THE WRONG KIND OF INFORMATION

Aditya Kuvalekar João Ramos Johannes Schneider*

October, 2020

Abstract

An agent decides whether to approve a project based on his information, some of which is verified by the principal. An honest agent wants to implement projects that are likely to succeed; a dishonest agent wants to implement any project. If the project fails, the principal examines the verifiable information and decides the punishment. The principal 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, while 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 punishments.[1] This phenomenon—the unintended consequences of a law or a regulation affecting activities outside its intended scope—is the *chilling effect.*

*Kuvalekar: U of Essex, Ramos: Queen Mary U, Schneider: U Carlos III de Madrid. We would like to thank Nageeb Ali, Heski Bar-Isaac, Dan Bernhardt, Dhruva Bhaskar, Antonio Cabrales, Odilon Câmara, Marco Celentani, Joyee Deb, Siddharth Hari, Chad Kendall, Nenad Kos, Antoine Loeper, Anthony Marino, John Matsusaka, Moritz Meyer-Ter-Vehn, Ignacio Ortuño Ortín, Harry Pei, Jacopo Perego and Maher Said for helpful comments and discussions. Aditya Kuvalekar and Johannes Schneider gratefully acknowledge support from the María de Maeztu grant MDM 2014-0431, the Ministerio Economia y competitividad grant PGC2018-09159-B-I00 and the Comunidad de Madrid, MadEco-CM(S2015/HUM-3444) grant.

[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.

Our paper focuses on 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 ex-post (e.g. in courts), others are not. For example, doctors, civil servants and politicians often possess extremely valuable situational knowledge based on their experience over and above the verifiable information such as exams and reports. The chilling effect is that individuals hesitate to rely on the unverifiable information if it contradicts the information that is verifiable.

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 welfare consequences. More precisely, welfare unambiguously increases in the precision of the *unverifiable information*, but, in contrast, welfare may decline in the precision of the *verifiable information*.

In the real world, there are constant improvements in the available information, of both verifiable and unverifiable nature—e.g., 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 welfare consequences.

Model: Motivated by the forces described above, we provide a simple framework capturing the following three features: First, some agents may have preferences misaligned with the society's. Second, agents act on information not all of which is verifiable ex-post. And third, the society wishes to deter the dishonest agents from taking bad decisions out of selfish interest while encouraging the honest agents to rely on the (often useful) unverifiable information.

More concretely, there are three players: a designer, a principal and an agent. The agent (honest or dishonest) decides whether to implement a risky project or to take the safe action. The agent relies on two conditionally independent binary signals about the binary state of the world to inform his decision.[2] The risky project succeeds when the state is good and fails when the state is bad. The honest agent's preferences are aligned with society's. He values successful projects and suffers from failures. The dishonest agent always wants to implement a project. If the agent implements the risky project, the state of the world is publicly realized. Thereafter, the principal (e.g. a court) decides whether to punish and what the punishment will be. The main innovation of the model is to consider that the two signals differ—one is verifiable by the principal, while the other is not.

[2] The binary signal structure is purely for convenience. Our main results—demonstrating how the effect of improving the quality of information depends on the nature of information—, as well as the underlying mechanism behind it, extend to richer environments.

The objective of the principal (who is essentially an adjudicator) is to punish the agent if, given the available evidence, she is sufficiently convinced that the agent is dishonest.[3] The designer (i.e. the society) selects the maximum punishment the principal can inflict upon the agent. The designer chooses this punishment to maximize welfare, defined as her ex-ante payoff.

If the agent implements the project and it fails, the principal observes the verifiable information and decides on the agent's punishment. To highlight the main mechanism, we focus on the case in which any positive signal (verifiable or unverifiable) is sufficient to make implementation interim efficient.[4]

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 honest 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 the dishonest agent—absent sufficient punishment, he implements the project even when it is inefficient to do so.

For instance, consider an honest 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 principal only observes the verifiable information. Naturally, if the threat of consequences is not too harsh, the agent may be willing to act.

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 lowers the agent's posterior belief about the project's likelihood of success. Therefore, the agent now 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.

This particular case, however, suggests a simple remedy—lower the punishment if the verifiable information becomes more precise. Therefore, a key contribution lies in exploring whether we can adjust the punishment scheme so that the benefits of improved information outweigh the potential cost of a stronger chilling effect.

Our main results, propositions 1 and 2, show that the stronger chilling effect may outweigh the benefits of improved information, even adjusting the punishment optimally. In particular, increasing the precision of the verifiable information can lead to welfare reductions in non-knife-edge cases, while, in contrast, increasing

[3] We also explore the consequences of the principal having a different objective than screening the dishonest agent in Section 4.3.

[4] The analyzes of all remaining cases are presented in the online appendix.

the precision of the unverifiable information always increases welfare.

We take this opportunity to 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 chose to present a binary, stylized model, the novel mechanism that drives this difference (discussed below) goes beyond it.

The mechanism: The main conflict in our environment is that, at times, we want the honest agent to decide in favor of implementing the project against the verifiable information, relying only on his unverifiable information. However, the associated cost is that the dishonest agent may implement a project when both the verifiable and unverifiable information suggest not doing so. Increasing the precision of unverifiable information helps the designer here. At once, it makes the honest agent more confident about implementing the project by trusting the unverifiable information whilst it discentivizes the dishonest agent from implementing it. Thus, welfare increases.

In contrast, increasing the precision of the verifiable signal 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, and only for the honest 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 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 lower welfare.

1.1 Examples

Real-world examples of our setting abound, as the central tension that motivates our exercise is relevant in various environments, from politics to inside a firm. Essentially, our model aims to capture scenarios with the following three ingredients: decisions are made in the face of uncertainty; poor outcomes involve consequences; and there is adverse selection on the part of the agent. We chose to focus the paper on the optimal punishment of a principal lacking commitment. However, we want to emphasize that our main comparative statics are not a direct result of the lack of commitment or of the optimal design of the consequences, but, rather, despite it. As discussed in Section 4.1, our environment can also accommodate situations in which formal contracts can be used to specify the punishments following negative outcomes, and, as discussed in Section 4.2, settings in which the punishment is independent of the quality of information (and therefore not optimally adapted to it).

First, consider a bureaucrat deciding whether to approve the expenditure on a certain project, which may be overpriced. Approving expenses involves taking a risk, as overpriced projects invite corruption charges. The bureaucrat relies on verifiable (e.g. reports) and unverifiable information (e.g. expertise) to inform her of whether the project is overpriced, and to decide whether to approve it or not. The punishment in case of overpricing depends on the reports provided, but the bureaucrat's expertise cannot be used in court. Lastly, different bureaucrats care differently about citizens' interests.

Second, consider a president deciding on a foreign policy issue. For instance, the decision of whether to impose sanctions to a country in response to an invasion of a neutral country. For the president this is a risky decision, as her electoral chances in the future might be compromised following negative outcomes. The safe option is to follow standard diplomatic procedures. Different politicians may value their future chances of election differently. The decision is taken based on top-secret intelligence reports (private information) as well as what is being reported in the news (public information). Voters might hold the president accountable, but only have access to the outcome and the public information.

As a third example, consider a doctor deciding on the delivery method of a child without obvious complications. To decide, the doctor relies on some verifiable information (for instance, ultrasound examinations indicating the child's position in the womb) and on some unverifiable information (for instance, 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 evidences supporting its choice, then the doctor may face legal and administrative consequences.

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 in the future, to decide whether to proceed with the acquisition. Finally, different CEO's might weigh long-run outcomes and short-run outcomes differently.

Although these examples are very different in nature—for instance, in the first and last examples the punishment magnitude may be adapted to the quality of information available and formal contracts may govern relationships in the last example, while in the second and third examples the punishment is fixed—our framework is insightful to study the consequences of better information in equilibrium in each of these cases.

1.2 Related Literature

At a high level, our paper contributes to the literature on deterrence and punishment following Becker (1968).[5] In our paper, the level of expected punishment influences how much an agent uses the information available to him when taking a risky decision. A consequence of this interaction is that changing the level of punishment can have unintended side effects: honest agents may be discouraged to take beneficial yet risky actions.

