Di Pei, Harry; Strulovici, Bruno

**Working Paper**

# Strategic abuse and accuser credibility

CSIO Working Paper, No. 0145

**Provided in Cooperation with:**
Department of Economics - Center for the Study of Industrial Organization (CSIO), Northwestern University

*Suggested Citation:* Di Pei, Harry; Strulovici, Bruno (2018) : Strategic abuse and accuser credibility, CSIO Working Paper, No. 0145, Northwestern University, Center for the Study of Industrial Organization (CSIO), Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/213452

# Strategic Abuse and Accuser Credibility

Harry Pei* Bruno Strulovici†

November 3, 2018
Preliminary, comments welcome.

**Abstract**

We study the interaction between a potential offender's (*principal*) incentive to commit crimes and the potential victims' (*agents*) incentive to report crime. The probability of crime and the credibility of reports are endogenously determined in equilibrium, and the principal is convicted if found sufficiently likely of having committed crime by a Bayesian judge. We show that when the punishment in case of a conviction is sufficiently large, the principal's decisions to commit crimes are strategic substitutes, while the agents' decisions to report crime are strategic complements. The tension between agents' coordination motive and the negative correlation of their private information causes their reports to become arbitrarily uninformative in equilibrium and lead to a significant probability of crime. The occurrence of crime and lack of report credibility can be mitigated by *reducing* the punishment to a convicted principal or by rewarding lone accusers.

**Keywords:** communication informativeness, coordination, negative correlation, information aggregation, law enforcement

**JEL Codes:** D82, D83, K42.

*Department of Economics, Northwestern University. harrydp@northwestern.edu
†Department of Economics, Northwestern University. b-strulovici@northwestern.edu

# 1 Introduction

Abuses and assaults are often hard to prove with incontrovertible evidence. To mitigate this difficulty, judges, firms, other organizations, and the public at large, often use the number of accusations leveled against an individual to assess their credibility. The presumption is that the accumulation of claims against an individual makes it more likely that he/she was guilty of at least some of these claims.

This paper revisits this presumption in an environment between a rational potential abuser (or *principal*) and a number of potential victims (or *agents*) who may have private motives to accuse the principal irrespective of the abuse. We demonstrate, perhaps paradoxically, that when the principal has much to lose from being convicted of assault, environments with a higher number of potential victims reduce, and can even destroy, the credibility of individual reports against the principal, and can dramatically increase the probability of abuse relative to an environment with fewer potential victims.[1]

In our setting, agents incur a cost from accusing the principal unless the latter is deemed sufficiently likely of having abused at least some agent. If, for instance, the agents are the principal's subordinates, they may face retaliation for accusing him unless he is convicted of abuse and removed from power. When considering whether to accuse the principal, an agent trades off the expected cost of retaliation with the benefit from punishing his abuser and, possibly, a private benefit of uncertain magnitude from harming or getting rid of the principal, which is independent of the abuse.

The principal's likelihood of guilt is a function of the number and credibility of agents' reports, which are endogenously determined in equilibrium. When more numerous reports of abuse increase the posterior likelihood of wrongdoing, a rational principal strategically commits fewer abuses in order to reduce the expected number of reports filed against him.

This strategic restraint by the principal causes agents' reports to become less credible and, paradoxically, increases the probability with which abuses happen in equilibrium. To see this, consider the case of two agents and suppose that two reports are required to get the principal convicted.[2] In this context, suppose that an agent has been abused. This agent knows that the other agent is

[1] Our equilibrium analysis sheds light on the incentives to commit and report crimes when all individuals understand the rules of the game and the consequences of their actions. As such, our results may be better suited to describe environments in which the institutions, laws, and norms that punish abusers have been in place for a long time and everyone understands them. This stands in contrast to situations involving recent institutional changes, in which the repercussions of an assault, a victims' ability and willingness to file a report against a presumed abuser, and the credibility of such report, may be incorrectly evaluated by some of the individuals involved in the interaction.

[2] In our analysis, the link between reports and conviction is determined by the posterior probability that the principal is guilty of at least one abuse after observing the reports. When the punishment following a conviction is sufficiently large, we show that in equilibrium two reports are required to convict the principal in all equilibria that survive a moderate refinement (namely, *monotonicity* and *responsiveness* defined in subsection 3.1).

unlikely of being abused, since the principal wishes to avoid generating two reports against him. The first agent thus knows that his accusation of the principal, should he formulate one, is unlikely to be complemented by a second accusation, and is thus more likely to result in retaliation against him than in punishment against the principal. This reduces the abused agent's incentive to file an accusation. A similar phenomenon arises in the reverse direction: an agent who has not been abused knows that the other agent is more likely of being abused and, hence, of accusing the principal. If the first agent holds an independent grudge against the principal, this increases his incentive to file an accusation.

In summary, a negative correlation emerges between agents' private information due to the principal's strategic restraint. This negative correlation, combined with the endogenous complementarity of the agents' reports, undermines the agents' credibility as they try to coordinate their reports regardless of whether abuses have actually taken place.

The lowered credibility of the agents' reports has a further perverse effect: the principal can more easily abuse one of the agents since reports against him carry less information. We show that the equilibrium probability of abuse is strictly higher in a setting with two agents compared to that in the single-agent benchmark. More generally, environments with a higher number of potential victims exhibit equilibria in which reports are less credible, less informative even after aggregating all agents' reports, and suffer from a higher probability of abuse.

This logic leads to an extreme result as the punishment in case the principal is convicted becomes arbitrarily large. When there are multiple agents, the agents' incentives to coordinate their reports and the negative correlation across their private experience of abuse cause their combined reports to become arbitrarily uninformative. As a result, the probability of abuse converges to a significant level. This result stands in sharp contrast with the single-agent case, in which a report of abuse becomes arbitrarily informative and the probability of abuse vanishes to zero as the punishment to a convicted principal becomes arbitrarily large.

Several remedies are studied to restore the informativeness of reports in environments with multiple agents, which offset agents' coordination motives or the negative correlation induced by the principal's strategic restraint. First, we consider the inclusion of transfers to the agents by a social planner. Intuitively, rewarding an agent who stands alone accusing the principal can offset the retaliation cost that he would face for failing to convict the principal. Rewards of this kind can restore the informativeness of the agents' reports close to the level in the single-agent setting.[3] However, these

[3] We will show that the equilibrium level of informativeness under this transfer scheme coincides with that under the single-agent benchmark when the punishment to the convicted principal becomes arbitrarily large.

transfer schemes are not budget balanced. They subject a social planner to losses and create incentives for collusion between the various individuals involved in the mechanism.

In fact, we show that no budget-balanced transfer scheme can fully address the weak credibility of reports in multi-agent settings. Intuitively, a budget-balanced transfer scheme must punish an agent who fails to accuse the principal when the other agent does. In doing so, the scheme reintroduces a coordination motive among the agents: conditional on an agent accusing the principal, it becomes beneficial for the other agent to also accuse the principal in order to avoid the negative transfer. Transfers thus have limited value for addressing the reduced credibility of reports and the increased probability of abuse that stems from an increase in the number of potential victims.

