline{F})$ be the probability that an agent of type ω acts, given signal realizations (x, t), and penal code $\overline{F}$. Then, $a^\omega(x, y; \overline{F}) = 1$ only if

$$\beta_{xy} u^\omega(1; 1) + (1 - \beta_{xy})(u^\omega(1; -1) - F(x)) \geq 0.$$

[11] Our results remain unchanged if we assume that the court is, instead, the designer and commits ex-ante to a punishment scheme rather than judging at an interim level. We discuss the alternative model in detail in Section 4.

Therefore, the agent follows a cutoff strategy. Let

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

The agent acts if $\beta_{xy} > \overline{\beta}^\omega(F(x))$ and does not act if $\beta_{xy} < \overline{\beta}^\omega(F(x))$. Note that, if the unbiased agent acts for some (x, y), the biased agent acts, too, since $\overline{\beta}^u(F(x)) > \overline{\beta}^b(F(x))$.

3.1 Optimal Penal Code

Whether an action is expected to increase welfare, given the received information, depends on the interim belief β_{xy}. Given S, we say that it is *(interim) efficient to act* on (x, y) if the unbiased agent would act on (x, y) when $\overline{F} = 0$—that is, if $\beta_{xy} \geq 1/2$.[12] In most of the main text, we focus on the environments S such that

$$\beta_{xy} \geq 1/2 \Leftrightarrow x + y > 0.$$

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. This case enables us to demonstrate how improving verifiable information can lower welfare (Proposition X).[13]

Basic Tradeoff. Our goal is to identify the optimal penal code—the maximum sentence, $\overline{F}$, that the court should impose to maximize welfare. If $\overline{F}$ is low, the biased agent is given a *free pass*—he acts even if it is inefficient to do so. If $\overline{F}$ is high, the unbiased agent suffers from the *chilling effect*—the fear of being punished that results in not acting when it is efficient to do so. The optimal penal code balances deterrence of the biased agent with encouragement of the unbiased agent.

Define,

$$\overline{W}(\overline{F}, S, \gamma) := \sup_{(a^u, a^b, F) \in \mathcal{E}(\overline{F})} W(a^u, a^b, F, \overline{F}, S, \gamma)$$
$$W^*(S, \gamma) := \sup_{\overline{F} \in [0,\infty)} \overline{W}(\overline{F}, S, \gamma),$$

where $\mathcal{E}(\overline{F}) := \{(a^u, a^b, F) : (a^u, a^b, F)$ constitute an equilibrium given $\overline{F}, S, \gamma\}$.

Definition 1 *Fix some (S, γ). If there exists $(\overline{F}^*, a^u, a^b, F)$ such that $(a^u, a^b, F) \in \mathcal{E}(\overline{F}^*)$ and $W(a^u, a^b, F, \overline{F}^*; S, \gamma) = W^*(S, \gamma)$, then we say that the equilibrium is a "(designer) optimal equilibrium," and $\overline{F}^*$ is an "optimal penal code."*

[12] Recall that "to act on (x, y)" means to act on the event $\{\boldsymbol{X} = x, \boldsymbol{Y} = y\}$.

[13] Our results are not specific to this case. The possibility result of Proposition X is true whenever $y \geq 0$ is sufficient for interim efficiency; the general result Proposition Y holds in all the cases. For a formal treatment, see Appendix D.

Optimal Penal Code. First, notice that it is efficient to act when $x = 1$ regardless of the realization y, and, therefore, the optimal sentence $F(1) = 0$. Conflict arises when $x = -1$. Here, it is efficient to act on $(-1, 1)$ and to not act on $(-1, -1)$. We first consider the case wherein we give a *universal free pass* to the agent by setting $\overline{F} = 0$. The unbiased agent is going to follow the efficient schedule, while the biased agent always acts. In particular, $a^b(-1, -1; 0) = a^u(-1, 1; 0) = 1$. Therefore, $\tilde{\gamma}(-1) = \frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)}$. If, $\frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)} > \gamma^*$, the court will not convict the agent upon seeing a negative outcome and negative verifiable information, for any $\overline{F} > 0$.

We now turn to the cases in which the universal free pass is not optimal—that is, $\frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)} < \gamma^*$.

Recall that if the unbiased type acts on some (x, y) with positive probability, the biased type does so with probability 1. Moreover, the unbiased type (whose payoffs coincide with society's) will never act on $(-1, -1)$. On the other hand, the biased type will act on $(-1, -1)$ if the sentence is low. Similarly, the unbiased type will act on $(-1, 1)$, as he should, if the sentence is low, but not otherwise. Therefore, given a verifiable negative signal, the following two questions guide our analysis.

1. What is the minimum sentence, F^b, that prevents the biased type from acting when receiving a bad unverifiable signal?
2. What is the maximum sentence, F^u, that will allow the unbiased type to act when receiving a good unverifiable signal?

From the agent's best response, we obtain,

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

The threshold punishments above depend on the relevant parameters, (α, p_x, p_y). We now present our first result. All proofs are in Appendix A.

Lemma 1 *The optimal penal code, $\overline{F}^*$, exists and, without loss of generality, $\overline{F}^* \in \{0, F^u, F^b\}$.*

To see the intuition, first notice that $F(-1)$ affects both $a^u(-1, 1)$ and $a^b(-1, -1)$. If $F^u < F^b$, it is impossible to induce $a^u(-1, 1; \overline{F}) = 1$ *and* $a^b(-1, -1; \overline{F}) < 1$ through any $F(-1)$. The optimal penal code, therefore, either allows a universal free pass that implies that $a^u(-1, 1; 0) = a^b(-1, -1; \overline{F}) = 1$, or, by setting $\overline{F} = F^b$, deters both, $a^u(-1, 1; F^b) = a^b(-1, -1; F^b) = 0$. Which of the two is optimal depends on the ex-ante probability of the agent being biased.

If $F^u > F^b$, it is possible to partially deter the biased type from acting on $(-1, -1)$ with probability 1, but to have the unbiased type act on $(-1, 1)$. However, it is impossible to have $a^u(-1, 1; \overline{F}) = 1$ *and* $a^b(-1, -1; p^{FB}) = 0$. If that were the case, the court would not convict on $x = -1$, and the biased type would have an incentive to deviate to $a^b(-1, -1) = 1$. The optimum implies either partial deterrence or a moderate chilling effect. If $\overline{F} = F^b$, the biased type is partially deterred from acting

on $(-1,-1)$. He acts with probability $\eta^b > 0$. If $\overline{F} = F^u$, there is a moderate chilling effect, as the unbiased type acts with probability $\eta^u < 1$ on $(-1,1)$. When $x = -1$, the court is indifferent about conviction, and convicts to ensure indifference on the (relevant) agent's side.

Table 1 summarizes the four possible optimal equilibria in terms of agent strategy profiles. In the appendix, we characterize all equilibrium objects in terms of primitives.

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

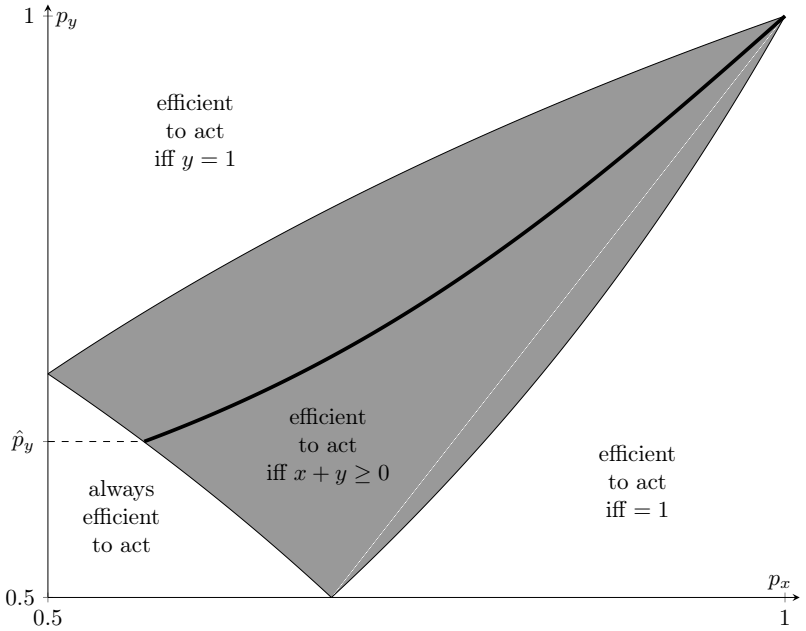
As we can see, the optimal equilibria are qualitatively different depending on how F^u and F^b are ranked. Since the focus of our paper regards comparative statics on p_x and p_y, which determine F^u and F^b, we now present the key step for our comparative static.

Lemma 2 *$F^b - F^u$ is continuous in p_x and p_y, and is increasing in p_x and decreasing in p_y.*

Intuition behind the proof: To understand why the difference is increasing in the precision p_x, first note that both F^b and F^u are decreasing in it. 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 different types in different ways. Due to 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 conviction effect). The unbiased agent suffers also directly—through the failure itself (the outcome effect).

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

Figure 2: Critical values of the quality of information



The shaded area is the parameter region (p_x, p_y) for 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 α and it is always efficient to act. The thick black line depicts the beliefs at which $F^b = F^u$, (p_x^*, p_y^*). Changes in p_x represent movements parallel to the $X-$axis, changes in p_y movements parallel to the $Y-$axis. Welfare drops for horizontal moves crossing the black line (see Figure 3 below). In this example, $\alpha = 9/13$.

3.2 Signal Precision

Suppose that the verifiable information becomes more precise; that is, p_x increases to some $p_x' > p_x$. By Lemma 2, the difference between the cutoff punishments, $F^b - F^u$, can switch its sign. Holding the other relevant parameters constant, we define a critical threshold of the precision of the verifiable information if, at that value, the cutoff punishments are equal, and it is efficient to act if either of the signals is positive. Formally, given (p_y, α), say that a precision p_x^* is a critical threshold of information quality if

1. $F^b(p_x^*, p_y, \alpha) = F^u(p_x^*, p_y, \alpha)$ (see (1));
2. (p_x^*, p_y, α) is in the interior of environments where it is efficient to act if and only if $x + y \geq 0$.