A small literature has identified other, orthogonal, side effects of punishment and is thus related in spirit. For example, 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 large punishment reduces the amount of crimes reported by witnesses, thereby reducing the cost of committing a crime. Intermediate punishments can deter some individuals from committing crimes, yet those that commit some crime are likely to commit several crimes.

More substantively, the chilling effect itself—deterrence of one action may unintentionally prevent other, desired actions—has been recognized in the literature before. 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.[6] We complement these models by introducing a novel channel (via the interaction of the verifiable and the unverifiable information) behind the chilling effect, and also by demonstrating the pernicious effects of the verifiable information through the chilling effect.

There is by now a large literature showing that, in different environments of strategic interactions, improving the quality of information available can harm welfare.[7] The substantive message of this paper, however, goes beyond highlighting this in a novel setting. More importantly, we show that the welfare effects of improving the precision of information depend on the nature of information—

[5] See also Lazear (2006), Mookherjee and Png (1994), Polinsky and Shavell (1984) for the economics literature and e.g. Hylton (2019) and Chalfin and McCrary (2017) and references therein for the connection to the legal literature.

[6] A couple of other examples include Prato and Strulovici (2017) (PS) and Brandenburger and Polak (1996) (BP). PS consider the effects of direct democracy on the incentives for politicians to implement projects. Their *negative incentive spiral* contains elements of the chilling effect. BP show how managers of publicly traded firms can ignore useful private information to maximize the current share price by affecting the markets' expectations.

[7] For instance,Che et al. (2013) who show that, in a cheap talk game where a biased sender recommends one of two projects to a decision maker, a publicly available signal about the relative ranking of the two projects may harm the decision maker. See Crémer (1995), Dewatripont et al. (1999), Holmström (1999), Kim (1995) for examples on moral hazard and career concerns environments.

verifiable or not.

A similar result appears in Vidal and Möller (2007) who study a principal-agent problem where the principal has two pieces of information, only one of which can be shared with the agent before the agent chooses his effort. They show that sharing information may harm welfare. The mechanism there is that the principal finds it harder to elicit high effort from the agent if his own action relies more on his non-shareable information. A crucial difference in our setting is that the player possessing the information is also the one taking the action. Our principal only chooses the punishment (endogenously) to discipline the agent.[8]

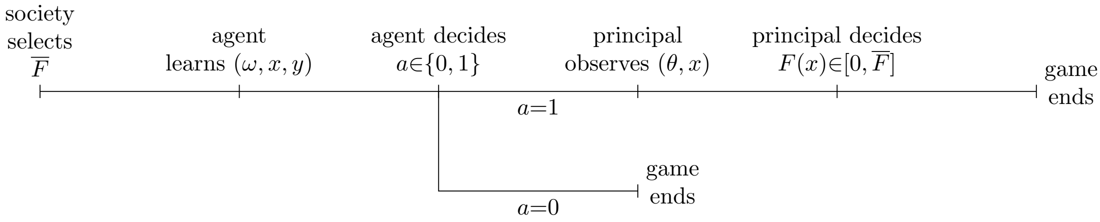
The closest papers in the contracting literature are Prendergast (1993) and Prat (2005). In a principal agent setting, Prendergast (1993) focuses on how to provide incentives for an agent to costly acquire relevant information. To provide incentives, the principal must evaluate the information obtained by the agent, thus the agent has incentives to inefficiently conform to what he expects the principal's prior to be. Instead of acquiring useful information about the state, the agent acquires (and relies on) useless information about the principal's belief of it. Also in a principal agent setting, 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). We highlight how the *nature of the information* about the same object—the quality of the project—affects welfare. While the results look similar at a superficial level, the main mechanism is different. In our world, better verifiable information affects honest and dishonest agents in different ways. An honest agent's direct payoff from taking an action changes with information, a dishonest agent's payoff does not. That difference leads to asymmetries in the effect of a change in precision.

Finally, our main comparative static is reminiscent of Morris and Shin (2002) if one views verifiable (unverifiable) information as public (private) information.[9] In contrast to our environment, coordination motives are the main driver of their results. 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, the 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.

[8] Frankel (2020) studies how to balance the emphasis on verifiable and unverifiable information when contracting with recruiters.

[9] In fact, Angeletos and Pavan (2007) offer a very general and unified treatment of the type of environments studied by Morris and Shin (2002). The key focus there is on the coordination motives.

Figure 1: Timing of the Game.


society selects $\overline{F}$
agent learns (ω, x, y)
agent decides $a \in \{0,1\}$
principal observes (θ, x)
principal decides $F(x) \in [0, \overline{F}]$
game ends
a=1
game ends
a=0

The society selects the maximum punishment, $\overline{F}$. The agent observes his type, ω, and 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 principal observes the project quality, θ, and the verifiable signal, x. If the project failed, θ= -1, the principal selects a punishment, $F(x) \in [0, \overline{F}]$. Then payoffs are realized.

2 Model

We consider a game with three players, an agent ("he"), a principal ("she") and the society (designer). The agent 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 principal, having observed a failure, decides whether and how to punish the agent. The maximum punishment available to the principal is determined by the designer ex-ante. We sketch the timing of the game in Figure 1.

Project Quality and Information. A project may be of good quality ($\theta = 1$) or bad quality ($\theta = -1$). If undertaken, a project of bad quality always fails and a project of good quality always succeeds. However, the project quality is unknown. The ex-ante probability that the project is good, $\theta = 1$, is denoted by α.

The agent does not know the project quality but has access to the realizations of two binary signals, $(X, Y) \in \{-1, 1\}^2$.The two signals are independent conditional on the state. The realization of the signal X, denoted by x, coincides with the true state θ with probability $p_x := \mathbb{P}(x = \theta)$ and is misleading with complementary probability, $1 - p_x = \mathbb{P}(x = -\theta)$. The signal Y has a similar structure with p_y being the likelihood of realization y being correct. We assume that both signals are informative about the state but are noisy; that is, $1 > p_x, p_y > 1/2$. Accordingly, we refer to p_i as the *precision* of signal i.

The Agent. The agent can be of two types, honest ($\omega = h$) or dishonest ($\omega = d$). The probability that the agent is honest is denoted by γ.

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

$$u^h(a; \theta) = a\theta, \qquad \text{and } u^d(a, \theta) = a.$$

Honest agents benefit from successful projects, but suffer from failed projects, while dishonest agents benefit from any project undertaken. Besides that, an agent's payoff is affected by the principal's decision on punishing the agent. If the agent is punished by the principal to 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).

The Principal. After the agent makes his decision, and the outcome of the project is revealed, the principal 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 highest possible punishment, $\overline{F}$, is chosen ex-ante by the designer to maximize societal welfare.

The principal observes the agent's action. If the agent implements the project, she observes the project's outcome and therefore the project's quality θ. In addition, she observes x. Because x is observed by the principal, we call x the *verifiable* information and y the *unverifiable* information. If no project is undertaken, the quality of the project remains unobserved.

Based on her information, $\{\theta, x\}$, the principal decides on the punishment to inflict the agent. We make two assumptions regarding the principal's objective and capacity to punish the agent.

First, we assume that the principal 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 really a 'safe' option for the agent. In reality, the reactions to a harmful action are indeed often more drastic than that 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 4.4, we extend the model to allow the principal to also punish an agent for inaction. We show that our results are unchanged.

Second, the principal aims to punish only dishonest agents. The principal infers the agent's preferences from the information available to her, and punishes only when she is sufficiently convinced that the agent is dishonest. Thus, the principal aims to punish wrongful intentions. Alternatively, we could consider a principal that aims to punish agents for taking the "wrong" action, irrespectively of the agents' intention. In Section 4.3, we discuss how our model outcomes remain largely unchanged under this different objective for the principal.[10]

Formally, the principal receives a benefit equal to the magnitude of the punishment, F, if she punishes a dishonest agent, and a loss, FL, with parameter $L > 0$,

[10] Since our goal is to identify the welfare-maximizing equilibrium (see below), in Section 4.3 we show that the welfare maximizing equilibrium leads to a higher welfare when the principal aims at screening out then dishonest agents than when the principal's objective is to punish the agents that took the "wrong" action.

if she inflicts punishment F on an honest agent. The principal's (expected) utility from punishing an agent that is honest with probability $\tilde{\gamma}$ is,

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

We interpret this tradeoff as a requirement of intent: the principal has to be sufficiently convinced that the agent's action was ill-intentioned; otherwise, she prefers not to punish the agent.