Second, we treat the punishment to the convicted principal as an instrument to lower the probability of abuse. Surprisingly, the probability of abuse may decrease when this punishment is reduced. Intuitively, a principal facing a lower punishment has less incentive to strategically restrain his abusing behavior, and in particular, his decisions to abuse different agents will become strategic complements. This leads to a positive correlation between the agents' private information and based on our previous reasoning, their coordination motives will lead to more credible reports and lower probability of abuse in equilibrium.

One concern of our baseline model is that it fails to take into account the potential heterogeneity in people's propensities to commit crimes. We extend the model to include the presence of *saints* who do not enjoy committing crimes and *serial assaulters* whose relative value from committing crime stands out. In our extension, the principal has private information about his marginal benefit from committing crimes and/or his (perceived) cost of being convicted. Under our formulation, saints have very low or even negative benefits from committing crimes and serial assaulters have high benefits from committing crimes or low perceived costs of being convicted. In equilibrium, the former commit no abuse, while the latter abuse all agents under his power. Our main findings, namely, the endogenous negative correlation between the agents' private information and their coordination motives in filing reports, still apply as long as saints and serial assaulters occur with a low enough probability.

The rest of this article is organized as follows. We conclude this section with a review of the relevant literature in order to clarify our contributions. We describe the baseline model in section 2 and state our main results in section 3. Section 4 examines two potential solutions to restore reporting informativeness. Section 5 studies extensions of the baseline model and section 6 concludes.

### 1.1 Related Literature

In considering the ability to aggregate and elicit information from multiple agents, our paper contributes to various literatures related to collective decision making, coordination, law and economics, and the part of mechanism design concerned with eliciting private information.

First, our analysis propose a novel explanation for the failure of information aggregation. The literature on such failures includes "herding" models, in which social learning is subject to informational externalities (Banerjee 1992, Bikhchandani, Hirshleifer and Welch 1992, Smith and Sørensen 2000) as well as payoff externalities (Scharfstein and Stein 1990, Ottaviani and Sørensen 2000). In these papers, agents may fail to act on their private information, but only because their predecessors' actions are informative. By contrast, in our model agents cannot observe one another, and an increase in the number of agents can result in agents' reports becoming *less* informative, even taken collectively. In herding models, moreover, agents' signals are conditionally independent or positively correlated, whereas in our setting agents' private information (about whether a crime has taken place or not) is negatively correlated in equilibrium. It is the combination of this negative correlation and of agents' motives to coordinate their reports that reduces the informativeness of agents' reports in equilibrium.[4]

Failures of information aggregation may also arise in voting models and, more generally, in the context of collective decision making. For example, it can be caused by informational effects when agent's decision is evaluated conditional on the agent being pivotal, as pointed out by Austen-Smith and Banks (1996) and Bhattacharya (2013), or by individual biases as in Morgan and Stocken (2008). When voters' payoffs from a reform are negatively correlated, Ali, Mihm and Siga (2018) point out that collective decisions may fail to reflect the social optimum for some supermajority rules. Intuitively, if a voter is pivotal, it suggests that many other voters are in favor of the reform, which by negative correlation means that he himself is unlikely to benefit from the reform. In Ali, Mihm and Siga (2018), agents' private benefits from the reform are negatively correlated. In our paper, the negative correlation concerns agents' signals and, other things equal, agents benefit from coordinating their actions. Put more broadly, a distinctive feature of our analysis is that both the voting rule and the correlation structure of the agents' private information are endogenous determined. This stands in contrast to most models on voting, such as the seminal contributions of Feddersen and Pesendorfer (1996,1997,1998), in which these ingredients are exogenous.

[4]Strulovici (2018) studies a sequential learning model in which agents' signals exhibit information attrition: an agent is less likely of having an informative signal, other things equal, if another agent has found such a signal. This information attrition may be viewed as a form negative correlation across agents' signals and also has adverse effects on learning.

Second, the use of normally distributed preference shocks and the presence of agents' coordination motives are reminiscent of Baliga and Sjöström (2004) and Chassang and Padró i Miquel (2010), who demonstrate the significant effects of incomplete information on the equilibrium outcomes of static and dynamic games. In contrast to these papers, our analysis focuses on the possibility of revealing information about an endogenously distributed state of the world (e.g., whether an abuse has taken place) that is orthogonal to those preference shocks, which is absent in their models. The logic underlying our results is absent from this literature. In particular, the lack of informativeness in our model is driven by the interactions between the endogenous negative correlation between the agents' private information and their coordination motives, while their results follow from the contagion arguments arising in global games, à la Carlson and Van Damme (1993), Morris and Shin (1998).

Third, the failure to elicit correlated private information from multiple informed agents stands in sharp contrast to the well-known result of Crémer and McLean (1985, 1988), who show that one can extract all agents' private information via a budget balanced mechanism under a convex independence condition. As noted, budget-balanced transfers fail to elicit whether abuses occurred in our setting. Intuitively, the difference arises because our agents' type is two-dimensional: one dimension concerns whether he was abused; the other concerns his private benefit from accusing the principal. This second dimension causes Crémer and McLean's convex independence condition to fail: agents with the same abuse status but different private benefits hold the same belief about the other agents' types.

Fourth, our paper contributes to the law and economics literature by studying the interactions between a potential criminal's incentives to commit crimes and the potential victims' incentives to report crimes. Recent work in this area includes Silva (2018) and Baliga, Bueno de Mesquita and Wolitzky (2018), who consider settings with multiple potential suspects. These suspects have negatively correlated types, because at most one of them has committed the crime of which they are accused. In these papers, there is only one class of players (potential criminals, unlike the potential criminal and and victims in our setting) and the results bear little resemblance to ours.[5]

Lastly, our inclusion of rational strategic abusers distinguishes our analysis from recent work focused on the incentives of potential victims, in which the potential criminal is non-strategic who mechanically commits crimes with some fixed probability irrespective of the stakes. Lee and Suen (2018),

[5]Other differences between our model and theirs include: In Silva (2018), the distribution of states is exogenous and the judge has access to an informative signal about the state, whereas in our model, the distribution of states is endogenous and the judge cannot use exogenous signals to evaluate the credibility of the agents' reports. In Baliga et al.(2018), the negative correlation between the agents' innocence is exogenous (only one criminal has the opportunity to attack) and the complementarity between the potential criminals' attacking decisions arises due to an imperfect attribution problem faced by the judge. In our model, such negative correlation arises endogenously in all equilibria and the complementarity between the potential victims' reports is driven by the endogenously determined voting rule.

in particular, study the timing of reports by victims and libelers. Their analysis and ours consider complementary aspects of reporting by potential victims. In particular, the strategic restraint which emerges endogenously in our model is necessary for the negative correlation across agents' reports that lies at the heart of our analysis.

## 2 Model

**Primitives:** Consider the following game between a principal, $n$ agents and an evaluator that unfolds in three stages. In stage 1, the principal chooses an $n$-dimensional vector $\boldsymbol{\theta} \equiv (\theta_1, ..., \theta_n) \in \{0,1\}^n$ where $\theta_i = 0$ is interpreted as the principal abusing/committing a crime against agent $i$ and vice versa. In stage 2, agent $i \in \{1, 2, ..., n\}$ privately observes the following three pieces of information before deciding whether to file a report against the principal or not:

1. the principal's choice of $\theta_i \in \{0,1\}$;
2. the realization of a random variable $\omega_i \in \mathbb{R}$;
3. whether he is strategic or mechanical.