The shaded region in Figure 2 depicts the (p_x, p_y) region such that it is efficient to act if and only if $x + y \geq 0$. The thick black line plots the critical information quality, $p_x^*(p_y)$, for a fixed α. Notice that when $p_y < \hat{p}_y$, a p_x that makes $F^b = F^u$ falls outside the shaded region, and, hence there is no critical belief when $p_y < \hat{p}_y$. However, as the figure shows, neither our baseline case nor the (p_y, α) region that admits a critical

belief is knife-edge.[14]

Figure 2 reinforces why we focus on the case in which it is efficient to act if, and only if, one of the signals is positive—i.e., $x + y \geq 0$. This is the case if the prior is high and both signals are informative to a similar extent. In that case, either signal being positive sways the posterior belief of an unbiased agent towards the risky action if there is no threat of punishment. As we can see, the critical threshold of information quality occurs precisely in this area, where the information quality of both signals is similar.

With some abuse of notation, let $W^*(p_x)$ denote $W^*(S, \gamma)$ for a fixed p_y, γ, α. Similarly, $W^*(p_y)$ stands for $W^*(S, \gamma)$ fixing p_x, γ, α.

Proposition X *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 belief given (p_y, α), then $\exists \epsilon > 0$ such that, $W^*(p_x) > W(p_x') \ \forall \ p_x \in (p_x^* - \epsilon, p_x^*), p_x' \in (p_x^*, p_x^* + \epsilon)$.*

Proposition X is driven by the sign change of $F^b - F^u$ around p_x^* as described in Lemma 2. For example, suppose that γ is high and p_x is slightly below p_x^*. Here, since $F^b < F^u$, the optimal equilibrium is as in Table 2b. In particular, it is possible to have the biased agent act with probability less than one on $(-1, -1)$ while 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 while having the unbiased agent act with a positive probability on $(-1, 1)$. Therefore, we are left with two options. Either we can give a universal free pass, or we can achieve partial deterrence of the biased agent at the cost of a chilling effect on the unbiased agent.

While 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 terms of welfare. Yet, the effect of such an improvement is continuous in p_x, while 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 this is a local comparative static. A sufficiently large increase of p_x obviously 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.

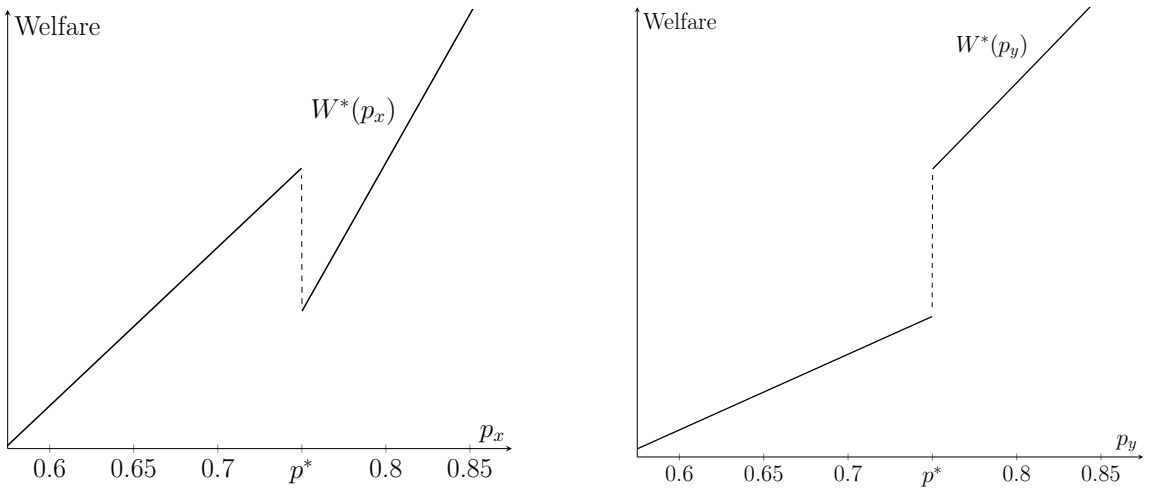
[14] The reason to emphasize this aspect is Proposition X: welfare declines when we move from $p_x^* - \epsilon$ to $p_x^* + \epsilon$ for a small $\epsilon > 0$. Also, while the figure does not vary α, continuity in α and, therefore, the claim of non-knife-edgedness is obvious.

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

Figure 3: *Welfare for different precision levels.*



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

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

Welfare 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 if and only if $x + y \geq 0$. Also, the penal code, $\overline{F}$, is chosen optimally throughout. Parameters: $\gamma^* = 1/2, \gamma = 11/20, \alpha = 9/13$ and $p_y = 3/4$ (left panel), $p_x = 3/4$ (right panel).

The main difference between p_x and p_y, and the driver of our results, is in their effect on $F^b - F^u$. While $F^b - F^u$ is increasing in p_x, it is decreasing in p_y. To better understand the source of this difference, we revisit the discussion in the introduction. As mentioned there, the main conflict in our environment is that, at times, we want the unbiased agent to decide in favor of implementing the project by going against a negative $\boldsymbol{X}$, the verifiable information, and relying only on $\boldsymbol{Y}$, the unverifiable information. However, the associated cost is that the biased agent may implement a project when both $\boldsymbol{X}$ and $\boldsymbol{Y}$ recommend not doing so. Increasing the precision of $\boldsymbol{Y}$ helps the designer here. At once, it makes the unbiased agent more confident about implementing the project by trusting a positive $\boldsymbol{Y}$, whilst it discentivizes the biased agent from implementing it despite a negative $\boldsymbol{Y}$. Thus, welfare increases.

In contrast, increasing the precision of $\boldsymbol{X}$ disincentivizes both types on $x = -1$. It increases the threat of punishment, as it indicates a higher chance of failure and, thus,

of punishment if the project is implemented. 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 implement, given a negative $\boldsymbol{X}$, go down in the signal's precision, regardless of the punishment. Due to 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 upwards. This results from the fact that, as the precision of the unverifiable information exceeds the critical threshold, the punishment needed to deter the biased agent creates, at most, a partial chilling effect on unbiased agents. Given the parameters of Figure 3, as the precision of the unverifiable information increases, 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

In this section, we highlight 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 the court aimed to punish the agent for acting against his better knowledge ('objective means rea')? Next, we consider punishment for inaction: what if the court can punish the agent for not acting? Finally, we extend the model to more than binary types of agents.

4.1 A Contracting Model

In this section, we discuss the fundamental force behind our results. To this end, we temporarily leave the application with its three players (designer, agent, court) and consider an abstract model between 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, before the agent has taken his action, the principal designs and commits to punishments for each of the possible realizations. Second, the principal decides, without constraints on the punishment, *ex post*—that is, after the agent has taken his action (the "ex post screening" case). We show that our result is strengthened in both cases.

Commitment. There is a principal (she) and an agent (he). Nature moves first and draws θ, ω, x, y according to a commonly known informational environment, S. The principal observes the realization of $\boldsymbol{X}$. Thereafter she commits to a punishment $F : \text{supp}(\boldsymbol{X}) \to \mathbb{R}_+$. The agent observes F, his type ω, and the realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. 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, both the size of the punishments from (1) and Lemma 1 apply. Thus, given S, F^b and F^u remain the same as in our baseline model, and it remains without loss to consider $F \in \{F^b, 0\}$ when $F^b > F^u$ and $\overline{F} \in \{F^b, F^u\}$ when $F^b < F^u$, which, in turn, implies that Lemma 2 holds.

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.

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

On the one hand, if S is such that $F^u > F^b$, committing to $F(-1) = F^b$ guarantees that the agent takes the (interim) efficient action regardless of his type. On the other hand, if S is such that $F^b > F^u$, the payoffs are identical to those in the baseline case. The unbiased agent is too pessimistic about the state and, thus, 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 or to avoid the chilling effect at the expense of no deterrence. We lose interim efficiency.

As the environment changes, from $F^b < F^u$ to $F^b > F^u$, welfare suffers a discrete loss due to the inability to implement the interim efficient action, which outweighs the marginal gain from better information. Because Lemma 2 applies, we move from the bottom 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.

The commitment case strengthens our substantive results. The only difference from the baseline model is that welfare in improves if $F^b < F^u$. Thus, commitment does not improve the outcome when p_x is relatively precise (top row of Table 2), but it improves the outcome if p_x is imprecise (bottom row of Table 2). In the latter case, commitment implies that the principal can align the preferences of the biased agent fully and at no cost to the unbiased agent. Thus, if welfare decreases when p_x increases to p_x' in the baseline case, the same is true in the commitment case.

Similarly, because welfare increases for any increase p_y to p'_y in the baseline case, which implies (at most) that we switch from the top row to the bottom row, welfare increases in the commitment case, too.

Ex Post Screening. There is 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 his type ω and the realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$. 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. Like the court in the baseline setting, the principal receives a benefit of 1 if she punishes a biased agent and suffers a loss L if she punishes an unbiased agent.

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

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

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

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 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 due to the nonexistence of a designer

to limit the punishment ex-ante. The bottom row, instead, yields the same welfare as the baseline case. We see that society's ability to limit the punishment is beneficial, particularly in the case in which the verifiable signal is precise. If verifiable information is precise and a principal screens ex-post, only full deterrence and full chilling are feasible, a situation worse than those in which the punishment is limited.

Relationship to the Baseline. The commitment case corresponds to strict liability in the legal setting. In certain situations—e.g., if the realization of the public information is some specific x—an action is 'per se' illegal. 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, e.g., the agent gets banned from continuing his job. In some of the examples we discuss in Section 5, such an exogenous punishment appears appropriate.