We denote the principal's strategy by $F : \{-1,1\} \to [0,\overline{F}]$, meaning that the principal inflicts punishment $F(x)$ when she sees $(a=1, \theta=-1, x)$.

Welfare. Our main objective is to understand the direct effect of the different natures of the information on 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.

We define the ex-ante welfare as follows:

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

That is, we assume that implemented good (bad) projects increase (decrease) welfare with respect to the status quo. The ex-post returns to projects coincide with those of an honest agent.[11]

Timing and Solution Concept. We focus on ex-ante welfare-maximizing perfect Bayesian equilibria. The timing is sketched in Figure 1: First, society chooses the maximum allowed punishment, $\overline{F}$. Then, the state of the world, (ω, θ, x, y), realizes. The agent observes (ω, x, y) and decides upon action a. If the agent implements a project, the principal observes (θ, x) and decides whether and how much punishment to inflict on the agent, $F(x) \leq \overline{F}$.[12]

3 Analysis

The analysis is done by backward induction. We first characterize the principal's best response—given that the agent chose to implement the project, the project's realization, and the verifiable information. Next, we focus on the agent's behavior. We characterize the agent's best response—given the punishment scheme, his type,

[11] In reality, punishment may also affect welfare. For example, a court sentence can imply an additional benefit (fine payments to the treasury) or an additional cost (providing a prison slot). Adding such effects to the model is, however, straightforward and does not alter the intuition.

[12] Our results remain unchanged if we assume that the principal commits ex-ante to a punishment scheme rather than acting at an interim level. We discuss the alternative model in detail in Section 4.

and the information available to him. In Section 3.1, we characterize the optimal punishment scheme, balancing the tradeoff of giving a free pass to the dishonest agent (lax punishment that results in the dishonest agent always acting) with the chilling effect that any threat of punishment may generate on the honest agent. Finally, Section 3.2 contains our main results.

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

Principal's Best Response. The principal's information set is (a, θ, x). As mentioned before, she can punish the agent only if $a = 1$, and $\theta = -1$. Let $\tilde{\gamma}(x) := \mathbb{P}(\omega{=}h|x, a{=}1, \theta{=}-1)$ be the principal's posterior belief of the agent being honest when the information set is $(a{=}1, \theta{=}-1, x)$. She wants to maximize $v(F) = F\left(1 - \tilde{\gamma}(x) - \tilde{\gamma}(x)L\right)$. Therefore, her best response is straightforward: impose a punishment of $F = \overline{F}$ if $\tilde{\gamma}(x) < \gamma^* := \frac{1}{1+L}$, and no punishment, henceforth called acquit, if $\tilde{\gamma}(x) > \gamma^*$. Also, she is indifferent between any punishment if $\tilde{\gamma}(x) = \gamma^*$.

Agent's Best Response. Assume that the principal inflicts on the agent a punishment of $F(x)$ for a given realization (x, y) if the project fails. Further, let $\beta_{xy} := \mathbb{P}(\theta = 1|x, y)$. Let $a^\omega(x, y)$ be the probability that an agent of type ω acts, given signals (x, y). Then, $a^\omega(x, y) = 1$ only if

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

Therefore, the agent follows a cutoff strategy: he acts if $\beta_{xy} > \overline{\beta}^\omega(F(x))$ and does not act if $\beta_{xy} < \overline{\beta}^\omega(F(x))$, where

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

Note that, if the honest agent acts for some (x, y), the dishonest agent acts, too, since $\overline{\beta}^h(F(x)) > \overline{\beta}^d(F(x))$.

3.1 Optimal Punishment

Basic Tradeoff. Our goal is to identify the optimal punishment scheme—the maximum punishment, $\overline{F}$, that the principal can impose. If $\overline{F}$ is low, the dishonest agent is given a *free pass*—he acts even if it is inefficient to do so. If $\overline{F}$ is high, the honest agent suffers from the *chilling effect*—the fear of being punished results in sometimes not acting even when it is efficient to act. The optimal punishment scheme balances deterrence of the dishonest agent with the encouragement of the honest agent.

The society maximizes $W(a^h, a^d, F; \overline{F})$ by choosing among equilibria given $\overline{F}$. Define,

$$\begin{aligned} \overline{W}(\overline{F}) &:= \sup_{(a^h,a^d,F)\in\mathcal{E}(\overline{F})} W(a^h, a^d, F; \overline{F}) \\ W^* &:= \sup_{\overline{F}\in[0,\infty)} \overline{W}(\overline{F}), \end{aligned}$$

where $\mathcal{E}(\overline{F}) := \{(a^h, a^d, F) : (a^h, a^d, F) \text{ constitute an equilibrium given } \overline{F}\}$.

Definition 1 *If there exists $(\overline{F}^*, a^h, a^d, F)$ such that $(a^h, a^d, F) \in \mathcal{E}(\overline{F}^*)$ and $W(a^h, a^d, F; \overline{F}^*) = W^*$, then we say that the equilibrium is a "(society) optimal equilibrium," and $\overline{F}^*$ is an "optimal punishment scheme."*

Cases. Since ex-post welfare is $a\theta$, it is interim efficient to act whenever $\mathbb{E}[a\theta|x, y] \geq 0$—i.e., whenever $\beta_{xy} \geq \frac{1}{2}$. The posterior belief, β_{xy}, depends on the parameters (p_x, p_y, α). We ignore the trivial cases in which signals are irrelevant, either because $\beta_{xy} \leq 1/2\ \forall(x, y)$ or because $\beta_{xy} \geq 1/2\ \forall(x, y)$. 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 $\Leftrightarrow x + y = 2$.

For the remainder of the paper, purely for expositional purposes, we focus on case 3: it is efficient to act iff at least one of x and y is 1.

Case 1 implies that x is more informative than 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. Figure 2 on page 15 shows these cases for a fixed α.

Case 1 presents no real tradeoffs. For a high $\overline{F}$, the principal sets $F(-1) = \overline{F}$ and $F(1) = 0$. Both types act efficiently. Cases 2, 3 and 4 are less straightforward. For the remainder of the main body of the paper, we focus on case 3, in which acting is efficient if and only if $\max\{x, y\} = 1$. Cases 2 and 4 follow in the online appendix. We would like to highlight that our main results, propositions 1 and 2, do not condition on any particular case.

Optimal Maximum Punishment. First, notice that it is efficient to act whenever $x = 1$, and, therefore, the principal should, optimally, not punish when $x = 1$. 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 honest agent is going to follow the efficient schedule, while the dishonest agent always acts. In particular, $a^d(-1, -1) = a^h(-1, 1) = 1$. 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 principal will not punish the agent upon seeing a negative outcome and a negative verifiable information, for any $\overline{F} > 0$.

Having solved this trivial case, the remainder of the paper focuses on the case in which the principal might punish the agent when the project fails and the verifiable information is adverse—that is, if $\frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)} < \gamma^*$.

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

1. What is the minimum punishment, F^d, that prevents the dishonest type from acting when receiving an additional bad unverifiable signal?
2. What is the maximum punishment, F^h, that will allow the honest type from action when receiving an additional good unverifiable signal?