We interpret $\omega_i$ as a utility shock that affects agent $i$'s preference towards the principal. We assume that $\omega_1, \omega_2, ..., \omega_n$ are independently and identically distributed according to $\mathcal{N}(\mu, \sigma^2)$ with $\mu \geq 0$ and $\sigma > 0$.[6] Let $\Phi(\cdot)$ be its cdf and $\phi(\cdot)$ be its pdf. Each agent is strategic with probability $\delta \in (0,1)$, in which case he can flexibly choose whether to file a report or not. With complementary probability, the agent is non-strategic and mechanically files a report with probability $\alpha$, where $\alpha$ is an arbitrary real number in $(0,1)$.[7] Whether an agent is strategic or mechanical is independent of $(\omega_1, ..., \omega_n)$ and is also independent of whether other agents are strategic or mechanical. We are primarily interested in settings where mechanical types are rare, namely, $\delta$ being close to 1.

In stage 3, the evaluator observes the vector of reports $\boldsymbol{a} \equiv (a_1, ..., a_n) \in \{0,1\}^n$ and updates his belief about $\Pi_{i=1}^n \theta_i$, i.e. whether the principal is *guilty* (in which case $\Pi_{i=1}^n \theta_i = 0$) or *innocent* (in which case $\Pi_{i=1}^n \theta_i = 1$). Then he chooses $s \in \{0,1\}$ where $s = 0$ means that the principal is convicted and $s = 1$ means that the principal is acquitted.

[6] The assumption that $\mu \geq 0$ is only needed for the results on comparative statics and is not required for the results on the $L \to \infty$ limit.

[7] The presence of the mechanical type is a technical assumption in order to guarantee the existence of non-trivial equilibria in settings with multiple agents. The details will be explained in Appendix B. In Appendix K, we show that our main insights are robust under alternative specifications of the mechanical type's strategy.

**Payoffs:** The evaluator maximizes the expected value of the following quadratic function:

$$-\left(s-(\pi^*-\frac{1}{2})-\Pi_{i=1}^n\theta_i\right)^2, \tag{2.1}$$

where $\pi^* \in (0,1)$ is an exogenous parameter. Intuitively, the evaluator will choose $s \in \{0,1\}$ in order to minimize the distance between $s$ and his bliss point, given by $\pi^* - \frac{1}{2} + \Pi_{i=1}^n\theta_i$. Therefore, he will adopt the following cutoff strategy:

$$s\begin{cases} =0 & \text{if} \quad \Pr\left(\Pi_{i=1}^n\theta_i=0\middle|\boldsymbol{a}\right)>\pi^* \\ \in[0,1] & \text{if} \quad \Pr\left(\Pi_{i=1}^n\theta_i=0\middle|\boldsymbol{a}\right)=\pi^* \\ =1 & \text{if} \quad \Pr\left(\Pi_{i=1}^n\theta_i=0\middle|\boldsymbol{a}\right)<\pi^*, \end{cases} \tag{2.2}$$

namely, he will convict the principal if and only if his posterior belief attaches probability greater or equal to $\pi^*$ to the principal being guilty. For future reference, let

$$l^* \equiv \frac{\pi^*}{1-\pi^*} \tag{2.3}$$

be the critical likelihood ratio for conviction.

The principal's payoff is:

$$\sum_{i=1}^n(1-\theta_i)-L(1-s), \tag{2.4}$$

where $L>0$ is the punishment he receives after being convicted. Strategic agent $i$'s payoff is:

$$s(\omega_i+b\theta_i-ca_i), \tag{2.5}$$

where $b>0$ measures the disutility he suffers from interacting with a principal who has abused him in the past and $c>0$ is his cost of filing a report, which is only incurred when the principal is acquitted but can be avoided if the principal is convicted.

To interpret these payoffs, the principal strict benefits from committing more crimes but will suffer great losses if he is believed to be guilty with high enough probability.[8] This reveals his trade-off when deciding whether to commit crimes or not as well as who to commit crimes against. For the

[8]Despite assuming constant marginal benefit of committing crimes, our results are even stronger when the principal faces decreasing returns from committing additional crimes. Our results also remain robust when the principal's benefit from committing crimes is 0 or even negative with positive probability, as long as the probability with which those types occur is strictly less than $1-\pi^*$.

most part of this paper, we will focus on cases where $L$ is large enough relative to the benefit from committing assaults. This models situations such as the principal is the head of an organization and being convicted can result in the loss of power, or the principal is an influential public figure and being convicted of misconduct could result in the loss of his reputation.

Every agent receives a status quo payoff 0 if the principal is convicted ($s = 0$) and his payoff when the principal is acquitted ($s = 1$) depends on three terms. First, it is affected by the heterogenous tastes towards the principal, which are modeled by the normally distributed preference shocks $\omega_1, \omega_2, ..., \omega_n$.

Second, each agent is more inclined to report against the principal when he has been abused, as captured by the strictly positive $b$. Since $s(\omega_i + b\theta_i - ca_i)$ is agent $i$'s continuation utility at the reporting stage *after* the abuse has taken place, $b$ does not capture the direct utility loss from the abuse, but rather, it should be interpreted as an agent's disutility from continuing to interact with a principal who has abused him in the past, his increased hazards of being abused again in the future, his preference for justice and vengeance, etc.

Third, filing reports are costly to the agents when they fail to convict the principal.[9] In applications where the principal is a powerful person or when the agents are the principal's subordinates, this cost stems from the principal's retaliation which is incurred only when he stays in power. In other applications, such as the principal is the agents' colleague, neighbor or friend, the cost can come from the social stigma on individuals who make public claims on sensitive topics or it can be the monetary cost of going through the judicial process (such as paying the lawyer fee). Common in these applications, a potential victim's cost of reporting is greater when he fails to convict the defendant.

**Remark:** In order to endogenously assess the credibility of reports, an important feature of our model is that agents can file reports no matter whether they have been abused or not. This captures situations where smoking-gun evidence is scarce and the potential victims' claims are hard to verify, which is applicable to workplace discrimination, verbal abuse, sexual harassment, etc.

Due to this reporting credibility problem, whether the principal is convicted or not depends on the probability with which he is guilty according to a Bayesian observer's posterior belief. Depending on the application, the punishment to the principal can be enforced by the criminal justice system (when the principal's misbehavior is against the law), or it can be carried out by the board of directors of a firm (when the manager's behavior hurts the firm's public image and reputation), or it can be dictated by public opinions and is implemented by the local communities via ostracism (when the principal's

[9]For our results to hold, we only need the cost of filing reports to be strictly lower when the principal is convicted compared to the case when the principal is acquitted.

misbehavior is legal but immoral). Therefore, $\pi^*$ measures the society's attitude towards the trade-off between convicting the guilty and acquitting the innocent, which we view as an *exogenous parameter* that reflects the prevailing ideology instead of an endogenous choice variable.[10]

# 3 Analysis & Results

In subsection 3.1, we introduce our solution concept *monotone-responsive equilibrium*, which refines the set of sequential equilibria and exist when $L$ is large enough.[11] To highlight the inefficiencies caused by coordination motives, we compare the case with one agent (subsection 3.2) to that with two agents (subsection 3.3). In particular, we focus on the informativeness of reports and the equilibrium probability of crime. We explain the economic mechanisms behind our results in subsections 3.3 and 3.4. Generalizations of our main insights to settings with more than two agents as well as comparative static results on the number of agents can be found in subsection 5.1.