4.2 Objective Mens Rea

In this section, we directly 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.

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. Under this assumption, the court wants to convict a "guilty mind"; i.e., 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. The legal philosophy calls 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. And the court wants to punish the agent for acting when the available information indicated that he should have exercised restraint. The legal philosophy calls this notion *objective mens rea.*

While 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.[15]

In addition, discrimination lawsuits also use type attributes to prove intentional

[15]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.

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),[16] the court (over)ruled that the 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 on the basis of the "discriminatory attitude" of the chief as "'direct evidence' of discriminatory animus."[17]

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. The second reason is that, as discussed above, the courts seem to employ subjective mens rea regularly in and outside of criminal law.[18] 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 the he took an action, $a = 1$, which 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—i.e., the agent's signal was $(-1, -1)$. That is, under objective mens rea the court punishes if $q := \mathbb{P}(\boldsymbol{Y} = 1 | \theta = -1, a = 1, \boldsymbol{X} = -1) \leq \gamma^*$.

Fixing all the parameters, including the loss L from punishing wrongly, we obtain our first result.

Proposition MR *Society's ex-ante expected welfare in the optimal equilibrium is weakly higher if the court employs subjective mens rea than if it employs objective mens rea.*

The intuition underlying Proposition MR 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

[16] See https://m.openjurist.org/55/f3d/126/wilson-v-susquehanna-township-police-department-1.

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

[18] We want to emphasize that we do not claim that subjective mens rea is employed more or less often than objective mens rea.

on $(-1,-1)$ is η_2, which is larger than η^b, used in the baseline case. These three properties imply Proposition MR.

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

We now present the effects of changes in information quality for the alternative specification of the court's objective.

Unlike in the baseline model, welfare may decline upon improving the precision of $\boldsymbol{Y}$, the unverifiable information, when the court adopt objective mens rea. If at all 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.

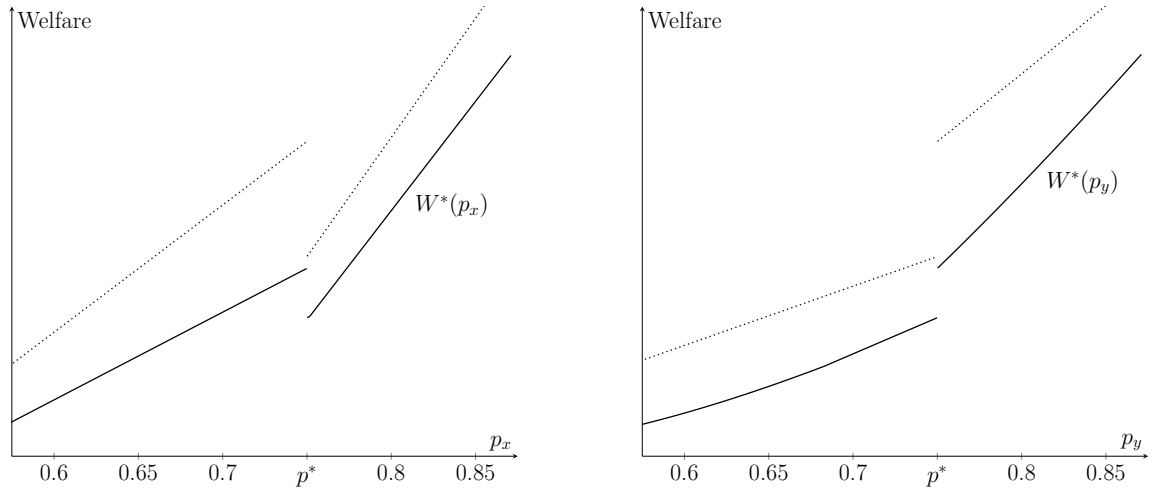
Other than at the critical level p_y^*, welfare is increasing in p_y. The following proposition characterizes when the positive effect of increasing p_y that mitigates the chilling effect outweighs the negative effect of reduced deterrences. To keep the formulations compact and to maintain focus, we introduce some additional notation.

Recall that $W^*(S,\gamma)$ denotes the welfare in the optimal equilibrium given the information structure $S = (\alpha, p_x, p_y)$. Let $\mathcal{Y}(\alpha, p_x)$ be the set of p_y's such that it is efficient to act iff $x + y \geq 0$ when the precision of $\boldsymbol{X}$ is p_x, and that of $\boldsymbol{Y}$ is p_y. Notice that $\mathcal{Y}(\alpha, p_x)$ is compact, and hence, we can define $\underline{p_y}(\alpha, p_x) := \min \mathcal{Y}(\alpha, p_x)$ and $\overline{p_y}(\alpha, p_x) := \max \mathcal{Y}(\alpha, p_x)$.

Proposition O *Suppose that it is efficient to act if and only if $x + y \geq 0$ and the court employs objective mens rea.*

1. ***Precision of $\boldsymbol{X}$**: Fix a (p_y, α) such and let p_x^* be the associated critical belief. Then, $\exists \epsilon > 0$ such that $W^*(\alpha, p_x, p_y, \gamma) > W^*(\alpha, p'_x, p_y, \gamma)$ for all $p_x \in (p_x^* - \epsilon, p_x*)$ and $p'_x \in (p_x^*, p_x^* + \epsilon)$.*
2. ***Precision of $\boldsymbol{Y}$**: Fix (α, p_x) such that $\mathcal{Y}(\alpha, p_x) \neq \emptyset$. Let p_y^* be the associated critical belief. If $p_y^* \notin \mathcal{Y}(\alpha, p_x)$, then $W^*(\alpha, p_x, \cdot, \gamma)$ is increasing on $\mathcal{Y}(\alpha, p_x)$. If*

Figure 4: *Welfare when the court aims to convict only agents that acted against better information.*



(a) Welfare changes in precision of **X**

(b) Welfare changes in precision of **Y**

Welfare 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: $\gamma^* = 1/2, \gamma = 11/20, \alpha = 9/13$ and $p_y = 3/4$ (left panel), $p_x = 3/4$ (right panel).

$p_y^* \in \mathcal{Y}(\alpha, p_x)$, *then*

(a) $W^*(\alpha, p_x, \cdot, \gamma)$ *is continuously increasing on* $[\underline{p_y}(\alpha, p_x), p_y^*)$ *and on* $(p_y^*, \bar{p_y}(\alpha, p_x)]$.[19]

(b) $W^*(\alpha, p_x, \cdot, \gamma)$ *is increasing at* p_y^*, *i.e.,*
$\lim_{p_y \uparrow p_y^*} W(\alpha, p_x, p_y, \gamma) \leq \lim_{p_y \downarrow p_y^*} W`*(\alpha, p_x, p_y, \gamma)$, *iff*

$$\frac{1-\gamma}{\gamma}(\eta_2(p_y^*) - \eta_1(p_y^*)) \leq \frac{\alpha p_y(1-p_x) - (1-\alpha)p_x(1-p_y)}{(1-\alpha)p_x p_y - \alpha(1-p_x)(1-p_y)} \tag{2}$$

Remark 1 *That is, the optimal welfare is increasing in* p_y *at all levels of precision except* p_y^*. *Equation* (2) *provides the condition that characterizes when the welfare is globally increasing in* p_y.

There are two economically interpretable sufficient conditions that arise from (2). First, if γ—the proportion of the unbiased agents in the society—is sufficiently large, then the society prefers not to deter the relatively fewer biased types from acting on $(-1, -1)$. The reason is the associated cost of chilling of 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 Y.

Another sufficient condition relates to the tolerance of the court. If $1 - \gamma^* > p_y^*$—i.e., the court convicts when it is sufficiently confident that the agent acted on

[19] We say that a function $f(z)$ is increasing on $[z_1, z_2]$ and on $[z_3, z_4]$ if $f(z) \geq f(z')$ whenever $z \geq z'$ and $z, z' \in [z_1, z_2]$ or $z, z' \in [z_3, z_4]$.

$(-1,-1)$—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.

4.3 Punishment for Inaction

Throughout the paper, we have assumed that the court can punish the agent only when $a = 1$ and $\theta = 0$. Our choice is motivated mainly by realism (for recent experimental evidence, see Cox et al., 2017).[20]

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 or not 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, unfortunately, 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 with strictly positive probability, a biased agent finds it optimal to act. Therefore, for any realization, inaction only increases the likelihood of the agent being unbiased, and the court's posterior over the agent being unbiased must be weakly higher than its prior. Because the prior $\gamma > \gamma^*$ by assumption, the court chooses to not punish the agent for inaction, even if allowed to do so.

4.4 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 society'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, while type K is the unbiased agent. Let μ_k denote the ex-ante probability of the agent being of type k. Even in this environment, the essential problem we face is the same: we want to define a penal code that incentivizes agents from acting on $(-1, 1)$ while disincentivizing the agents from acting in $(-1, -1)$. Given any F, the following fact is immediate:

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

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 X and Y. To see why, recall that the main driver of those results is Lemma 2, which established that $F^b - F^u$

[20] There is a debate over whether not punishing on inaction stems from a cognitive bias or from rational behavior (for an overview, see Woollard, 2019).

is increasing in p_x and decreasing in p_y. Similar 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

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

Therefore, $F_{-1}^{K^{-1}} - F_1^{K^1}$ is increasing in p_x and decreasing in p_y, as in Lemma 2. As a consequence, both Propositions X and Y 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 binary types. Similarly, the effect of increasing p_y is also identical to that in our main model.

5 Applications

Before concluding, we wish to illustrate some specific settings to which our model applies within and outside the formal legal system. Although *actus reus* and *mens rea* are not formally defined in some of these cases, they are explicitly or implicitly followed.