From the agent's best response, we obtain,

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

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 maximum punishment, $\overline{F}^*$, exists and, without loss of generality, $\overline{F}^* \in \{0, F^h, F^d\}$.*

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

If $F^h > F^d$, it is possible to partially deter the dishonest type from acting on $(-1,-1)$ with probability 1, yet having the honest type act on $(-1,1)$. However, it is impossible to have $a^h(-1,1) = 1$ *and* $a^d(-1,-1) = 0$. If that were the case, the principal would not punish on $x = -1$, and the dishonest type would have an incentive to deviate to $a^d(-1,-1) = 1$. The optimum implies either partial deterrence or a moderate chilling effect. If $\overline{F} = F^d$, the dishonest type is partially deterred from acting on $(-1,-1)$. He acts with probability $\eta^d > 0$. If $\overline{F} = F^h$, there is a moderate chilling effect, as the honest type acts with probability $\eta^h < 1$ on $(-1,1)$. On $x = -1$, the principal is indifferent about punishing and not (which pins down η^d). Therefore, she punishes 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 $F^d > F^h$

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

(x,y)	a^h	a^d
(-1,-1)	0	1
(-1,1)	1	1

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

(x,y)	a^h	a^d
(-1,-1)	0	0
(-1,1)	0	1

When $F^h > F^d$

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

(x,y)	a^h	a^d
(-1,-1)	0	η^d
(-1,1)	1	1

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

(x,y)	a^h	a^d
(-1,-1)	0	0
(-1,1)	η^h	1

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

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

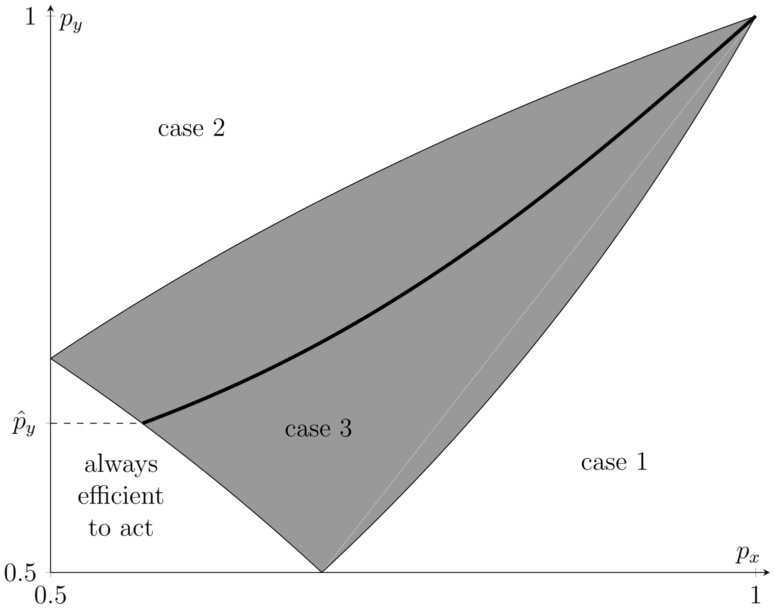
Intuition behind the proof: To understand why the difference is increasing in the precision p_x, first note that both F^d and F^h are decreasing in it. The likelihood of failing conditional on $x = -1$ increases with p_x, and, thus, the agent expects a higher punishment. However, the increase in the punishment and in the probability of failure affect different types in different ways. Due to the misalignment of preferences between the honest and the dishonest agent, F^h falls faster than F^d.

The dishonest agent only suffers indirectly from the higher failure rate—through the higher punishment (the conviction effect). The honest agent suffers also directly—through the failure itself (the outcome effect).

The reasons why the difference is decreasing in p_y is more direct. Following an increase of p_y both outcome and conviction effect encourage the honest agent to act on $(-1, 1)$, thus F^h increases. The conviction effect discourages the dishonest agent from acting on $(-1, -1)$, thus F^d decreases.

Since Lemma 2 drives our main results, we want to point out that the main mechanism behind Lemma 2—when x becomes more precise, the honest agent suffers from two effects while the dishonest agent suffers from only one—is equally applicable beyond the binary signal structure we have. And therefore, our main results can extend to richer signal structures too.

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 (case 3)). On the top left of the shaded region, it is efficient to act iff $y \geq 0$ (case 2); and on the bottom right iff $x \geq 0$ (case 1). 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 critical belief p_x^*. In case 3, and for a given p_y, welfare discontinuously drops at p_x^* (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, $F^d - F^h$, 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. $F^d(p_x^*, p_y, \alpha) = F^h(p_x^*, p_y, \alpha)$ (see (1)). And,
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 $F^d = F^h$ 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 are knife-edge.[13]

Figure 2 reinforces why we chose to highlight 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 relevant case if both the signals are of reasonable and similar quality, and the prior is high. 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 the signals is similar.

Recall that $W^*(p_x, p_y, \gamma, \alpha)$ is a function of the environment $(p_x, p_y, \gamma, \alpha)$. With some abuse of notation, $W^*(p_x)$ denotes $W^*(p_x, p_y, \gamma, \alpha)$ for a fixed p_y, γ, α. Similarly, $W^*(p_y)$ stands for $W^*(p_x, p_y, \gamma, \alpha)$ for a fixed p_x, γ, α.

Proposition 1 *An increase in the precision of the verifiable signal can reduce welfare in non-knife-edge cases. Formally, if p_x^* is a critical belief given (p_y, α), then $\exists \epsilon > 0$ such that, $W^*(p_x) > W(p_x') \; \forall \; p_x \in (p_x^* - \epsilon, p_x^*), p_x' \in (p_x^*, p_x^* + \epsilon)$.*

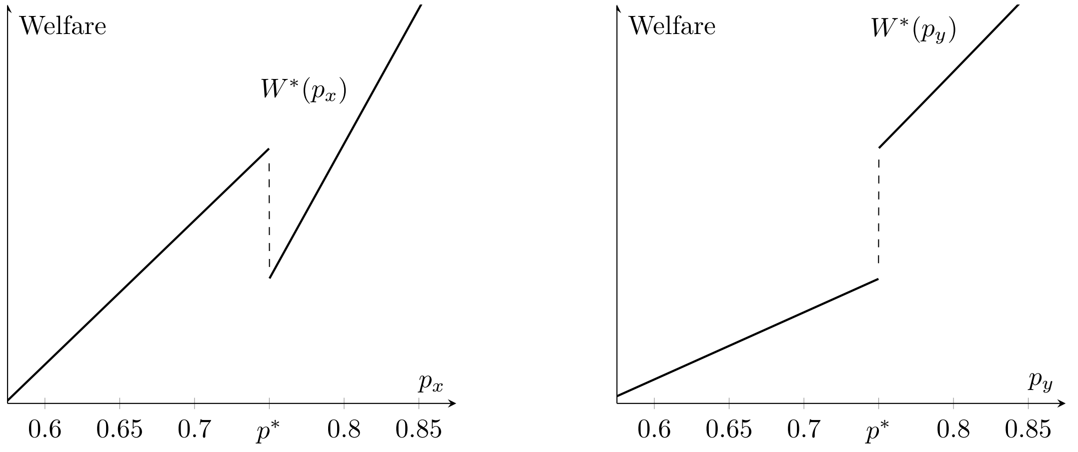
Proposition 1 is driven by the change of sign of $F^d - F^h$. Suppose γ is sufficiently high and p_x is slightly below p_x^*. Then, $F^d < F^h$, and, hence, the optimal maximum punishment restricts the free pass partially without inducing a chilling effect (Table 1b). The dishonest agent is deterred from acting (with probability $1 - \eta^d$ independent of p_x).

Increasing p_x to slightly above p_x^* ceteris paribus implies that $F^d > F^h$. We can no longer restrict the free pass to the dishonest agent without suffering from the chilling effect. We are left with two options. Give a universal free pass or suffer from the chilling effect. Either way, 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 $p_x \to 1$,

[13] The reason to emphasize this aspect is Proposition 1: welfare reduces 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.

Figure 3: *Welfare consequences of improved precision.*



(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). The discontinuity is at the point in which $F^h = F^d$ 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 maximum punishment, $\overline{F}$, 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 project is implemented if, and only if, it is good, and any adverse-selection problem disappears.

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 welfare, as a result of changes in the optimal punishment. For verifiable signals less informative than p_x^*, the optimal punishment was able to partially deter the dishonest agent without inducing any chilling effect. However, when the quality of the verifiable signal improves, to deter the dishonest agent implies a complete chilling effect. Although, given our parameters, that outcome applies in the best equilibrium, it is inferior to the situation with less-precise verifiable information.

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

Proposition 2 *An increase in the precision of the unverifiable signal never reduces welfare. That is,* $W^*(p_y') \geq W^*(p_y)\ \forall p_y' > p_y$.

The main difference between p_x and p_y, and the driver of our results, is in their effect on $F^d - F^h$. While $F^d - F^h$ 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 honest agent to decide in favor of implementing the project by going against x, the verifiable information, relying only on y, the unverifiable information. However, the associated cost is that the dishonest agent may implement a project when both x and y suggest not doing so. Increasing the precision of y helps the designer here. At once, it makes the honest agent more confident about implementing the project by trusting y whilst it discentivizes the dishonest agent from implementing it. Thus, welfare increases.