## 3.1 Solution Concept

Let $q : \{0,1\}^n \to [0,1]$ be the mapping from the vector of reports to the conviction probabilities. The solution concept is monotone-responsive equilibria (henceforth *equilibrium* for short), which are sequential equilibria that satisfy the following two additional requirements:

1. Responsiveness: $q(0,0,...,0) = 0$.
2. Monotonicity: For every $\mathbf{a}, \mathbf{a}' \in \{0,1\}^n$ with $\mathbf{a} \succeq \mathbf{a}'$, we have $q(\mathbf{a}) \geq q(\mathbf{a}')$.

To understand the implications of these refinements, notice that first, responsiveness requires that if no report is filed against the principal, then he will not be convicted. Economically, this fits well into the motivating applications of committing and reporting crimes, i.e. no report will not lead to conviction. Theoretically, it rules out uninteresting equilibria in which the principal commits crimes against all agents with probability 1 and he is convicted with probability 1 under every reporting profile. The assumption that $\alpha \in (0,1)$ implies that every reporting profile will occur with strictly positive probability. As a result, the ex ante probability with which the principal is guilty is strictly less than 1 in every equilibrium that satisfies the responsiveness criteria.

[10] As will become clear later, despite a decrease in $\pi^*$ will reduce the probability of crime, it will increase the fraction of innocent people among those that have been convicted. Moreover, maximizing the informativeness of reports is equivalent to minimizing the probability of crime under the constraint that the ratio of innocent people among those being convicted is below an exogenous upper bound $1 - \pi^*$.

[11] The existence of monotone-responsive equilibria will be addressed in the follow-up subsections both in the single-agent benchmark and in the two-agent scenario.

The monotonicity requirement rules out equilibria in which some agents are more likely to report when they have not been abused and (or) they enjoy interacting with the principal (i.e. their $\omega_i$ and $\theta_i$ are high). Economically, such equilibria are unreasonable since the agents who like the principal the most and have not been abused will file reports, suffer the reporting costs themselves for the purpose of helping the principal to stay in power. This is at odds with the interpretation that $c$ is the loss from the principal's retaliation, the cost of going through the judicial system, etc.

For a theoretical justification, notice that all equilibria are monotone if the principal can privately commit to a retaliation plan against each agent before the game starts, as in Chassang and Padró i Miquel (2018). To be more precise, suppose the principal can choose $\widetilde{c} \equiv (\widetilde{c}_1, ..., \widetilde{c}_n)$, where $\widetilde{c}_i : \{0,1\}^n \to [0,c]$ is a mapping from the set of reporting profiles to the losses suffered by agent $i$ from retaliation and $c$ is the maximal damage the principal can inflict on each individual agent. Agent $i$ can only observe $\widetilde{c}_i$ but not the retaliation plans against other agents. In this scenario, the principal's optimal retaliation plan is bang-bang and he will retaliate to the maximum against messages that increase the evaluator's belief about $\Pi_{i=1}^n \theta_i = 0$ and will not retaliate against other messages.

To conclude this subsection, we present an observation that crime will occur with strictly positive probability in every equilibrium:

**Lemma 3.1.** *In every equilibrium,* $\Pi_{i=1}^n \theta_i = 0$ *occurs with strictly positive probability.*[12]

*Proof of Lemma 3.1:* Suppose towards a contradiction that $(\theta_1, ..., \theta_n) = (1, 1, ..., 1)$ occurs with probability 1. Since a mechanical type always reports and a strategic type with $\omega_i > 0$ has a strictly dominant strategy of not reporting, every $\mathbf{a} \in \{0,1\}^n$ occurs with strictly positive probability. Given the prior probability of $\Pi_{i=1}^n \theta_i = 0$ is 0, the posterior probability of $\Pi_{i=1}^n \theta_i = 0$ under any reporting profile is also 0. Therefore, $s = 1$ occurs with probability 1, which gives the principal a strict incentive to choose $\theta_i = 0$ for every $i \in \{1, 2, ..., n\}$, leading to a contradiction.[13] □

[12] This is related to a well-known conclusion on inspection games (Dresher 1962). In those models, crime cannot happen with zero probability since otherwise, the inspector will have no incentive to conduct costly inspections which will provide the suspect a strict incentive to commit crimes. In our model, crime cannot happen with zero probability since the evaluator's posterior belief will never reach $\pi^*$ no matter how many reports he has observed, and this will provide the principal a strict incentive to commit crimes, leading to a contradiction.

[13] The insight from Lemma 3.1 remains robust in all sequential equilibria and also applies when there are no mechanical types. To see this, for $(\theta_1, ..., \theta_n) = (1, 1, ..., 1)$ to occur with probability 1 in some equilibrium, it must be the case that there exists $i$ such that agent $i$ reports with probability 0 on the equilibrium path. For the principal to have an incentive not to abuse him, he will report with strictly positive probability after being abused and the principal will be convicted with positive probability. However, this implies that agent $i$ will also have a strict incentive to report when $\omega_i$ is sufficiently low, leading to a contradiction.

### 3.2 Single-Agent Benchmark

When there is only one agent, the principal is convicted with strictly positive probability only when the agent files a report. Let $q_s$ be the probability with which $s = 0$ conditional on $a = 1$. If the agent has been abused ($\theta = 0$), then he has an incentive to file a report only when:

$$\omega \leq (1 - q_s)(\omega - c)$$

or equivalently,

$$\omega \leq \omega_s^* \equiv -c\frac{1 - q_s}{q_s}. \tag{3.1}$$

Similarly, if he has not been abused ($\theta = 1$), then he has an incentive to file a report only when:

$$\omega + b \leq (1 - q_s)(\omega + b - c)$$

or equivalently,

$$\omega \leq \omega_s^{**} \equiv -b - c\frac{1 - q_s}{q_s}. \tag{3.2}$$

Comparing (3.1) to (3.2), the distance between $\omega_s^*$ and $\omega_s^{**}$ equals to $b$. The informativeness of the agent's report is measured by the ratio between the probability with which he files a report conditional on $\theta = 0$ and the probability with which he files a report conditional on $\theta = 1$, given by:

$$\mathcal{I}_s \equiv \frac{\Pr(\text{ the agent reports } \mid \theta = 0)}{\Pr(\text{ the agent reports } \mid \theta = 1)} = \frac{\delta\Phi(\omega_s^*) + (1 - \delta)\alpha}{\delta\Phi(\omega_s^{**}) + (1 - \delta)\alpha}. \tag{3.3}$$

In the limiting economy where $\delta \to 1$, this informativeness ratio equals to $\Phi(\omega_s^*)/\Phi(\omega_s^{**})$. Let $\widetilde{\pi}_s$ be the equilibrium probability of crime. The following proposition provides conditions on the existence and uniqueness of equilibrium and offers a characterization:

**Proposition 1.** *A monotone-responsive equilibrium exists if and only if:*

$$\delta L\Big(\Phi(0) - \Phi(-b)\Big) \geq 1. \tag{3.4}$$

*When (3.4) holds with strictly inequality, there exists a unique equilibrium characterized by the quadruple* $(\omega_s^*, \omega_s^{**}, q_s, \widetilde{\pi}_s) \in \mathbb{R} \times \mathbb{R} \times (0, 1] \times [0, \pi^*]$ *which satisfies (3.1), (3.2),*

$$\delta q_s\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) = 1/L \tag{3.5}$$

*and*

$$\frac{\widetilde{\pi}_s}{1-\widetilde{\pi}_s} = \frac{l^*}{\mathcal{I}_s}. \tag{3.6}$$

*Proof of Proposition 1:* When (3.4) fails, then the principal's cost of committing a crime in any monotone-responsive equilibrium is at most:

$$\delta q_s L\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) \leq \delta L\big(\Phi(0) - \Phi(-b)\big),$$

with the maximum on the RHS attained when $q_s = 1$, $\omega_s^* = 0$ and $\omega_s^{**} = -b$. The value of the RHS is strictly less than 1, his benefit from committing a crime, leading to a contradiction.

When (3.4) holds, then for a fixed $c$, notice that $\omega_s^*$ and $\omega_s^{**}$ are strictly increasing in $q_s$. Since $\omega_s^*, \omega_s^{**} \leq 0$, the distance between $\omega_s^*$ and $\omega_s^{**}$ is $b$ and the pdf of $\mathcal{N}(\mu, \sigma^2)$ is strictly increasing in $\omega$ when $\omega < 0$, we know that $\Phi(\omega_s^*) - \Phi(\omega_s^{**})$ is also strictly increasing in $q_s$. This implies that the LHS of (3.5), which is the cost of abusing the agent, is strictly increasing in $q_s$. Inequality (3.4) ensures the existence and uniqueness of $(\omega_s^*, \omega_s^{**}, q_s)$ that solves (3.1), (3.2), (3.5) and (3.6) since the LHS of (3.5) is strictly less than $1/L$ when $q_s = 0$ and is weakly more than $1/L$ when $q_s = 1$.

When (3.4) holds with strictly inequality, the equilibrium level of $q_s$ is interior. Therefore, the probability with which the principal is guilty according to the evaluator's posterior belief equals to $\pi^*$. As a result, $\widetilde{\pi}_s$ is uniquely pinned down by (3.6). □

Next, we assume (3.4) holds with strict inequality and perform comparative statics with respect to the agent's loss from retaliation $c$ and the principal's loss from being convicted $L$. We also examine the limiting scenario where $L \to \infty$ and $\delta \to 1$, which will be useful once comparing it to the two-agent case in the next subsection:

**Proposition 2.** *In the single agent benchmark,*

1. *When $c$ increases, $q_s$ increases, both $\omega_s^*$ and $\omega_s^{**}$ decrease, $\widetilde{\pi}_s$ decreases.*

2. *When $L$ increases, $q_s$ decreases, both $\omega_s^*$ and $\omega_s^{**}$ decrease, $\widetilde{\pi}_s$ decreases.*

3. *Fix $L$ and let $c \to \infty$, we have $q_s \to 1$ and $\omega_s^* \to \omega(L)$ where $\omega(L) \in \mathbb{R}_-$ is pinned down by:*

$$\delta\Big(\Phi\big(\omega(L)\big) - \Phi\big(\omega(L) - b\big)\Big) = 1/L.$$

*and* $\widetilde{\pi}_s \to \widetilde{\pi}(L)$ *where* $\widetilde{\pi}(L)$ *is pinned down by:*

$$\frac{\Phi(\omega(L))}{\Phi(\omega(L)-b)} = l^* \Big/ \frac{\widetilde{\pi}(L)}{1-\widetilde{\pi}(L)}.$$

4. *Fix* $c$ *and let* $L \to \infty$*, we have* $q_s \to 0$*,* $\omega_s^*, \omega_s^{**} \to -\infty$*. Furthermore in the limiting economy where* $\delta \to 1$*, we have* $\mathcal{I}_s \to \infty$ *and* $\widetilde{\pi}_s \to 0$.[14]

The proof is in Appendix A. Before we proceed, it is worth commenting on the limiting result when $L \to \infty$. In particular, the proof makes use of the tail property of normal distributions that for every $b > 0$,

$$\lim_{\omega \to -\infty} \Phi(\omega) \big/ \Phi(\omega - b) \to \infty.$$

That is to say, the tail events are arbitrarily informative. Our finding that the agent's report becomes arbitrarily informative as well as the probability of crime vanishes in the limit extends to any distribution of $\omega$ as long as it has full support with the left tail thinner than exponential distributions.[15]

### 3.3 Main Results: Two-Agent Scenario

In this subsection, we characterize and analyze the game's equilibrium outcomes when there are two agents and compare them to the single-agent benchmark. First of all, we establish the existence of an equilibrium that survives our refinements when $L$ is large enough:

**Proposition 3.** *A monotone-responsive equilibrium exists when* $L$ *is large enough.*

The proof is in Appendix B, which makes use of the Brouwer's fixed point theorem. This is the only part where we need the existence of mechanical types. The proof can be generalized by allowing for any number of agents and alternative specifications of the mechanical types' strategies.[16]

In what follows, we will focus on parameter values under which an equilibrium exists. We state our first main result that establishes some common properties of equilibria. We will also compare the resulting equilibrium outcome to the single-agent benchmark in terms of the agents' strategies, the informativeness of their reports and the equilibrium probability of crime.

[14] For the second part of this statement, we require $\delta \to 1$ at a faster rate compared to $L \to \infty$.

[15] When the distribution of $\omega_i$ is exponential, the value of $\Phi(\omega)/\Phi(\omega - b)$ is constant as $\omega \to -\infty$.

[16] See Proposition B' for the statement of the generalization. We allow, for example, the mechanical type to adopt reporting strategies that are contingent on $(\theta_i, \omega_i)$ as long as the reporting probabilities conditional on each realization of $\theta_i$ is strictly bounded away from 0.

**Theorem 1.** *There exists* $\overline{L} : \mathbb{R}_+ \times (0,1) \to \mathbb{R}_+$ *such that when* $L > \overline{L}(c,\delta)$*, every equilibrium is characterized by a quadruple* $(\omega_m^*, \omega_m^{**}, q_m, \widetilde{\pi}_m) \in \mathbb{R}_- \times \mathbb{R}_- \times (0,1) \times (0,\pi^*)$ *such that:*

1. *For every* $i \in \{1,2\}$*, agent* $i$ *will report in either one of the following two events:* $\Big\{\omega_i \geq \omega_m^*$ *and* $\theta_i = 0\Big\}$ *and* $\Big\{\omega_i \geq \omega_m^{**}$ *and* $\theta_i = 1\Big\}$.

2. *The principal chooses* $(\theta_1,\theta_2) = (1,1)$ *with probability* $1 - \widetilde{\pi}_m$*,* $(\theta_1,\theta_2) = (1,0)$ *with probability* $\widetilde{\pi}_m/2$ *and* $(\theta_1,\theta_2) = (0,1)$ *with probability* $\widetilde{\pi}_m/2$.