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

Second, consider a doctor deciding on the delivery method of a child without obvious complications. To decide, the doctor relies on some verifiable information (e.g., examinations indicating the child's position in the womb) and on some unverifiable information (e.g., his expertise and how he feels the child inside the womb). While C-Sections might be the best method in some cases, choosing an unnecessary C-Section risks dire consequences for the mother and the child. Doctors value money and their own time differently, and C-sections are scheduled and pay substantially higher. If a C-Section leads to complications, and if there are no verifiable evidence supporting its choice, the doctor may face legal and administrative consequences. In such a case, if the examiner (court or hospital administration) conludes 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.

As a third example, consider a president (the agent) deciding on a foreign policy issue: for instance, the decision of whether to impose sanctions on a country in response to an invasion of a neutral country. For the president, this is a risky decision, as his electoral chances in the future 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 interests groups.

The decision is made based on top-secret information (unverifiable information), as well as on what is being reported in the news (verifiable information). Voters (the court) might hold the president accountable and might want to re-elect him only if he values their interests above those of particular groups. However, voters have access only to the outcome and the verifiable information.

Finally, consider a CEO of a firm deciding whether or not to acquire a smaller firm. The acquisition is risky and may affect the firm's value and stock prices, and the CEO's compensation package and future remuneration from other firms may depend on its outcome. The CEO relies on hard information about the firm to be acquired, as well as on private information about the synergies between the firms and the general outlook of the market, to decide whether to proceed with the acquisition. 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, while they may wish to retain one that is interested in the long-run development of the firm.

6 Conclusion

This paper highlights that it is not merely the amount of information, but also its nature, that has important welfare consequences. As the main conflict studied here is between the agent and the court that tries to discipline his behavior, we focus on information of two different natures, depending on whether it is verifiable in court. We show that increasing the information available to the agent might have different consequences, depending on its nature. While increasing the precision of unverifiable information always increases welfare, increasing the precision of verifiable information may have dire consequences.

Our findings directly extend to a variety of settings well beyond legal systems. Whether we consider politicians seeking reelection, CEOs wanting to extend their contracts, or public officials with career concerns, our results apply whenever the principal's ex-post evaluation of a risky decision is based only on parts of the information available to the agent. The principal has to balance the chilling effect against a free pass, and changes in the information structure influence the ability to do so. We show that our findings are robust to a variety of assumptions: while details in the timeline

or the principal's choice set may differ, the main message remains. The welfare effects of a change in the precision of the information differs in the underlying nature of that information.

The main driver behind our mechanism—the chilling effect—has been extensively documented in the legal and management literatures, as well as in the popular press.[21] Our results show that, whether the chilling effect is pronounced enough to overturn informational gains 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.

A Main Results: Proofs

A.1 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 α 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 α is low and that a necessary and sufficient condition for efficiency is that *both* signal realizations are positive.

Notation. Fix $\overline{F}$. Let $q^u := a^u(-1,1;\overline{F})$ and $q^b := a^b(-1,-1;\overline{F})$. 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,

$$\frac{\gamma(1-p_y)}{\gamma(1-p_y) + (1-\gamma)(1-p_y+p_y\eta^b)} = \gamma^* \tag{3}$$

$$\frac{\gamma\eta^u}{\gamma\eta^u + (1-\gamma)} = \gamma^* \tag{4}$$

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

A.2 Proof of Lemma 1

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

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

Proof. $q^b = 0 \implies \tilde{\gamma}(-1) = \gamma > \gamma^*$. Therefore, $F(-1) = 0$. Therefore, D would deviate to play $q^b = 1$. Therefore, $q^b > 0$. Also, $q^b = 1 \implies \tilde{\gamma}(-1) = \frac{\gamma(1-p_y)}{\gamma(1-p_y)+1-p_y} < \gamma^*$. Therefore, $F(-1) = \overline{F}$. Notice that $\overline{F} < F^b \implies q^b = 1$. And, if $\overline{F} \geq F(-1) > F^b \implies q^b = 0 \implies \tilde{\gamma}(-1) = \gamma > \gamma^*$. This would imply that $F(-1) = 0$, a contradiction. Therefore, if $\overline{F} > F^b$, the D type would mix to have $\tilde{\gamma}(-1) = \gamma^*$—i.e., $q^b = \eta^b$, 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 citizen preferred equilibrium. □

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

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

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

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

If $F(-1) \in (F^b, F^u)$, then $q^u = 1$ and $q^b = 0$. But then, $\tilde{\gamma}(-1) = \gamma > \gamma^* \implies F(-1) = 0$, 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, $\tilde{\gamma}(-1) > \gamma^*$, 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,

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

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

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

A.3 Proof of Proposition X and Lemma 2

Now we are equipped to present our main comparative static. To this end, let $W^*(p_x, p_y, \gamma, \alpha) := \overline{W}(\overline{F}^*)$ denote the optimal equilibrium given the signal structure (p_x, p_y). Let $\delta(p_x, p_y) := F^b - F^u$.

Proof of Lemma 2.

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

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

Proof of Proposition X. At p_x^*, $F^b(p_x^*, p_y, \alpha) = F^u(p_x^*, p_y, \alpha)$. Henceforth, we will suppress the dependence on (p_y, α).

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 2.

Case 1: $\overline{F}^*(p_2) = 0$.[22]

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 (3) features no dependence on p_1 and η^b is strictly less than 1.

Let $W_1 := \overline{W}(F^b(p_1))$ and $W_2 := \overline{W}(0) = W^*(p_2, p_y, \gamma, \alpha)$.

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

Therefore,

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

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

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

Since it is inefficient to act on $(-1,-1)$, $\beta_{-1,-1} = \frac{\alpha(1-p_y)(1-p_1)}{\alpha(1-p_y)(1-p_1)+(1-\alpha)p_y p_1} < \frac{1}{2}$. Equivalently, $(1-\alpha)p_y p_1 > \alpha(1-p_y)(1-p_1)$. Therefore, $W_1 > W_2$. Lastly, if $\overline{F}^*(p_1) = F^u(p_1)$, then $W^*(p_2, p_y, \gamma, \alpha) \geq W_1 > W_2 = W^*(p_2, p_y, \gamma, \alpha)$.

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* $q^u(p_1) = \eta^u > 0$. Since 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, $W^*(p_1) > W^*(p_2)$ when $p_1 = p_2 - \delta$ for a small δ. □

A.4 Proof of Proposition Y

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 the online appendix.

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 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.

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

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. Since η^b decreases in p_y, the biased agent's actions on y improve from an efficiency perspective, while the unbiased agent's decisions can only improve by Lemma 2. The welfare increase. 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 since, by construction, $\overline{F} = F^u$ dominates $\overline{F} = F^b$, and the proof is complete. □

B Objective Mens Rea: Characterisation and Proofs

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

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

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

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

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

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

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, an 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 sub-optimal. No punishment yields a better outcome than the 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 MR. 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 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.[23] Since acting is inefficient for the information $(-1,-1)$, the alternative objective is welfare-inferior. If $F^b > F^u$, the agent chooses

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

Again, the alternative objective is welfare-inferior.

Proof of Proposition O. 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 X. We omit it, as it provides no further insight. We discuss the second part below.

When is welfare unambigously increasing in the precision of p_y?

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

Using a similar argument $\overline{W}^*(\alpha, p_x, \cdot, \gamma)$ is continuously increasing on $(p_y^*, \overline{p_y}(\alpha, p_x]$. 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 society 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

$$\alpha\Big(p_y(1-p_x) + (1-\gamma)(1-p_x)(1-p_y)\Big) - (1-\alpha)\Big(p_x(1-p_y) + (1-\gamma)p_xp_y\Big) >$$
$$\alpha\Big(p_y(1-p_x)(1-\gamma) + (1-\gamma)(1-p_x)(1-p_y)\eta_1\Big) - (1-\alpha)\Big(p_x(1-p_y)(1-\gamma) + (1-\gamma)p_xp_y\eta_1\Big)$$

which can be simplified to

$$\frac{1-\gamma}{\gamma}(1-\eta_1) > \underbrace{\frac{\alpha p_y(1-p_x) - (1-\alpha)p_x(1-p_y)}{(1-\alpha)p_xp_y - \alpha(1-p_x)(1-p_y)}}_{:=\widehat{\Delta}(p_y)} > 0 \tag{5}$$

[23] 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".

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

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

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,

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

or equivalently

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

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

$$\frac{(1-\gamma)}{\gamma}(\eta_2-\eta_1) \leq \underbrace{\frac{\alpha(1-p_x)p_y^* - (1-\alpha)p_x(1-p_y^*)}{(1-\alpha)p_xp_y^* - \alpha(1-p_x)(1-p_y^*)}}_{=\widehat{\Delta}(p_y^*)}. \tag{6}$$

Notice that if condition (6) 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 (6) implies (5). 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 γ^*. Moreover,

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

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

Thus, independent of γ^*, (6) holds if

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

which can be simplified to

$$\frac{1-\alpha}{\alpha}\frac{p_x}{1-p_x} \leq \frac{1-\gamma(1-p_y^*)}{1-p_y^*\gamma}. \tag{7}$$

Because $p_y > 1/2$ the RHS of the above larger than 1 which in turn implies that if $\alpha > p_x$, then Proposition 2 holds for any (γ, γ^*).

Moreover, $p_y > 1/2$ implies that the RHS of condition (7) is increasing in γ with limit $p_y/(1-p_y)$ as $\gamma \to 1$

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

which holds because we are in the case in which a any positive signal makes it efficient to act. Thus, for any $(\alpha, p_x, p_y, \gamma^*)$ 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 (7) fails, there always is a threshold $\underline{\gamma}^* > 1 - p_y^*$ such that Proposition 2 holds if $\gamma^* < \underline{\gamma}^*$. The reason is that for γ^* low enough $\eta_2 - \eta_1 = 0$.

C General Information Structure

Throughout the paper, we have assumed that the verifiable and the unverifiable signals are binary. Here we address how our results carry over to more general information structures.