In contrast, increasing the precision of x 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, and only for the honest 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 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 discontinuous change in welfare. However, in contrast to panel (a), the change is an increase in welfare. This results from the fact that as the quality of unverifiable information increases above the critical threshold, the punishment needed to deter the dishonest agent creates, at most, a partial chilling effect on honest 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.

Finally, we can use Proposition 2 to discuss the efficiency of the optimal equilibrium. For any triple, (p_x, γ, α), welfare coincides with the interim efficient payoff when $p_y \in \{0.5, 1\}$. When the signal is uninformative ($p_y = 0.5$), there is no use for unverifiable information, and punishing the agent for implementing a project that fails when the verifiable signal is bad restores efficiency. On the opposite end, when $p_y = 1$, the agent should rely on the unverifiable information, but, as y is always correct, punishing all failures restores full efficiency. In contrast, for some environments with intermediate values of precision, interim efficiency cannot be reached due to the chilling effect. Therefore, the efficiency loss in the best equilibrium is not monotonic in p_y.

4 Extensions

In this section, we highlight the robustness of our main message to different model specifications. In particular, we consider three alternatives to our baseline model.

Mainly, these extensions offer more flexibility to the principal in deciding when and how to convict the agent.

First, we consider an alternative timeline for the principal to make decisions: what if she could commit to a punishment scheme rather than making a sequentially rational decision? Second, we address the scenarios when the punishment is exogenous: What we cannot adapt the maximum sentence to the precision of the information available? Finally, we change the objective of the principal: what if the principal aimed at punishing the agent for acting against better knowledge?

4.1 The Role of Principal Commitment

While the available punishment is decided upon ex-ante, the principal has no commitment power in our model. The principal first observes the outcome and then decides whether to punish the agent. Her decision is driven by equilibrium reasoning wherein the agent's verifiable information has to be sufficient to induce a pessimistic posterior belief about the agent. A natural question to ask is whether and how our results change if the principal could instead *commit to* a punishment scheme ex-ante.

While the equilibrium behavior changes, we show that our main results are unaffected by the commitment assumption. The agent's choice problem is identical, and, therefore, it still suffices to consider $\overline{F} \in \{F^d, 0\}$ when $F^d > F^h$ and $\overline{F} \in \{F^d, F^h\}$ when $F^d < F^h$. The size of the punishment is also unaffected by the commitment assumption. So are the qualitative results in Table 1.

The main departure from the baseline model is that the principal need not be indifferent anymore. Consequently, the welfare optimal equilibrium implies that $\eta^d = 0$ and $\eta^h = 1$. These changes, in turn, imply that the outcome under $F^d < F^h$ (the bottom row of Table 1) yields a higher welfare than the baseline model. The outcome under $F^d > F^h$ (the top row of Table 1) is identical to that in the baseline case.

Our main results continue to hold for the same reasons as before. A ceteris paribus increase of $p_x < p_x^*$ to $p_x' > p_x^*$ implies a change from the bottom row to the top row of Table 1. Hence, whenever welfare drops discontinuously at p_x^* in the baseline case, welfare drops discontinuously even if the principal had the ability to commit to a punishment scheme. Similarly, a ceteris paribus increase from precision $p_y < p_y^*$ to $p_y' > p_y^*$ implies a change from the top row to the bottom row of Table 1—welfare always increases.

4.2 Exogenous Maximum Punishment

We now look at the case in which the maximum punishment is ex-ante fixed, and thus, not tailored to the precision of the information available. For some applications mentioned in Section 1.1, such an assumption seems realistic. In particular, we are interested in large punishments $\overline{F} \geq \max\{F^d, F^h\}$. We show that our main results—Propositions 1 and 2—remain under such alternative model.

The only difference to our baseline model is that punishment in case of failure is less flexible. It turns out that for $F^h > F^d$, the two models are isomorphic (panels (c) and (d) of Table 2). Under the alternative assumption the welfare maximizing equilibrium implies an effective punishment $\lambda\overline{F} = \overline{F}^*$, where λ is the probability that the principal punishes the agent after observing a failure, and $\overline{F}^*$ is the optimal punishment from the baseline case.

If $F^d > F^h$, results differ slightly (panels (a) and (b) of Table 2). Call $F^d_{y=1}$ the punishment that effectively deters the dishonest agent from acting on $(x, y) = (-1, 1)$. It is straightforward to see that $F^d_{y=1} > F^d$. If $\overline{F} < F^d_{y=1}$, the optimal equilibrium implies the same action profile as the equilibrium with $\overline{F} = F^d$. If $\overline{F} > F^d_{y=1}$ we obtain the deterrence of both, the desired and the undesired actions.

Table 2 is the analogue to Table 1 in the baseline case. The second row is equivalent to the baseline case, the first row differs.

Table 2: Strategy profiles in the optimal equilibria

When $F^d > F^h$

(a) When $\overline{F} > F^d_{y=1}$

(x, y)	a^h	a^d
(-1,-1)	0	0
(-1,1)	0	0

(b) When $\overline{F} \leq F^d_{y=1}$

(x, y)	a^h	a^d
(-1,-1)	0	0
(-1,1)	0	1

When $F^h > F^d$

(c) $\lambda = F^d/\overline{F}$

(x, y)	a^h	a^d
(-1,-1)	0	η^d
(-1,1)	1	1

(d) $\lambda = F^h/\overline{F}$

(x, y)	a^h	a^d
(-1,-1)	0	0
(-1,1)	η^h	1

Proposition 1 and 2 remain as they are: welfare is identical to the baseline case if $F^d < F^h$, but (weakly) worse when $F^d > F^h$. Therefore, the reduction in welfare becomes stronger compared to the baseline case as we move from $p_x < p_x^*$ to $p_x > p_x^*$. Similarly, the increase in welfare becomes stronger compared to the baseline case for an increase in p_y.

Comparing the two models provides an additional insight concerning when a flexible punishment scheme mitigates the chilling effect. If deterrence implies a strong chilling effect (that is $F^d > F^h$), having the flexibility to design the punishment scheme allows the principal to choose between suffering from the chilling effect—by choosing to allow for punishment—and suffering from lack of deterrence—by choosing to give a free pass. That option is not available if the punishment is exogenously fixed.

4.3 Alternative Specification of Principal's Objective

Throughout the paper, we have focused on a principal that aims to infer the *agent's preferences* from the information available to her, and she wants to punish only the dishonest agent. That is, she wants to punish the agent only if she is sufficiently convinced that the agent caused harm because his preferences are not aligned with society's. An alternative specification could be to assume that the principal tries to infer the agent's *(non-verifiable) information* from the (verifiable) information available to her, and wants to punish the agent for acting when he should have exercised restraint.[14]

In this paper, we chose the first specification for the principal's objective primarily for economic reasons: we are interested in the welfare-maximizing equilibria. We show below that welfare under the specification in our baseline model is higher than under the latter. Furthermore, in this section, we also demonstrate that the intuition behind our model and our main comparative statics hold regardless of which payoff specification is used.

Let the agent be punished if the principal is sufficiently convinced that the agent's signal was $(-1,-1)$ when the agent took an action, $a = 1$, resulting in a bad outcome, $\theta = -1$. That is, the principal punishes if $q := \mathbb{P}(y = 1|\theta = -1, a = 1, x = -1) \leq \gamma^*$.

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

Proposition 3 *Ceteris paribus, society's ex-ante welfare in the welfare-maximizing equilibrium is weakly higher if the principal aims at punishing dishonest agents than if the principal aims at punishing the agent for acting against better information.*

The intuition behind the result comes from the equilibrium action profile described in Table 3 below.

The three main differences in the equilibrium behavior are (i) if $F^d > F^h$ and $F = F^d$, the dishonest agent acts with positive probability η_1 (as opposed to 0-probability in the baseline case) on $(-1,-1)$; (ii) if $F^d < F^h$, the optimal punishment is always F^d; and (iii) the probability with which the dishonest agent

[14] Within legal philosophy, the specification used in this paper is called *subjective mens rea*, while the alternative is called *objective mens rea*. The philosophical debate regarding which is a better notion of mens rea has not been settled to the best of our knowledge. See, for example, Brudner (2008), Roach (2012), Simester and Smith (1996). An overview of the concepts and their philosophical foundations is in Tebbit (1999). A general discussion of objectivism and subjectivism is, e.g., in Gordon (1974), Norrie (1992).