3. $q(0,0) = q(0,1) = q(1,0) = 0$ *and* $q(1,1) = q_m$.

4. *Compared to the unique equilibrium in the single-agent benchmark, we have:* $\omega_m^* > \omega_s^*$*,* $\omega_m^{**} > \omega_s^{**}$*,* $q_m > q_s$*,* $\widetilde{\pi}_m > \widetilde{\pi}_s$ *and*

$$\mathcal{I}_s > \mathcal{I}_m \equiv \frac{\Pr(\textit{two agents report} \mid \theta_1\theta_2 = 0)}{\Pr(\textit{two agents report} \mid \theta_1\theta_2 = 1)}.$$

The proof of Theorem 1 consists of three parts, which can be found in Appendices C (symmetry), D (conviction probabilities) and E (comparison to the single-agent benchmark). According to Theorem 1, all equilibria share the following properties when $L$ is large enough. First of all, the agents' strategies are symmetric as they adopt the same pair of reporting thresholds. Second, the principal abuses the two agents with equal probability but will never abuse both at the same time. Third, the principal is convicted with strictly positive probability only when there are two reports. Fourth, compared to the single-agent benchmark, strategic agents are more likely to file reports. Moreover, conditional on all agents unanimously report, the probability of conviction increases. Paradoxically, the equilibrium probability of crime also increases. This is because the aggregate informativeness of the agents' reports, measured by $\mathcal{I}_m$, is strictly lower compared to $\mathcal{I}_s$, the informativeness of report in the single-agent benchmark.

Our second main result examines the informativeness of the agents' reports and the equilibrium probability of crime in the limiting scenario where $L \to \infty$.

**Theorem 2.** *For every* $(c,\delta) \in \mathbb{R}_+ \times (0,1)$ *and* $\epsilon > 0$*, there exists* $\overline{L}_\epsilon(c,\delta) \geq \overline{L}(c,\delta)$*, such that for every* $L > \overline{L}_\epsilon(c,\delta)$*, every equilibrium under parameter configuration* $(L,c,\delta)$ *satisfies*

$$\omega_m^*, \omega_m^{**} < -1/\epsilon, \quad \mathcal{I}_m < 1 + \epsilon \quad \textit{and} \quad \widetilde{\pi}_m \geq \pi^* - \epsilon.$$

The proof is in Appendix F. According to Theorem 2, as the punishment to the convicted becomes sufficiently harsh, the agents' reports are becoming arbitrarily uninformative about the principal's innocence even at the aggregate level. As a result, the probability of crime converges to $\pi^*$. Our conclusion applies as long as the agents' cost of reporting is strictly positive and moreover, it is not sensitive to the order of limits. For example, it applies when the agents' loss from the principal's retaliation is arbitrarily small (i.e. $c$ is small), $\theta_i$ significantly affects agent $i$'s payoff (i.e. $b$ is large) and the mechanical types are arbitrarily rare (i.e. $\delta \to 1$) as long as $L \geq \overline{L}(c, \delta)$. This contrasts to the limiting scenario in the single-agent benchmark where the informativeness of the agent's report converges to infinity and the probability of crime vanishes to 0.

Theorems 1 and 2 suggest that the informativeness of reports decreases and the probability of crime increases when the number of potential victims increases. The comparison becomes stark when the punishment to the convicted is large enough. Next, we argue that these effects are driven by the coordination motives among the agents as well as the negative correlation between their private information, i.e. $\theta_1$ and $\theta_2$, both of which arise endogenously in all monotone-responsive equilibria when $L$ is large enough.

To begin with, when $L$ is large enough, the principal will only be convicted when both agents file reports. This is because if one report is sufficient to convict the principal, then he will have a strict incentive not to commit any crimes which will contradict the conclusion in Lemma 3.1. Therefore, $q(0,0) = q(1,0) = q(0,1) = 0$ and $q(1,1) \in (0,1)$ in all equilibria when $L$ is large enough.

Next, whether principal's choices of $\theta_1$ and $\theta_2$ are strategic complements or substitutes is determined by the sign of:

$$q(1,1) + q(0,0) - q(1,0) - q(0,1). \tag{3.7}$$

As will be formally shown in Lemma D.1, if (3.7) is positive, then $\theta_1$ and $\theta_2$ are strategic substitutes; if (3.7) is negative, then $\theta_1$ and $\theta_2$ are strategic complements. As we have established that (3.7) is strictly positive from the previous step, we know that the principal will have a strict incentive not to abuse agent $j$ once he has already abused agent $i$ and vice versa. This leads to an endogenous negative correlation between $\theta_1$ and $\theta_2$.

From the perspective of an individual agent, he has more incentives to report when he believes that the other agent is more likely to report, as it increases his chances of avoiding the reporting cost $c$. The endogenous negative correlation between $\theta_1$ and $\theta_2$ implies that if $\theta_1 = 0$, then agent 1 believes that agent 2 is abused with probability 0 which decreases his incentives to report; if $\theta_1 = 1$, then agent

1 believes that agent 2 is abused with significant probability which increases his incentives to report. As a result, the coordination motives among agents will undermine the credibility of their reports, which explains why $\mathcal{I}_m < \mathcal{I}_s$.

In the limiting scenario where $L \to +\infty$, two competing effects arise as $\omega_m^*$ and $\omega_m^{**}$ converge to $-\infty$. First, the probability of false positive reports vanishes to 0 as $\omega_m^{**} \to -\infty$, which increases informativeness. This is the only force at work in the single-agent benchmark. Second, the distance between the two cutoffs vanishes to 0 as both of them converge to $-\infty$, which decreases informativeness. Theorem 2 implies that when there are two agents, the second effect dominates even if their losses from miscoordination $c$ is arbitrarily small. This is because as $L$ goes to infinity, the probability of conviction (conditional on two reports) vanishes. This magnifies the cost of miscoordination (which is scaled up by $1-q$) relative to the benefits from reporting (which is scaled up by $q$). The comparison between these magnitudes explains why any strictly positive $c$ suffices.

Nevertheless, given the presence of mechanical types whose reports transmit no information, one may suspect that $\mathcal{I}_m \to 1$ is driven by the scarcity of reports filed by the strategic types rather than their coordination motives. To address this concern, first of all, notice that the informativeness ratio converges to 1 and the probability of crime converges to $\pi^*$ in the limit where $\delta \to 1$ and $L \to \infty$ as long as the relative rate of convergence satisfies $L \geq \overline{L}(c, \delta)$. Next, the comparison between $(\omega_m^*, \omega_m^{**})$ and $(\omega_s^*, \omega_s^{**})$ suggests that the strategic agent's reporting thresholds are *strictly higher* when there are two agents. This implies that for every given report, the probability with which it is filed by a strategic type is strictly higher in the two-agent case compared to the single-agent benchmark. As the agent's report becomes arbitrarily informative in the limit of the single-agent benchmark, the result that the aggregate of the two agents' reports becoming arbitrarily uninformative in the limit cannot be driven by the scarcity of reports from the strategic types. This establishes the causality between the coordination motives among agents and the uninformativeness of reports.