We proceed as follows. First, we discuss how the model translates from the binary case to the more general case. Second, we provide a notion of being *"spread out"* and discuss how it is a meaningful notion of precision in our context. The spreading order is stronger than the, more familiar, Blackwell order of informativeness. That is, if one distribution is more spread out than another, it is also Blackwell more informative. The reverse, however, does not hold. In the spirit of our results from the baseline model, we then show how a more spread out verifiable information may lead to lower welfare. In comparison, a more spread out unverifiable information always increases welfare. Finally, we show that, the same does not hold for the weaker order of Blackwell informativeness. In particular it is not true that *all* Blackwell better unverifiable information (weakly) improves welfare.

Model. Let $\boldsymbol{\theta} \in \{-1, 1\}$ be a binary random variable describing the state of the world as before.[24] In the spirit of our baseline model, there are two signals. The public signal is a $[0, 1]$-valued random variable denoted by $\boldsymbol{X}$. $\boldsymbol{X}$ captures the posterior probability that $\boldsymbol{\theta} = 1$, i.e., $\boldsymbol{X} = \mathbb{P}(\boldsymbol{\theta} = 1|\boldsymbol{X})$. Rather than modelling the private signal, $\boldsymbol{Y}$, separately, we denote the agent's total information, again a $[0, 1]$-valued random variable, by $\boldsymbol{\mu}$. That is, given the agent's information, $\boldsymbol{\mu} = \mathbb{P}(\boldsymbol{\theta} = 1|\boldsymbol{X}, \boldsymbol{Y})$.

[24] Unless stated explicitly, bold faces denote random variables and normal fonts denote their realizations.

Let $G_{\boldsymbol{X}}(\cdot)$ denote the CDF of $\boldsymbol{X}$ and let $G_{\boldsymbol{\mu}_{|x}}(\cdot)$ denote the conditional CDF of $\boldsymbol{\mu}$ given $\boldsymbol{X} = x$. With some abuse of notation, we will denote by $\boldsymbol{\mu}_{|x}$ a random variable with a CDF $G_{\boldsymbol{\mu}_{|x}}$. Finally, let $G_{\boldsymbol{\mu}}(\cdot)$ denote the unconditional CDF of $\boldsymbol{\mu}$. Bayes plausibility implies that $\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{\mu}] = \alpha$. Moreover, $\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{X}] = \boldsymbol{X}$. Let the space of all public signals be denoted by $\mathcal{X}$. Since we want to demonstrate the results in the spirit of our main model for richer information structures, let $\mathcal{I}$, be the space of signal structures $(\boldsymbol{X}, \boldsymbol{Y})$ such that $G_{\boldsymbol{\mu}}(\cdot)$ has no mass points. Unless stated otherwise, we assume that $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{I}$.

First, since $\boldsymbol{X}$ is a public signal, the optimal punishment can condition on the realization x. That is, we can treat the problem for each realization of $\boldsymbol{X}$ separately. Given a punishment $F(x) \in \mathbb{R}_+$, let $\mu^\omega(F(x))$ denote the critical posterior—the posterior above which an agent of type ω acts. We obtain the following.

$$F(x) = \frac{1}{1 - \mu^b(F(x))} = -2 + \frac{1}{1 - \mu^u(F(x))}$$

In general, if some punishment implies a critical posterior of μ^u for the unbiased type. With some abuse of notation, the critical posterior, μ^b, for the biased type is given by,

$$\mu^b(\mu^u) = \frac{1}{2}\left[3 - \frac{1}{2\mu^u - 1}\right] \tag{8}$$

In the spirit of our main model, for $\omega \in \{u, b\}$, define

$$F^\omega(\mu) := -2\mathbb{1}_{\omega=u} + \frac{1}{1-\mu}.$$

As in the baseline model, the designer's problem is separable in the realization of $\boldsymbol{X}$. That is, the designer can choose a punishment $F(x)$ for each realization of $x \in \text{supp}(\boldsymbol{X})$ to provide incentives to both types. Therefore, as in the baseline model, the designer's problem is to choose a function $F : [0, 1] \to \mathbb{R}_+$ to maximize welfare.

In the baseline model, we demonstrated how the welfare effects of improving the precision of information depend on the nature of the information. While capturing precision through a scalar for each signal was natural in the case of binary signals, there does not seem to be a natural analog of the same in the case of richer signals. However, motivated by our setting, the spread of information (made precise momentarily) captures the main ideas in the baseline model. It highlights the difference between verifiable and unverifiable information. Recall that, from the perspective of a risk-neutral society, the decision rule in a binary action space is simple: Act if the project is more likely to be good, $\mu > 1/2$, do not act if the project is likely to be bad, $\mu < 1/2$.

Spread of a signal around $\frac{1}{2}$ captures the degree of certainty in each of the two cases. An increase in the spread means that the mass attributed to posteriors far from $\mu = 1/2$ increases. A simple way to formalize this is to think about "quantile-preserving" spreads around a posterior $\hat{\mu}$.

Definition 2 (Spread.) *We say that $\boldsymbol{X}$ is* "more spread out" *around $\hat{\mu}$ than $\boldsymbol{X}'$ if $G_{\boldsymbol{X}}(a) \geq G_{\boldsymbol{X}'}(a)$ for all $a \leq \hat{\mu}$ and $G_{\boldsymbol{X}}(a) \leq G_{\boldsymbol{X}'}(a)$ if $a > \hat{\mu}$.*

We say that $\boldsymbol{Y}$ *is "more spread out" around* $\hat{\mu}$ *than* $\boldsymbol{Y}'$ *if* $\boldsymbol{\mu}_{|x}$ *is more spread out around* $\hat{\mu}$ *than* $\boldsymbol{\mu}'_{|x}$ *for a.e.* x.[25]

As we are interested in the spread around 1/2, we shorten notation using $\succeq_{sp}$ to denote more spread out around $\hat{\mu} = 1/2$.

Remark 2 *Notice that, while the measure of spread of* $\boldsymbol{X}$ *only considers the distribution of* $\boldsymbol{X}$*, the same for* $\boldsymbol{Y}$ *is defined for each realization of* $\boldsymbol{X}$*. This is motivated by the fact that the designer's problem is separable across the realizations of* $\boldsymbol{X}$*, and therefore, the measure of spread of the private information conditions on these realizations. Morever, notice that, if* $\boldsymbol{Y}$ *is more spread out around* $\hat{\mu}$ *than* $\boldsymbol{Y}'$*, then* $G_{\boldsymbol{\mu}}(a) \geq G_{\boldsymbol{\mu}'}(a)$ *for all* $a \leq \hat{\mu}$*, and* $G_{\boldsymbol{\mu}}(a) \leq G_{\boldsymbol{\mu}'}(a)$ *for all* $a \geq \hat{\mu}$*.*

We will now see that, in the spirit of the main results from our baseline model, a more spread out (around $\frac{1}{2}$) $\boldsymbol{X}$ can reduce welfare, while a more spread out (around $\frac{1}{2}$) $\boldsymbol{Y}$ always increases welfare.

Proposition X' *There exist information structures* $S_1 := (\boldsymbol{X}', \boldsymbol{Y})$ *and* $S_2 := (\boldsymbol{X}, \boldsymbol{Y})$ *such that* $\boldsymbol{X}' \succeq_{sp} \boldsymbol{X}$*, and the payoff to the designer under* S_1 *is strictly lower than under* S_2*.*

Sketch of proof. We will only outline the structure of the proof here and omit some straightforward arguments to conserve space and notation. The arguments are nearly identical to the proof of Proposition X.

To this end, suppose that $\boldsymbol{Y}$ is a binary random variable as in the baseline case with precision p_y.[26] Therefore, for each realization of $x \in \text{supp}(\boldsymbol{X})$, there are two possible posteriors, $\{\underline{\mu}(x), \bar{\mu}(x)\}$. Moreover, for any x we have

$$\Delta(x) := F^b(\underline{\mu}(x)) - F^u(\bar{\mu}(x)) = 2 + \frac{x}{1-x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right]$$

As in Lemma 2, $\Delta(x)$ is decreasing in x and x^* is the belief such that $\Delta(x^*) = 0$. Suppose that $\alpha > x^*$, the threshold belief where $F^b(\underline{\mu}(x)) = F^u(\bar{\mu}(x))$. We will suppress the dependence on x henceforth and just denote these objects by F^b and F^u. Fix an $\epsilon > 0$ that is sufficiently small. Consider two signals as follows.

$$G_{\boldsymbol{X}}(x) = \begin{cases} 0 & \text{if } x < x^* \\ \frac{x-x^*}{2\epsilon} & \text{if } x \in [x^*, x^* + \epsilon] \\ \frac{1}{2} & \text{if } x^* + \epsilon < x < x_1 \\ \frac{1}{2} + \frac{x-x_1}{2\epsilon} & \text{if } x \in [x_1, x_1 + \epsilon] \\ 1 & \text{if } x > x_1 + \epsilon \end{cases}$$

where x_1 is such that $\frac{1}{2}[x^* + \frac{\epsilon}{2}] + \frac{1}{2}[x_1 + \frac{\epsilon}{2}] = \alpha$. And,

[25] Here, $\boldsymbol{\mu}'_{|x}$ means a random variable $\boldsymbol{\mu}' := \mathbb{P}(\boldsymbol{\theta} = 1 | \boldsymbol{X} = x, \boldsymbol{Y}')$ with a CDF $G_{\boldsymbol{\mu}'_{|x}}$.

[26] As the following discussion will demonstrate, the result is not driven by the binary nature of $\boldsymbol{Y}$.

$$G_{\boldsymbol{X}'}(x) = \begin{cases} 0 & \text{if } x < x^* - \epsilon \\ \frac{x - x^* + \epsilon}{2\epsilon} & \text{if } x \in [x^* - \epsilon, x^*] \\ \frac{1}{2} & \text{if } x^* < x < x_2 \\ \frac{1}{2} + \frac{x - x_2}{2\epsilon} & \text{if } x \in [x_2, x_2 + \epsilon] \\ 1 & \text{if } x > x_2 + \epsilon \end{cases}$$

where x_2 is such that $\frac{1}{2}[x^- \frac{\epsilon}{2}] + \frac{1}{2}[x_2 + \frac{\epsilon}{2}] = \alpha$.