Table 3: Strategy profiles in the optimal equilibria

When $F^d > F^h$

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

(x, y)	a^h	a^d
(-1,-1)	0	1
(-1,1)	1	1

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

(x, y)	a^h	a^d
(-1,-1)	0	η_1
(-1,1)	0	1

When $F^h > F^d$

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

(x, y)	a^h	a^d
(-1,-1)	0	η_2
(-1,1)	1	1

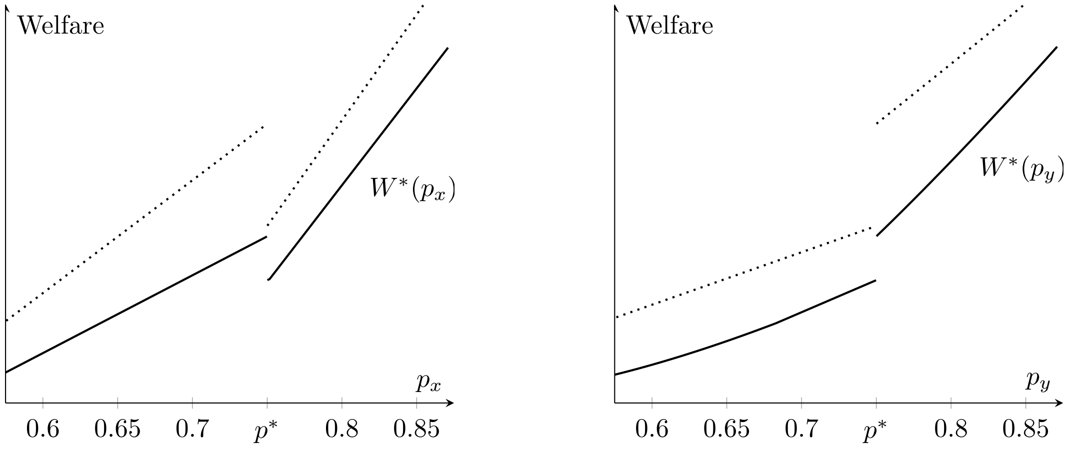
acts on $(-1, -1)$ if $F^d < F^h$ is η_2 which is larger than η^d that we used in the baseline case. The three properties immediately imply Proposition 3. We discuss the differences in turn.

First, assume that $F^d > F^h$ and $\overline{F} = F^d$. Then, the dishonest agent acts on $(-1, -1)$ with positive probability. In our baseline case, we have only the dishonest agent acting on $x = -1$. A punishment of F^d deters the honest agent effectively. In the optimal equilibrium, the dishonest agent acts only when $y = 1$ and, thus, when it is efficient to act. If the principal instead aims at punishing for acting against better information, she does not punish the agent despite being certain that his preferences are not aligned with society's. If the dishonest agent is not punished, however, he also acts on information $(-1, -1)$. Thus, the equilibrium of the baseline case is not sustainable anymore. Instead, the dishonest agent chooses to act with positive probability on the information $(-1, -1)$ to make the principal indifferent between punishing and not.

Second, if $F^d < F^h$, the optimal punishment is always F^d. For any $F < F^h$, regardless of the principal's behavior, agents' behavior will be governed by Table 1a for the same reasons as before. This punishment is obviously suboptimal because, by setting $\overline{F} = F^d$, we can have $a^d(-1, -1) = \eta_2 < 1$ and $a^h(-1, 1) = 1$. The dishonest agent acts on $(-1, -1)$ to make the principal indifferent between punishing and not. In the baseline case, the likelihood of a dishonest type acting on $(-1, -1)$ and a dishonest type acting on $(-1, 1)$ (which was 1) has to balance the likelihood of an honest type acting on $(-1, 1)$. Under the new objective, the likelihood of a dishonest type acting on $(-1, -1)$, has to balance the likelihood of *any* type acting on $(-1, 1)$, which is obviously larger. As a result, $\eta_2 > \eta^d$.

To summarize, if the principal's objective is to convict only agents for implementing the project on information that recommends not doing so, then the probability of the dishonest agent implementing a project is higher than if the principal's objective is to convict only dishonest agents. Reduction in welfare follows.

Figure 4: *Welfare consequences of improved precision when the principal 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 are the value for when the principal aims to convict only agents that acted against better information. Dotted lines are the values for the baseline case. Parameters: $\gamma^* = 1/2, \gamma = 11/20, \alpha = 9/13$ and $p_y = 3/4$ (left panel), $p_x = 3/4$ (right panel).

We now present the effects of changes in information quality for the alternative specification of the principal's objective, our second result for this extension.

Proposition 4 *Suppose that it is efficient to act if and only if $x + y \geq 0$. Ceteris paribus, an increase in precision from $p_x < p_x^*$ to $p_x' > p_x^*$ may reduce welfare. Also, ceteris paribus and for a wide range of parameters, an increase in precision of the unverifiable information leads to welfare increase; yet for* some *(p_x, α, γ), there exists a range $\gamma^* \in (\underline{\gamma}^*, \overline{\gamma}^*)$ such that welfare decreases for an increase from some $p_y < p_y^*$ to some $p_y' > p_y^*$. Moreover, $\underline{\gamma}^* \geq 1 - p_y^*$ and $\overline{\gamma^*} < 1/2$.*[15]

We provide a tight characterization in the appendix. The reasoning for welfare losses as we increase the precision of the verifiable information is identical to that for the baseline setting. In contrast, while increases in the precision of the unverifiable information are unambiguously good for welfare in the baseline model, they may harm welfare under the alternative payoff specification for some parameter values. The previous comparative statics hold, however, if the principal is lenient—i.e., γ^* is low—or strict—i.e., γ^* is high.

If γ^* is low, to punish the agent the principal has to be almost sure he received the

[15] In these comparative statics, it is assumed that we continue to remain in the baseline case: it is efficient to act if and only if $x + y \geq 0$.

signal $(-1,-1)$. In response, the agent acts with probability 1 on the signal. Then, the increase in precision of unverifiable information implies a pure information effect, which is unambiguously good.

If γ^* is high, to punish the agent a slight belief of the principal that the agent received a negative unverifiable signal is sufficient. In that case, the agent's additional potential to leverage against his own latent information is limited, and the results from the baseline case remain.

Whether either of the above cases applies *for all* γ^* or whether there is a range of γ^* for which an increase in the precision p_y can harm welfare, too, depends on the particular parameters. However, we want to emphasize that the range of γ^* where an increase in p_y may harm welfare is empty for a wide range of parameters. For example, if $\alpha \geq p_x$, welfare unambiguously increases in p_y. We provide a detailed derivation in the appendix.

Figure 4 is the analogue of Figure 3 for this alternative payoff specification for the principal. For comparison, we plot the results of the baseline case in the same graph.

4.4 When the principal can punish for inaction

Throughout the paper, we have assumed that the principal can only punish the agent when $a = 1$ and $\theta = 0$. Our choice is mainly motivated by realism (for recent experimental evidence, see Cox et al., 2017). There is still an ongoing debate concerning whether not punishing on inaction stems from a cognitive bias or rational behaviour (for an overview, see Woollard, 2019). Thus, it is a worthwhile exercise (at least theoretically) to allow the principal to punish the agent for not acting. That is, suppose that the principal always see θ and x regardless of whether the agent acted or not. The principal would ideally like to punish the agent for displaying excessive caution by not acting. However, given the lack of commitment on our principal'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 honest agent acts with strictly positive probability, a dishonest agent finds it optimal to act. Therefore, for any realization, inaction only increases the likelihood of the agent being honest, and the principal's posterior over the agent being honest must be weakly higher than her prior. Therefore, under the baseline assumption of γ being sufficiently high, the principal chooses to not punish the agent, even if allowed to do so.