In Appendix K, we show that the above insight is robust against alternative specifications of the mechanical types' strategies. In particular, even when the mechanical types' reports are informative about $\theta$ (for example, the mechanical type can adopt different reporting thresholds when $\theta = 0$ and $\theta = 1$ with the reporting threshold under $\theta = 0$ being strictly higher), it will be overturned by the strategic types' coordination motives when the mechanical types are sufficiently rare (i.e. $\delta \to 1$) and the punishment to the convicted is sufficiently harsh (i.e. $L \to \infty$).

**Remark:** Despite our results are obtained in static environments where reports are submitted simultaneously, the economic forces behind them are applicable to dynamic settings in which reports arrive sequentially. To see this, notice that first, the negative correlation between the agents' private information (namely their $\theta_i$) will arise endogenously whenever an opportunistic principal is aware of the serious consequences of being convicted. Second, an individual agent will have incentives to coordinate with other agents whenever he is unsure about whether his report is pivotal or not. Such concerns can arise when there is a *cold start* (i.e. very few people have reported before as none of them wants to be the first). It can also occur when an agent has observed many reports but he is unsure about the number of reports needed to convict the principal (for example, he faces uncertainty about $\pi^*$). Both of these are realistic concerns under which the coordination inefficiencies pointed out by our results remain valid in more complicated dynamic environments.

### 3.4 Equilibrium Analysis

To better understand Theorems 1 and 2, we perform some analysis of the two-agent model in order to unveil the coordination motives among agents and explain why their reports are becoming arbitrarily uninformative in the $L \to \infty$ limit. We focus on cases where $L > \overline{L}(c, \delta)$ such that all equilibria possess the properties stated in Theorem 1, i.e. symmetric strategies, two reports are required to convict the principal and can be characterized by a quadruple $(\omega_m^*, \omega_m^{**}, q_m, \widetilde{\pi}_m)$.

We start from the agents' incentives. For every $i \in \{1, 2\}$, if $\theta_i = 0$, then agent $i$ has an incentive to report if and only if:

$$\omega \leq \omega_m^* \equiv -c\frac{1 - q_m Q_0}{q_m Q_0} = c - \frac{c}{q_m Q_0}, \tag{3.8}$$

where $Q_0$ is the probability with which agent $j$ ($\neq i$) reports *conditional on* $\theta_i = 0$. Similarly, if $\theta_i = 1$, then agent $i$ has an incentive to report if and only if:

$$\omega \leq \omega_m^{**} \equiv -b - c\frac{1 - q_m Q_1}{q_m Q_1} = -b + c - \frac{c}{q_m Q_1}, \tag{3.9}$$

where $Q_1$ is the probability with which agent $j$ reports *conditional on* $\theta_i = 1$. The expressions for $Q_0$ and $Q_1$ are given by:

$$Q_0 = \delta\Phi(\omega_m^{**}) + (1 - \delta)\alpha \tag{3.10}$$

and

$$Q_1 = \delta\Big(\beta\Phi(\omega_m^{**}) + (1 - \beta)\Phi(\omega_m^*)\Big) + (1 - \delta)\alpha, \tag{3.11}$$

respectively, where

$$\beta \equiv \frac{1-\widetilde{\pi}_m}{1-\widetilde{\pi}_m/2}$$

is the probability that agent $j$ is not abused conditional on agent $i$ is not abused.

The comparisons between $Q_0$ and $Q_1$ as well as $\omega_m^*$ and $\omega_m^{**}$ reveal an important difference between the two-agent case and the single-agent benchmark. Instead of having a constant distance $b$, the distance between $\omega_m^*$ and $\omega_m^{**}$ equals to:

$$b - \frac{c}{q_m} \cdot \frac{-1+Q_1/Q_0}{Q_1}. \tag{3.12}$$

Lemma 3.2 shows that the value of the above expression is strictly less than $b$:

**Lemma 3.2.** *When there are two agents, $Q_1 > Q_0$ and $\omega_m^* - \omega_m^{**} \in (0, b)$.*

*Proof of Lemma 3.2:* According to (3.10) and (3.11), $\omega_m^* - \omega_m^{**} > 0$ is equivalent to $Q_1 > Q_0$. Suppose towards a contradiction that $Q_1 \leq Q_0$, then (3.12) implies that $\omega^* \geq \omega^{**} + b > \omega^{**}$. The comparison between (3.8) and (3.9) then yields $Q_1 > Q_0$, leading to a contradiction. Since $Q_1 > Q_0$, the term $\frac{-1+Q_1/Q_0}{Q_1}$ is strictly positive, and therefore, $\omega_m^* - \omega_m^{**} < b$. □

Intuitively, an agent can avoid the reporting cost $c$ only when the principal is convicted and the latter can only happen when both agents report. Therefore, coordination motives among the agents arise endogenously. Since the decisions to abuse agents are strategic substitutes from the principal's perspective, $\theta_1$ and $\theta_2$ are negatively correlated. Therefore, $Q_1 > Q_0$ and agent $i$'s desire to coordinate with agent $j$ increases his reporting threshold when $\theta_i = 1$ and decreases his reporting threshold when $\theta_i = 0$. This reduces the distance between the two cutoffs making it strictly less than $b$.

Next, we examine how the reduction in $\omega_m^* - \omega_m^{**}$ affects the informativeness of the agents' reports as well as the equilibrium probability of crime. First, we provide an expression for $\mathcal{I}_m$, the measure of reporting informativeness when two reports are required to convict the principal:

$$\mathcal{I}_m \equiv \frac{\Pr(\text{two agents report} \mid \theta_1\theta_2 = 0)}{\Pr(\text{two agents report} \mid \theta_1\theta_2 = 1)}$$

$$= \frac{\Big(\delta\Phi(\omega_m^*) + (1-\delta)\alpha\Big)\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)}{\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)^2} = \frac{\delta\Phi(\omega_m^*) + (1-\delta)\alpha}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha}. \tag{3.13}$$

Since $q_m \in (0,1)$, the evaluator's posterior belief attaching to $\theta_1\theta_2$ equals to $\pi^*$ after observing two

reports, which implies that the equilibrium probability of crime $\widetilde{\pi}_m$ solves:

$$\frac{\widetilde{\pi}_m}{1-\widetilde{\pi}_m} = \frac{l^*}{\mathcal{I}_m}. \tag{3.14}$$

One can then express $\beta$ and $1-\beta$ as functions of $\mathcal{I}_m$, which are given by:

$$\beta = \frac{2\mathcal{I}_m}{l^*+2\mathcal{I}_m} \text{ and } 1-\beta = \frac{l^*}{l^*+2\mathcal{I}_m}. \tag{3.15}$$

Therefore,

$$\frac{Q_1}{Q_0} = \beta + (1-\beta)\mathcal{I}_m = \frac{(l^*+2)\mathcal{I}_m}{l^*+2\mathcal{I}_m}, \tag{3.16}$$

with the RHS being a strictly increasing function of $\mathcal{I}_m$. Applying (3.8) and (3.9) and plug in the expression for $Q_1/Q_0$ in (3.16), we obtain:

$$\frac{|\omega_m^* - c|}{|\omega_m^{**} - c + b|} = \frac{-c/q_m Q_0}{-c/q_m Q_1} = \frac{Q_1}{Q_0} = \frac{(l^*+2)\mathcal{I}_m}{l^*+2\mathcal{I}_m}. \tag{3.17}$$