Notice that $x_2 > x_1$, and hence, $\boldsymbol{X}' \succeq_{sp} \boldsymbol{X}$ if α such that $x_1 > \frac{1}{2}$. Suppose that α and p_y are such that the following restrictions hold.

1. $\bar{\mu}(x^* - \epsilon) > \frac{1}{2} > x^* + \epsilon$. Therefore, $\underline{\mu}(x^* + \epsilon) < \frac{1}{2}$. This implies that it is efficient to act on $\bar{\mu}(x)$ for all $x \in \text{supp}(\boldsymbol{X})$ as well as for all $x \in \text{supp}(\boldsymbol{X}')$.
2. $\underline{\mu}(x_1) > \frac{1}{2}$. Notice that $x_2 > x_1$. Therefore, this assumption implies that it is efficient to whenever $\boldsymbol{X} \geq x_1$ or $\boldsymbol{X}' \geq x_2$.

The reason why the welfare with $(\boldsymbol{X}', \boldsymbol{Y})$ is lower than with $(\boldsymbol{X}, \boldsymbol{Y})$ is nearly identical to the baseline model. In the case of $\boldsymbol{X}'$, with probability $\frac{1}{2}$, the posterior is above x^*. In this case, $F^b(\underline{\mu}(x)) > F^u(\bar{\mu}(x))$. The designer cannot induce the unbiased agent to act on $\bar{\mu}(x)$ while preventing the biased agent from acting on $\underline{\mu}(x)$. Thus, for a sufficiently high γ, the optimal equilibrium would entail $F(x) = 0$ and the equilibrium would be as in Table 2a. The biased agent acts with probability 1 for all the realizations. The unbiased agent acts whenever it is efficient to do so. On the other hand, in the case of $\boldsymbol{X}$, the posterior is always above x^* and the equilibrium would be as in Table 2c. Here, as in the baseline model, the biased agent acts with an interior probability, η^b, on $\underline{\mu}(x)$, while the unbiased agent acts with probability 1 whenever it is efficient to do so. Recall that the equilibrium in Table 2a yields a strictly lower welfare than the one in Table 2c. Therefore, replicating the argument in Proposition X, for a small ϵ, the reduction in welfare follows, i.e., welfare under $(\boldsymbol{X}, \boldsymbol{Y})$ is strictly larger than that under $(\boldsymbol{X}', \boldsymbol{Y})$. □

We now turn to the remaining point: To show is that a more spread out $\boldsymbol{Y}$ implies an unambiguous increase in welfare as in Proposition Y. To this end, fix a public signal, $\boldsymbol{X}$. Let

$$\mathcal{Y}^{\boldsymbol{X}} := \{\boldsymbol{Y} : \forall x \in \text{supp}(\boldsymbol{X}), \quad G_{\boldsymbol{\mu}|x}(\cdot) \text{ has no atoms.}\}.$$

Let $F(\cdot)$ be some punishment function. Let $\{\mu^u(F(x)), \mu^b(\mu^u(F(x)))\}$ be the posteriors that $F(x)$ makes indifferent for type u and b respectively. Henceforth, we will simply denote them by μ^b and μ^u. Notice that, without loss, $\mu^b \leq \frac{1}{2} \leq \mu^u$ with at least one inequality being strict. The biased type acts whenever $\mu \geq \mu^b$ while the unbiased type acts whenever $\mu \geq \mu^u$. Therefore, we can define the welfare given a signal $\boldsymbol{Y}$ (suppressing the dependence on $\boldsymbol{X}$, and its realization, x) to be

$$\Pi(\boldsymbol{Y}) := (1 - \gamma) \int_{\mu^b}^{1} (2\mu - 1) \mathrm{d}G_{\boldsymbol{\mu}|x}(\mu) + \gamma \int_{\mu^u}^{1} (2\mu - 1) \mathrm{d}G_{\boldsymbol{\mu}|x}(\mu).$$

Proposition Y' *Consider $\boldsymbol{X} \in \mathcal{X}$, and $\boldsymbol{Y}, \boldsymbol{Y}' \in \mathcal{Y}^{\boldsymbol{X}}$ such that $\boldsymbol{Y} \succeq_{sp} \boldsymbol{Y}'$. Then $\Pi(\boldsymbol{Y}') \geq \Pi(\boldsymbol{Y})$.*

Proof. We want to establish the following. For all $x \in \text{supp}(\boldsymbol{X})$,

$$\begin{aligned}\Pi(\boldsymbol{Y}) - \Pi(\boldsymbol{Y}') &= (1-\gamma)\int_{\mu^b}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \\ &\quad + \gamma\int_{\mu^u}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \\ &\geq 0.\end{aligned}$$

Let $I_1 := \int_{\mu^b}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu)$. Define $h(\mu,\mu_1) := (2\mu-1)\mathbb{1}_{\mu\geq\mu^1} + (2\mu^1-1)\mathbb{1}_{\mu<\mu^1}$. Notice that $h(\cdot,\mu^1)$ is an increasing and a convex function in its first argument. Moreover, since $\boldsymbol{Y}$ is more spread out around $\frac{1}{2}$ than $\boldsymbol{Y}'$, $(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\cdot)$ changes sign exactly once at $\frac{1}{2}$. Therefore, by Theorem 3.A.44 of Shaked and Shanthikumar (2007), $\boldsymbol{\mu}_{|x} \geq_{cx} \boldsymbol{\mu}'_{|x}$, where $\geq_{cx}$ denotes the convex order.[27] Therefore,

$$\begin{aligned}&\int_0^1 h(\mu,\mu^b)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0 \\ \implies &h(\mu^b,\mu^b)(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu^b) + \int_{\mu^b}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0\end{aligned}$$

Since $\mu^b < \frac{1}{2}$, $h(\mu^b,\mu^b) < 0$ and $(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu^b) \geq 0$. Therefore,

$$I_1 = \int_{\mu^b}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0.$$

Similarly,

$$\begin{aligned}&\int_0^1 h(\mu,\mu^u)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0 \\ \implies &h(\mu^u,\mu^u)(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu^u) + \int_{\mu^u}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0\end{aligned}$$

Since $\mu^u > \frac{1}{2}$, $h(\mu^u,\mu^u) > 0$ and $(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu^b) \leq 0$. Therefore,

$$I_2 = \int_{\mu^u}^{1}(2\mu-1)\mathrm{d}(G_{\boldsymbol{\mu}_{|x}} - G_{\boldsymbol{\mu}'_{|x}})(\mu) \geq 0.$$

Therefore, $\Pi(\boldsymbol{Y}) - \Pi(\boldsymbol{Y}') = (1-\gamma)I_1 + \gamma I_2 \geq 0$. □

Remark 3 *Notice that Proposition Y' establishes a ranking for an arbitrary punishment for a given realization x. Since $\Pi(\boldsymbol{Y}) \geq \Pi(\boldsymbol{Y}')$ for all $F(x)$, the welfare ranking analogous to Proposition Y follows by taking the supremum over $F(x)$ on both sides for each x.*

[27] Given two random variables, Z_1, Z_2, we say that $Z_1 \geq_{cx} Z_2$ if $\mathbb{E}[\phi(Z_1)] \geq \mathbb{E}[\phi(z_2)]$ for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$.

Intuition. Recall the intuition from the baseline (binary) model. Increasing the verifiable information may lead to a qualitative change in the type of equilibria the designer can support. As p_x increases, it may not be possible to effectively separate the critical type-information pairs: the information $(-1,-1)$ for the biased agent and the information $(-1,1)$ for the unbiased agent. As p_x increases, there ceases to be a punishment that deters the former yet encourages the latter. The same mechanism drives Proposition X' for a more general information structure. As $\boldsymbol{X}$ becomes more spread out, a higher mass on posteriors prevents the designer from separating the critical type-information pairs. The welfare reduction follows.

For the unverifiable information, the intuition follows, again, by considering the binary model. In that model, for any given realization of $\boldsymbol{X}$, there are 2 realizations $\boldsymbol{Y}$: one good, one bad. As we want to— in our baseline case at least—motivate the unbiased agent to act after a good $\boldsymbol{Y}$ realization, and deter the biased agent from acting after a bad $\boldsymbol{Y}$ realization, the posterior after the good (bad) signal lies above (below) 1/2. As p_y increases, the distance between the two critical posteriors increases. Therefore it becomes easier to separate the two. That exact logic carries over to the general setting. With more spread out signals, there is more mass on the "easier" cases which facilitates separation. The result follows.

Informativeness. At first glance, it may seem surprising that other (and perhaps more familiar) notions of "better information" such as e.g., Blackwell informativeness do not deliver our result. The reason is that the Blackwell informativeness order is too weak. Hence, under the Blackwell order, better unverifiable information may also reduce welfare. The spread order strengthens the Blackwell order adapting it to our specific decision problem.[28] Because the spread order is stronger than the Blackwell order, the possibility result in Proposition X' holds. However, Proposition Y' need not hold under improvements according to the Blackwell order. We show this below.

To this end, as before, we say that $\boldsymbol{Y}$ *is more informative than* $\boldsymbol{Y}'$ if $\boldsymbol{\mu}_{|x}$ is Blackwell more informative than $\boldsymbol{\mu}'_{|x}$ for all $x \in \text{supp}(\boldsymbol{X})$.