5 Conclusion

This paper highlights that it is not merely the amount of information, but also its nature, that has important welfare consequences. The main conflict we studied is between the agent and the principal that tries to discipline his behavior. We focus on information of two different natures, depending on whether it is verifiable by the principal. 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 increases welfare, increasing the precision of verifiable information may have dire consequences.

Our findings straightforwardly extend to a variety of settings. A standard application would be to consider legal systems, with the principal representing the court. Beyond this setting, whether we consider politicians aiming for 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 her 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 and in the popular press.[16] Our results highlight 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 of its interaction with the provision of information of different natures.

A Appendix

Let $q^h := a^h(-1,1)$ and $q^d := a^d(-1,-1)$. Notice that if $F(-1) < (>) F^d$, then $q^d = 1(0)$, and if $F(-1) < (>) F^h$, then $q^h = 1(0)$. Let η and η_2 be defined by,

$$\frac{\gamma(1-p_y)}{\gamma(1-p_y)+(1-\gamma)(1-p_y+p_y\eta^d)} = \gamma^* \tag{2}$$

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

If $q^h = 1$ then $q^d = \eta^d \implies \tilde{\gamma}(-1) = \gamma^*$, making the principal indifferent between any punishment. If $q^d = 0$ and $a^d(-1,1) = 1$, then $q^h = \eta^h \implies \tilde{\gamma}(-1) = \gamma^*$.

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

A.1 Proof of Lemma 1

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

Proof. $q^d = 0 \implies \tilde{\gamma}(-1) = \gamma > \gamma^*$. Therefore, $F(-1) = 0$. Therefore, D would deviate to play $q^d = 1$. Therefore, $q^d > 0$. Also, $q^d = 1 \implies \tilde{\gamma}(-1) = \frac{\gamma(1-p_y)}{\gamma(1-p_y)+1-p_y} < \gamma^*$. Therefore, $F(-1) = \overline{F}$. Notice that $\overline{F} < F^d \implies q^d = 1$. And, if $\overline{F} \geq F(-1) > F^d \implies q^d = 0 \implies \tilde{\gamma}(-1) = \gamma > \gamma^*$. This would imply that $F(-1) = 0$, a contradiction. Therefore, if $\overline{F} > F^d$, the D type would mix to have $\tilde{\gamma}(-1) = \gamma^*$—i.e., $q^d = \eta^d$, so that $F(-1) = F^d$. In the case when $F = F^d$, $q^d \in [\eta^d, 1]$. In this case, $q^d = \eta^d$ is the society's preferred equilibrium. □

Claim 2 $F^h > F^d$ *and* $q^d > 0 \implies q^h = 1$.

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

Claim 3 *If* $F^h > F^d$, $\overline{F}^* \in \{F^d, F^h\}$.

Proof. First, notice that $F(-1) < F^d \implies q^h = q^d = 1$. Instead, with $F(-1) = F^d \implies q^h = 1, q^d = \eta^d$, giving us a strict welfare improvement.

If $F(-1) \in (F^d, F^h)$, then $q^h = 1$ and $q^d = 0$. But then, $\tilde{\gamma}(-1) = \gamma > \gamma^* \implies F(-1) = 0$, a contradiction. Therefore, $F(-1) \notin (F^d, F^h)$ in equilibrium.

If $F(-1) > F^h$ then $q^h = q^d = 0$. Instead, $F(-1) = F^h$ provides a strict welfare improvement by having $q^h \in [0, \eta^h], q^d = 0$. The welfare-maximizing choice is to have $q^h = \eta^h$. $q^h \leq \eta^h$ because, otherwise, $\tilde{\gamma}(-1) > \gamma^*$, and, therefore, $F(-1) = 0$, a contradiction. □

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

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

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

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

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

A.2 Proof of Proposition 1 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 $\delta(p_x, p_y) := F^d - F^h$.

Proof of Lemma 2.

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

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

Proof of Proposition 1. At p_x^*, $F^d(p_x^*, p_y, \alpha) = F^h(p_x^*, p_y, \alpha)$. Henceforth, we will suppress the dependence on (p_y, α).

Suppose that $p_1 < p_x^* < p_2$. Therefore, $F^d(p_1) < F^h(p_1)$ *and* $F^d(p_2) > F^h(p_2)$ by Lemma 2.

Case 1: $\overline{F}^*(p_2) = 0$.[17]

Hence, $q^d(p_2) = q^h(p_2) = 1$, By Claim 3, $\overline{F}^*(p_1) \in \{F^h(p_1), F^d(p_1)\}$. Suppose that $F(-1) = F^d(p_1)$. Therefore, $q^d(p_1) = \eta^d$ and $q^h(p_1) = 1$. Notice that (2) features no dependence on p_1 and η^d is strictly less than 1.

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

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

Therefore,

$$\begin{aligned}
W_1 - W_2 =& (p_1 - p_2)[\alpha(1-p_y) + (1-\alpha)p_y] \\
& + (1-\gamma)\Bigg[\eta^d\left[\alpha(1-p_1)(1-p_y) - (1-\alpha)p_1p_y\right] \\
& - \left[\alpha(1-p_2)(1-p_y) - (1-a)p_2p_y\right]\Bigg]
\end{aligned}$$

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

$$W_1 - W_2 = (1-\gamma)(1-\eta^d)[(1-\alpha)p_yp_1 - \alpha(1-p_y)(1-p_1)] + o(\delta).$$

Since it is inefficient to act on $(-1,-1)$, $\beta_{-1,-1} = \frac{\alpha(1-p_y)(1-p_1)}{\alpha(1-p_y)(1-p_1)+(1-\alpha)p_yp_1} < \frac{1}{2}$. Equivalently, $(1-\alpha)p_yp_1 > \alpha(1-p_y)(1-p_1)$. Therefore, $W_1 > W_2$. Lastly, if $\overline{F}^*(p_1) = F^h(p_1)$, then $W^*(p_2, p_y, \gamma, \alpha) \geq W_1 > W_2 = W^*(p_2, p_y, \gamma, \alpha)$.

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

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

Therefore, $q^d(p_2) = q^h(p_2) = 0$. Setting $F(-1) = F^h(p_1)$, we have $q^d(p_1) = 0$ *and* $q^h(p_1) = \eta^h > 0$. Since the only change is that the honest type acts on $(-1, 1)$ with probability η^h, 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.3 Proof of Proposition 2

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

The principal observes the realization of 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 principal provides a free pass on $x = 1$ 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 welfare.

Argument 2 **($x = -1$).** Compare two environments with p_y, p'_y such that $p'_y > p_y$. First, assume that $\overline{F}^* = 0$ for both levels. Increasing precision does not change $a^d(\cdot, \cdot)$, but projects implemented by the honest agent fail less often. Second, assume that $\overline{F}^* = F^h$ 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 $\overline{F}^* = F^d$ for both levels. Since η^d decreases in p_y, the dishonest agent's actions on y improve from a welfare perspective, while the honest agent's decisions can only improve by Lemma 2. Welfare increases. What remains is to show that welfare improves as we move from $\overline{F} = 0$ to $\overline{F} = F^\omega$. A change from $\overline{F} = F^0$ to $\overline{F} = F^d$ occurs either if $F^d > F^h$ or if $F^d = F^h$. In the former case, both equilibria are available, and the switch occurs because $\overline{W}(F^d)$ overtakes $\overline{W}(0)$, a welfare improvement. In the latter case, welfare improves because the only behavioral change is that the dishonest agent selects the inefficient action less often. Finally, a change from $\overline{F} = 0$ to $\overline{F} = F^h$ can occur only at $F^d = F^h$, and since, by construction, $\overline{F} = F^h$ welfare dominates $\overline{F} = F^d$, and the proof is complete. □

B Additional Material on Section 4

Equilibrium Characterization

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

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

If $\overline{F} = F^d < F^h$, the optimal equilibrium implies that $a^h(-1,1) = 1$, $a^d(-1,1) = 1$ and $a^d(-1,-1) = \eta_1$ with

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

If $F^d > F^h$, any punishment below F^d implies that the dishonest agent is never deterred from acting. If, in addition, $\overline{F} > F^h$, the honest agent is deterred from acting on $(-1,1)$, which is clearly worse. Thus, an optimal equilibrium exists for either $\overline{F} = 0$ or $\overline{F} = F^d$. The principal's indifference condition implies η_1.

If $F^d < F^h$, a punishment above F^d does not improve upon F^d, as it would lead to only the honest type acting on $x = -1$, which, in turn, implies that the principal does not punish. Conditional on not facing punishment, the dishonest 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. The only equilibrium with *effective* punishment is F^d (through mixing). It yields no better outcome than the optimal equilibrium under $\overline{F} = F^d$. Thus, it is sufficient to consider $\overline{F} = F^d$ only if $F^d < F^h$. The principal's indifference condition implies η_2.

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

To see that result, observe that action profiles are identical, apart from the dishonest agent's decision on $(-1,-1)$. If $F^d < F^h$ she chooses $\eta_1 > 0$ for the principal's objective assumed in this section (punishing for acting on wrong information) and 0 under the principal's objective in the baseline model.[18] Since acting is inefficient for the information $(-1,-1)$, the alternative objective is welfare-inferior. If $F^d > F^h$, the agent chooses

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

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

Again, the alternative objective is welfare-inferior.

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

When is welfare unambigously increasing in the precision of p_y?

Straightforward comparison yields that the only case we need to consider is the case in which $\overline{F} = F^d$ 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.

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.

Define

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

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

$$(1-\gamma)(\eta_2-\eta_1) \leq \frac{\gamma[\alpha(1-p_x)p_y^* - (1-\alpha)p_x(1-p_y^*)]}{[(1-\alpha)p_xp_y^* - \alpha(1-p_x)(1-p_y^*)]}. \tag{4}$$

The RHS of the above is independent of γ^* and positve. Moreover,

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

Solving 4 implies precise conditions on primitives as to whether welfare is increasing around p_y^*. With some algebra, we obtain a set of potential (non-binding) sufficient conditions for $\Delta > 0$. These are—$\Delta > 0$ if any of the following holds,

- $\gamma^* > 1/2$
- $\alpha > p_x$
- $\alpha(1-p_x)p_y^* - (1-\alpha)p_x(1-p_y^*)] - (1-\gamma)[(1-\alpha)p_xp_y^* - \alpha(1-p_x)(1-p_y^*)] > 0.$

References

G.-M. Angeletos and A. Pavan. Efficient use of information and social value of information. *Econometrica*, 75(4):1103–1142, 2007.

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

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

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.

A. Brandenburger and B. Polak. When managers cover their posteriors: Making the decisions the market wants to see. *The RAND Journal of Economics*, pages 523–541, 1996.

A. Brudner. Subjective fault for crime: A reinterpretation. *Legal Theory*, 14:1, 2008.

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

Y.-K. Che, W. Dessein, and N. Kartik. Pandering to persuade. *American Economic Review*, 103(1):47–79, 2013.

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.

J. Crémer. Arm's length relationships. *The Quarterly Journal of Economics*, 110 (2):275–295, 1995.

M. Dewatripont, I. Jewitt, and J. Tirole. The economics of career concerns, part i: Comparing information structures. *The Review of Economic Studies*, 66(1): 183–198, 1999.

A. Frankel. Selecting applicants. *Working Paper, University of Chicago*, 2020.

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

G. H. Gordon. Subjective and objective mens rea. *Crim. LQ*, 17:355, 1974.

B. Holmström. Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182, 1999. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2566954.

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.

S. K. Kim. Efficiency of an information system in an agency model. *Econometrica*, 63(1):89–102, 1995. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951698.

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.

E. P. Lazear. Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics*, 121(3):1029–1061, 2006.

D. Mookherjee and I. P. L. Png. Marginal deterrence in enforcement of law. *Journal of Political Economy*, 102(5):1039–1066, 1994. doi: 10.1086/261963. URL https://doi.org/10.1086/261963.

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

A. Norrie. Subjectivism, objectivism and the limits of criminal recklessness. *Oxford Journal of Legal Studies*, 12(1):45–58, 1992. ISSN 01436503, 14643820. URL http://www.jstor.org/stable/764569.

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

A. Polinsky and S. Shavell. The optimal use of fines and imprisonment. *Journal of Public Economics*, 24(1):89 – 99, 1984. ISSN 0047-2727. doi: https://doi.org/10.1016/0047-2727(84)90006-9. URL http://www.sciencedirect.com/science/article/pii/0047272784900069.

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

C. Prato and B. Strulovici. The hidden cost of direct democracy: How ballot initiatives affect politicians' selection and incentives. *Journal of Theoretical Politics*, 29(3):440–466, 2017.

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

K. Roach. *Criminal law*. Irwin Law eLibrary. Irwin Law, Toronto [Ont.], 5th ed. edition, 2012. ISBN 1-283-69440-9.

A. P. Simester and A. Smith. Harm and culpability. 1996.

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

M. Tebbit. *The Philosophy of Law: An Encyclopedia*. 1st ed.. edition, 1999. ISBN 9780203800225.

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

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.

C Online Appendix

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 $y = 1$.

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.

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

We call this case as the "pivotal intuition" case. In this case, the goal is to deter the D type from acting on $(-1, -1)$ and $(1, -1)$ while incentivizing the H 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 goal of this section is to demonstrate that a similar comparative static—increasing p_x can reduce welfare but increasing p_y always improves welfare—holds in this environment too.

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 Claim 3 and 4 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 $F^h(1, 1)$ denote the largest fine up to which the honest type will act on $(1, 1)$ and $F^d(1, -1)$ denote the smallest fine necessary to deter the dishonest type from acting on $(1, -1)$.

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

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

As in the baseline case, for the main comparative static, the difference between $F^d(-1, -1)$ and $F^h(-1, 1)$, as well as the difference between $F^d(1, -1)$ and $F^h(1, 1)$

matters in determining the optimal fine. Define,

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

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 1 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 $a^d(-1,-1) = \eta^d < 1$ *and* $a^h(-1,1) = 1$. 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 $a^h(-1,1) = 1$ *and* $a^d(-1,-1) < 1$. In particular, if γ is sufficiently high we set $\bar{F}^*(-1) = 0$ and obtain $a^h(-1,1) = a^d(-1,-1) = 1$. That is, the honest type acts on $(-1,1)$ but the price we pay is that the dishonest type acts with probability 1 on $(-1,-1)$.

Notice that this reasoning is identical, when $x = -1$, in the case where intuition is pivotal. Also, if $\Delta^1(p_x,p_y) < 0$, then we can have $a^d(1,-1) = \eta^d$ *and* $a^h(1,1) = 1$. 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. $F^d(-1,-1) = F^h(-1,1)$, 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 1 and 2 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 5 *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,γ,α) society's welfare $W^*(p_x,p_y,\gamma,\alpha)$ is non-decreasing in p_y.*

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

First of all, in this case, we can set F to be larger than $F^d(-1,1)$ 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 $a^h(1,1) = 1$ and $a^d(1,-1) = 0$.

Unsurprisingly,
ranked. Lastly, since $\Delta^1(p_x, p_y)$ is decreasing in p_x, increasing p_x cannot reduce welfare in this case.

C.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 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 a. **It is efficient to act iff $x = 1$.** In this case, Argument 1 of Appendix A.3 applies for both $x = 1$ and $x = -1$

Case b. **It is efficient to act iff $y = 1$.** For $x = -1$ this case is identical to the baseline case. Argument 2 of Appendix A.3 applies. Also conditional on $x = 1$ the situation is as in the baseline case for $x = -1$. Since $\Delta^1(p_x, p_y)$ decreases in p_y, Argument 2 of Appendix A.3 applies directly.

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

Case d. **It is efficient to act iff $x+y = 2$.** The case for $x = 1$ is as in the previous case and Argument 2 of Appendix A.3 applies. For $x = -1$ Argument 1 of Appendix A.3 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 $\mathbb{P}(\theta = 1 | x = -1, y = 1) = 1/2$. 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 dishonest type without any chilling effect is possible by the fine F^d conditional on $x = -1$ since the honest type (and the society) are indifferent between taking an action or not taking an action. Yet, F^d is also feasible. Thus, the transition is continuous.

From 1 to 4. Take $\hat{p}_y$ such that $\mathbb{P}(\theta = 1 | x = 1, y = -1) = 1/2$. Then, the society is indifferent between everyone acting on $(-1, 1)$, no one acting on it, or only the dishonest 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.