To see how a more informative $\boldsymbol{Y}$ can harm welfare, suppose that $\boldsymbol{X}$ is a binary random variable as in our baseline model, and, suppose that $|\text{supp}(\boldsymbol{Y})| = |\text{supp}(\boldsymbol{Y}')| = 3$. For the sake of concreteness, suppose that $\text{supp}(\boldsymbol{X}) = \{-1,1\}$ and $\text{supp}(\boldsymbol{Y}) = \text{supp}(\boldsymbol{Y}') = \{-1,0,1\}$. With some abuse of notation, we denote by $\mu(x,y)$ to mean $\boldsymbol{\mu}$ after observing $\boldsymbol{X} = x$ and $\boldsymbol{Y} = y$. Similarly, $\mu'(x,y)$ denotes $\boldsymbol{\mu}'$ after observing $\boldsymbol{X} = x$ and $\boldsymbol{Y}' = y$. Suppose that the following holds: Let the signals, $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Y}'$ be such that $\mu(1,\cdot), \mu'(1,\cdot) > \frac{1}{2}$. Moreover, suppose that the following holds:

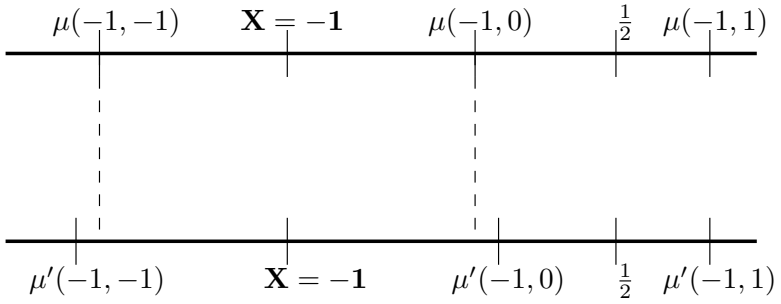
1. $\boldsymbol{\mu}(1,\cdot) > \frac{1}{2}, \boldsymbol{\mu}'(1,\cdot) > \frac{1}{2}$. That is, it is efficient to act whenever $\boldsymbol{X} = 1$ regardless of the unverifiable information.
2. (i) $\boldsymbol{\mu}(-1,-1) < \boldsymbol{\mu}(-1,0) < \frac{1}{2} < \boldsymbol{\mu}(-1,1)$ and,
   (ii) $\boldsymbol{\mu}'(-1,-1) < \boldsymbol{\mu}'(-1,0) < \frac{1}{2} < \boldsymbol{\mu}'(-1,1)$.
   Together, these imply that it is efficient to act on $(-1,1)$ for both $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Y}')$, and inefficient to act on $(-1,-1)$ and $(-1,0)$.

[28] Blackwell informativeness is defined by a mean-preserving spread instead, that is, if F and G have the same mean, F is a mean preserving spread of G if $\int_{-\infty}^{x} F(s)ds \geq \int_{-\infty}^{x} G(s)ds$ for any x strictly inside the joint support. It is straightforward to see, that the spread order satisfies that property. However, not every mean preserving spread is more spread out. Below we discuss such an example.

3. Finally, suppose that $\mu'(-1,0) = \mu(-1,0) + \epsilon_1$ and $\mu'(-1,-1) = \mu(-1,-1) - \epsilon_2$ and $\mu'(-1,1) = \mu(-1,1)$ for a small $\epsilon_1, \epsilon_2 > 0$.

Figure 5 illustrates the posteriors when $\boldsymbol{X} = -1$.

Figure 5: Posterior beliefs with $\boldsymbol{Y}$ and $\boldsymbol{Y}'$



By construction, $\boldsymbol{Y}'$ is Blackwell more informative than $\boldsymbol{Y}$ as $\boldsymbol{Y}'$ is a mean-preserving spread of $\boldsymbol{Y}$. Finally, to see how this can, potentially, reduce welfare, the designer wishes to induce the unbiased agent to act on $(-1,1)$ while preventing the biased agent from acting on $(-1,0)$. If, say, $\mu^b(\mu(-1,1)) = \mu(-1,0)$ (recall Equation 8), then this is feasible. However, when $\mu(-1,0)$ increases to $\mu'(-1,0)$ with $\boldsymbol{Y}'(-1,0)$, then the designer can no longer have the unbiased agent act on $(-1,1)$ with positive probability while preventing the biased agent from acting on $(-1,0)$ with probability 1. The welfare reduction follows due to identical reasoning as in the baseline model.

We conclude with a brief discussion about information structures in light of these results. Blackwell informativeness has indeed too many degrees of freedom. It cannot determine whether an improvement in either type of information, verifiable and unverifiable, increases or decreases welfare without further qualifications. When two experiments are ranked according to the Blackwell ranking, we obtain a ranking across *arbitrary* decision problems. Instead, our problem has more structure and allows a stronger definition of 'better information.' The central friction of our model is that a biased agent may act even when he believes that the project may fail. Likewise, the unbiased agent may refuse to act even when he believes that the project succeeds. More spread out information implies overall more convinced agents. Our finding shows that more spread out unverifiable information is beneficial—the fear (confidence) of the biased (unbiased) increases. The same does not hold for verifiable information—being more spread out. Here, it may lead to less trust in unverifiable but valuable signals causing welfare losses even if the punishment tries to account for it.

D Cases not discussed in the main text

As mentioned in the main text, we have the following four cases depending on where it is interim efficient to act.

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

In the main text, we analyzed case 3 where it is efficient to act iff either x or y is 1. We will refer to this case as the baseline case henceforth. Case 1 is straightforward, as mentioned in the main text, by setting $F(-1)$ to be very large and $F(1) = 0$. We now analyze the remaining 2 cases.

D.1 Efficient to act iff $y = 1$.

In this case, the goal is to deter the b type from acting on $(-1,-1)$ and $(1,-1)$ while incentivizing the u type from acting on $(-1,1)$ and $(1,1)$. An optimal policy trades off between different costs just like in the baseline case. The main difference in this case is that we have to choose $F(-1)$ and $F(1)$. Recall that in the baseline case $F(1)$ was 0 as acting on $x = 1$ was efficient.

The key idea in this case is that by setting $\bar{F}$ high enough, we can essentially treat the analysis of $x = 1$ separately from when $x = -1$. In fact, on each of these, the reasoning that guides us to $\bar{F}^*(-1)$ and $\bar{F}^*(1)$ is identical to those from the baseline case. We state the claims for this environment below. The proofs are identical and hence we have chosen to skip them.

Let $F^u(1,1)$ denote the largest punishment up to which the unbiased type will act on $(1,1)$ and $F^b(1,-1)$ denote the smallest punishment necessary to deter the biased type from acting on $(1,-1)$.

Claim 5 *If* $F^u(-1,1) > F^b(-1,-1)$ *then* $\bar{F}^*(-1) \in \{F^b(-1,-1), F^u(-1,1)\}$. *If* $F^u(-1,1) < F^b(-1,-1)$ *then* $\bar{F}^*(-1) \in \{0, F^b(-1,-1)\}$.

Claim 6 *If* $F^u(1,1) > F^b(1,-1)$ *then* $\bar{F}^*(1) \in \{F^b(1,-1), F^u(1,1)\}$. *If* $F^u(1,1) < F^b(1,-1)$ *then* $\bar{F}^*(1) \in \{0, F^b(1,-1)\}$.

As in the baseline case, for the main comparative static, the difference between $F^b(-1,-1)$ and $F^u(-1,1)$, as well as the difference between $F^b(1,-1)$ and $F^u(1,1)$ matters in determining the optimal fine. Define,

$$\Delta^1(p_x,p_y) := F^b(1,-1) - F^u(1,1) = 2 + \frac{\alpha}{1-\alpha}\frac{p_x}{1-p_x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right]$$

$$\Delta^{-1}(p_x,p_y) := F^b(-1,-1) - F^u(-1,1) = 2 + \frac{\alpha}{1-\alpha}\frac{1-p_x}{p_x}\left[\frac{1-p_y}{p_y} - \frac{p_y}{1-p_y}\right]$$

Below we state a straightforward result, exactly as in Lemma 2 from the main text for the baseline case.

Lemma 3 *$\Delta^1(p_x,p_y)$ is decreasing in p_x, p_y. $\Delta^{-1}(p_x,p_y)$ is increasing in p_x and decreasing in p_y. Moreover, $\Delta^{-1}(p_x,p_y) > \Delta^1(p_x,p_y)$ for all $p_x, p_y \in (\frac{1}{2}, 1)$.*

Recall that we obtained Proposition X thanks to the following observation: Start with a p_x such that $\Delta^{-1}(p_x,p_y) < 0$ but is close to 0. Then, it is possible to have $a^b(-1,-1) = \eta^b < 1$ *and* $a^u(-1,1) = 1$. However, a slight increase from p_x to $p'_x > p_x$, we can have $\Delta^{-1}(p'_x,p_y) > 0$. In this case, we can no longer have an equilibrium where $a^u(-1,1) = 1$ *and* $a^b(-1,-1) < 1$. In particular, if γ is sufficiently high we set $\bar{F}^*(-1) = 0$ and obtain $a^u(-1,1) = a^b(-1,-1) = 1$. That is, the unbiased type acts on $(-1,1)$ but the price we pay is that the biased type acts with probability 1 on $(-1,-1)$.

Notice that this reasoning is identical, when $x = -1$, in the case where y is pivotal. Also, if $\Delta^1(p_x,p_y) < 0$, then we can have $a^b(1,-1) = \eta^b$ *and* $a^u(1,1) = 1$.

Moreover, the crucial point to note is that $\Delta^1(p_x, p_y)$ is decreasing in p_x, and is smaller than $\Delta^{-1}(p_x, p_y)$. Therefore, if $\Delta^{-1}(p_x, p_y) < 0$ then $\Delta^1(p_x, p_y) < 0$. And, for any $p'_x > p_x$, $\Delta^1(p'_x, p_y) < 0$. As a consequence, if we have a critical belief p^*_x, i.e. $F^b(-1,-1) = F^u(-1,1)$, then $\Delta^1(p^*_x, p_y) < 0$, and will continue to be so in a neighbourhood of p^*_x.

Therefore, replicating the construction as in the baseline case, we can obtain a similar result as in Proposition X and Y in this environment as well. That is, there exist a set of parameters where increasing p_x can reduce welfare but increasing p_y can never harm welfare. We state them formally below.

Proposition $\widehat{XY}$ *There exists (non knife-edge) environments,* $(p_x, p'_x, p_y, \gamma, \alpha)$ *, such that* $p_x > p'_x$ *and* $W^*(p'_x, p_y, \gamma, \alpha) > W^*(p_x, p_y, \gamma, \alpha)$*. Moreover, for all environments* $(p_x, p_y, \gamma, \alpha)$ *society's welfare* $W^*(p_x, p_y, \gamma, \alpha)$ *is non-decreasing in* p_y*.*

D.2 Efficient to act iff $x = y = 1$

First of all, in this case, we can set F to be larger than $F^b(-1,1)$ and convict on $x = -1$. This way, we ensure that no type acts on $x = -1$. Therefore, what remains is the case when $x = 1$. Here, we want to have $a^u(1,1) = 1$ and $a^b(1,-1) = 0$. Unsurprisingly, the possibility of this depends on how $F^u(1,1)$ and $F^b(1,-1)$ are ranked. Lastly, since $\Delta^1(p_x, p_y)$ is decreasing in p_x, increasing p_x cannot reduce welfare in this case.

D.3 General Proof of Proposition 2

First step: Inside each case. The main observation is that the cases $x = 1$ and $x = -1$ can be addressed separately since the principal can condition on the realization of $\boldsymbol{X}$ and by Lemma 3 above $\Delta^1(p_x, p_y)$ and $\Delta^{-1}(p_x, p_y)$ decrease in p_y.

We restate below the four cases mentioned earlier in this appendix and the main text.

Case 1 **It is efficient to act iff** $z = 1$. In this case, Argument 1 of Appendix A.4 applies for both $z = 1$ and $z = -1$

Case 2 **It is efficient to act iff** $y = 1$**.** For $z = -1$ this case is identical to the baseline case. Argument 2 of Appendix A.4 applies. Also conditional on $z = 1$ the situation is as in the baseline case for $z = -1$. Since $\Delta^1(p_x, p_y)$ decreases in p_y, Argument 2 of Appendix A.4 applies directly.

Case 3 **It is efficient to act iff** $z + y \geq 0$**.** This is the baseline case. We showed it in Appendix A.4.

Case 4 **It is efficient to act iff** $z + y = 2$**.** The case for $z = 1$ is as in the previous case and Argument 2 of Appendix A.4 applies. For $z = -1$ Argument 1 of Appendix A.4 applies

Second Step: Accross cases. As we keep (p_x, γ, α) fixed and increase p_y, we can move across cases. In particular, the following relation holds.

- Case 2 is absorbing—any increase in p_y keeps us in this case.
- Case 3 can only transition to case 2.
- Case 4 can go to case 2 directly or through case 3.
- Case 1 can go through case 3 or case 4.
- In knife-edge cases, a direct transition from case 1 to case 2 is possible.

We show that $W^*(p_x, p_y, \gamma, \alpha)$ is continuous at the boundaries and thus, by the first step, welfare improves.

From 1 to 3. Take $\hat{p}_y$ such that $\mathbb{P}(\theta = 1|\boldsymbol{X} = -1, \boldsymbol{Y} = 1) = 1/2$. Then, for any $p_y < \hat{p}_y$, we are in case 1 and for any $p_y > \hat{p}_y$ we are in case 3. For $\hat{p}_y$, full deterrence of the biased type without any chilling effect is possible by the punishment F^b conditional on $x = -1$ since the unbiased type (and the society) are indifferent between taking an action or not taking an action. Yet, F^b is also feasible. Thus, the transition is continuous.

From 1 to 4. Take $\hat{p}_y$ such that $\mathbb{P}(\theta = 1|\boldsymbol{X} = 1, \boldsymbol{Y} = -1) = 1/2$. Then, the society is indifferent between everyone acting on $(-1, 1)$, no one acting on it, or only the biased type acting on it. It is feasible by setting $F(-1) = 1$ for example. It covers all the potential action profiles in case 4. Thus, welfare is continuous at the boundary.

From 3 to 2. Analogous to the case from 1 to 4.

From 4 to 2. Analogous to the case from 1 to 3.

E Extensions not Discussed in the Main Text

E.1 Asymmetric Precision

In the baseline model, we assumed that the precision of a signal, $\boldsymbol{X}$ or $\boldsymbol{Y}$, is independent of the state. That is, p_x is the probability that $\boldsymbol{X}$ matches the state regardless of the actual realization of the state. We now relax this. Let $p_x^i := \mathbb{P}(\boldsymbol{X} = \theta|\theta = i)$ and p_y^i analogously. Straightforward calculations show that

$$F^b - F^u = -2 + \frac{\alpha}{1-\alpha}\frac{1-p_x^1}{p_x^{-1}}\left[\frac{1-p_y^1}{p_y^{-1}} - \frac{p_y^1}{1-p_y^{-1}}\right].$$

Therefore, $F^b - F^u$ is increasing in p_x^1, p_x^{-1} and decreasing in p_y^1, p_y^{-1}.[29] The comparative statics follow from this monotonicity.

E.2 Conditionally Dependent Signals

In the baseline model, we assumed that signals $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent. Relaxing this while retaining a structure that allows us to perform comparative statics wherein we improve the precision of one signal while keeping the precision of the other signal the same is not straightforward. Here we provide one specific example. Suppose that $\boldsymbol{X}$ (the public signal) is a binary signal which has a precision of p_x. Moreover, $\boldsymbol{Y}$ is another binary signal that is equal to $\boldsymbol{X}$ with probability ρ. More precisely, conditional on the state θ, $\boldsymbol{Y}$ has the following distribution:

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{X} & \text{w.p.} \rho \\ \theta & \text{w.p. } p_y(1-\rho) \\ -\theta & \text{w.p. } (1-p_y)(1-\rho) \end{cases}$$

[29] To be more precise, the said monotonicity holds when, as in our main model, we assume that $p_x^1, p_x^{-1}, p_y^1, p_y^{-1} \geq \frac{1}{2}$.

In this case, we have,

$$F^b - F^u = -2 + \frac{\alpha}{1-\alpha}\frac{1-p_x^1}{p_x^{-1}}\left[\underbrace{\frac{1-(1-\rho)p_y}{\rho+p_y(1-\rho)} - \frac{p_y}{1-p_y}}_{A}\right]$$

It is easy to check that $A \leq 0$ for any $\rho \in [0,1]$. Therefore, as in Lemma 2, $F^b - F^u$ is increasing in p_x and decreasing in p_y. The comparative statics follow from this monotonicity.

References

J. Baron, I. Ritov, et al. Reference points and omission bias. *Organizational behavior and human decision processes*, 59:475–475, 1994.

E. Bernstein. The transparency trap. *Harvard Business Review*, 92(10):58–66, 2014.

J. F. Bibby. Committee characteristics and legislative oversight of administration. *Midwest Journal of Political Science*, 10(1):78–98, 1966.

J. Blanes i Vidal and M. Möller. When should leaders share information with their subordinates? *Journal of Economics & Management Strategy*, 16(2):251–283, 2007.

J. Bull and J. Watson. Statistical evidence and the problem of robust litigation. *The RAND Journal of Economics*, 50(4):974–1003, 2019.

P. Cane. Mens rea in tort law. *Oxford Journal of Legal Studies*, 20(4):533–556, 2000.

A. Chalfin and J. McCrary. Criminal deterrence: A review of the literature. *Journal of Economic Literature*, 55(1):5–48, 2017.

J. C. Cox, M. Servátka, and R. Vadovič. Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission. *Experimental Economics*, 20(1):1–18, 2017.

Fowler v. Padget. 7 term rep 509. Technical report, 7 Term Rep 509; 101 ER 1103, 1798.

N. Garoupa. The economics of political dishonesty and defamation. *International Review of Law and Economics*, 19(2):167–180, 1999.

K. N. Hylton. Economic theory of criminal law. 2019.

L. Kaplow. On the optimal burden of proof. *Journal of Political Economy*, 119(6): 1104–1140, 2011.

L. Kaplow. Optimal multistage adjudication. *The Journal of Law, Economics, and Organization*, 33(4):613–652, 2017a.

L. Kaplow. Optimal design of private litigation. *Journal of Public Economics*, 155: 64–73, 2017b.

R. Lagunoff. A Theory of Constitutional Standards and Civil Liberty. *The Review of Economic Studies*, 68(1):109–132, 01 2001. ISSN 0034-6527. doi: 10.1111/1467-937X.00162. URL https://doi.org/10.1111/1467-937X.00162.

B. Lester, N. Persico, and L. Visschers. Information acquisition and the exclusion of evidence in trials. *The Journal of Law, Economics, & Organization*, 28(1):163–182, 2012.

S. Morris and H. S. Shin. Social value of public information. *American Economic Review*, 92(5):1521–1534, 2002.

H. Pei and B. Strulovici. Crime entanglement, deterrence, and witness credibility. *mimeo*, 2019.

A. Prat. The wrong kind of transparency. *American economic review*, 95(3):862–877, 2005.

C. Prendergast. A theory of "yes men". *The American Economic Review*, pages 757–770, 1993.

C. W. Sanchirico. Character evidence and the object of trial. *Columbia Law Review*, 101(6):1227–1311, 2001. ISSN 00101958. URL http://www.jstor.org/stable/1123746.

J. Schrag and S. Scotchmer. Crime and prejudice: The use of character evidence in criminal trials. *Journal of Law, Economics, & Organization*, pages 319–342, 1994.

M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Science & Business Media, 2007.

G. J. Stigler. The optimum enforcement of laws. *Journal of Political Economy*, 78(3): 526–536, 1970.

E. Wang. Frightened mandarins: the adverse effects of fighting corruption on local bureaucracy. *Available at SSRN 3314508*, 2019.

F. Woollard. *Doing and allowing harm*. Number 1. Oxford University Press„ 2019.