This ratio between the absolute values of cutoffs leads to the following lemma, which reveals another distinction between the single-agent benchmark and the multi-agent scenario:

**Lemma 3.3.** *When $\omega_m^* \to -\infty$, we have $\mathcal{I}_m \to 1$ and $\widetilde{\pi}_m \to \pi^*$.*

*Proof of Lemma 3.3:* Since $\omega_m^* - \omega_m^{**} \in (0, b)$, the difference between $|\omega_m^* - c|$ and $|\omega_m^{**} - c + b|$ is at most $b$. That is to say, the LHS of (3.17) converges to 1 as $\omega_m^* \to -\infty$. Since the RHS of (3.17) is strictly increasing in $\mathcal{I}_m$, we know that the limiting value of $\mathcal{I}_m$ equals to 1, and according to (3.14), the limiting value of $\widetilde{\pi}_m$ converges to $\pi^*$. □

Finally, we argue that both $\omega_m^*$ and $\omega_m^{**}$ will converge to $-\infty$ as $L \to \infty$. This is driven by the principal's indifference condition:

$$\frac{1}{\delta L} = q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big). \tag{3.18}$$

As $L \to \infty$ and suppose towards a contradiction that $\omega_m^*$ and $\omega_m^{**}$ converge to some bounded number, then either $q_m \to 0$ or $\omega_m^* - \omega_m^{**} \to 0$. If $q_m \to 0$, then (3.8) and (3.9) imply that $\omega_m^*$ also converges to $-\infty$. If $\omega_m^* - \omega_m^{**} \to 0$, namely the distance between the reporting cutoffs converges to 0 while the two cutoffs converge to an interior number, then the agents' coordination motives disappear so the distance between their two reporting cutoffs should be $b$, leading to a contradiction.

# 4 Restore Reporting Credibility

In this section, we explore how to restore the informativeness of reports in order to decrease the probability of crime. We proceed along two directions: (1) mitigating the punishment to the convicted principal and (2) reshaping the agents' incentives via report-contingent monetary transfers.

Our results imply that (1) when punishments are moderate, the principal's decisions to abuse agents are strategic complements and coordination motives among agents will increase the informativeness of their reports; (2) monetary transfers can restore reporting informativeness by compensating an agent when he is the only one that reports; (3) when transfer schemes are required to be budget balanced, reporting informativeness is uniformly bounded from above. This unveils the tension between eliminating coordination inefficiencies and discouraging false positive reports.

## 4.1 Mitigating Punishment to the Convicted

In this subsection, we show that one can restore the informativeness of reports and reduce the probability of crime by mitigating the punishment to the convicted principal. This is because under a moderate $L$, the principal's decisions to abuse the two agents are strategic complements and coordination motives among the agents will increase the informativeness of reports.

Due to the multiplicity problem, computing the optimal $L$ requires to take a strong stance on equilibrium selection among those monotone-responsive equilibria. To circumvent this issue, we show that for every $\delta$, one can find an open interval of $L$ such that the probability of crime vanishes to 0 as $c$ becomes sufficiently large. This is stated as Proposition 4:

**Proposition 4.** *For every $\epsilon > 0$, there exists $\overline{c} > 0$ such that for every $c > \overline{c}$, there exists an open interval $(\underline{L}, \overline{L})$ such that for every $L \in (\underline{L}, \overline{L})$, there exists a symmetric monotone-responsive equilibrium in which the probability of crime is less than $\epsilon$.*

The proof is in Appendix I.1 that constructs a symmetric equilibrium in which the principal's decisions are strategic complements and the convicting probabilities are such that $q(1,1) = 1$, $q(1,0) = q(0,1) \in [1/2, 1)$ and $q(0,0) = 0$.[17] In equilibrium, the principal will either abuse no agent or will abuse both agents at the same time. Consequently, $\theta_1$ and $\theta_2$ will be positively correlated. The coordination motives among the agents will *increase* the distance between the two reporting thresholds, making it strictly larger than $b$. This increases the informativeness of reports and decreases the probability

[17] In order to address the concerns for equilibrium multiplicity, we will show in Appendix I.2 that for any given $(c, \delta)$, there exists an open interval of $L$ under which the value of (3.7) is non-positive in *all* monotone-responsive equilibria.

of crime. In particular, the informativeness of any agent's report will converge to infinity as his cost from miscoordination, $c$, increases.

Proposition 4 and our results in subsection 3.3 imply that in order to minimize the probability of crime, the optimal $L$ is interior when there are multiple potential victims, as opposed to the single-agent benchmark in which the probability of crime decreases as $L$ increases. This provides a novel rationale for mitigating the punishment to the convicted compared to the alternative explanations in the law and economics literature. Our logic applies to settings where objective evidence is scarce and the potential victims' claims are hard to verify: due to the lack of evidence, the potential victims wish to coordinate their reports in order to improve the chances of conviction. Moderating the punishment to the convicted enhances reporting credibility by endogenously generating positive correlations between the potential victims' private information. Under this positive correlation, an agent's coordination motive will discourage him from reporting when he has not witnessed a crime and will encourage him to report when he has witnessed a crime.

### 4.2 Monetary Transfers

In this subsection, we maintain the assumption that $L$ is large and explore the possibility of mitigating coordination inefficiencies via monetary transfers that can be contingent on the vector of reports. Let $t_i(\mathbf{a}) \in \mathbb{R}$ be the transfer to agent $i$ under reporting profile $\mathbf{a}$. A transfer scheme achieves *budget balance* if and only if $\sum_{i=1}^{2} t_i(\mathbf{a}) = 0$ for all $\mathbf{a} \in \{0,1\}^2$.

We focus on equilibria where $q(0,0) = q(1,0) = q(0,1) = 0$ under a given transfer scheme $t : \{0,1\}^2 \to \mathbb{R}^2$, which is without loss of generality when $L$ is large enough. Since transfer schemes can be asymmetric across agents, it is no longer without loss to focus on symmetric equilibria. To address this issue, we start from analyzing the principal's incentives in order to provide the right formula for the informativeness of reports by taking these potential asymmetries into account.

To start with, the informativeness of reports is measured by:

$$\mathcal{I}_m \equiv \frac{\Pr(a_1 = a_2 = 1 |\ \theta_1\theta_2 = 0)}{\Pr(a_1 = a_2 = 1 |\ \theta_1\theta_2 = 1)},$$

as two reports are required to convict the principal. For every $i \in \{1,2\}$, let $\Psi_i^*$ be the probability with which agent $i$ reports conditional on $\theta_i = 0$ and let $\Psi_i^{**}$ be the probability with which he reports

conditional on $\theta_i = 1$. Let $\mathcal{I}_i \equiv \Psi_i^*/\Psi_i^{**}$. One can write $\mathcal{I}_m$ as a convex combination of $\mathcal{I}_1$ and $\mathcal{I}_2$:

$$\mathcal{I}_m \equiv \frac{p_1}{p_1+p_2}\mathcal{I}_1 + \frac{p_2}{p_1+p_2}\mathcal{I}_2 \tag{4.1}$$

where $p_i$ is the probability with which agent $i$ is mistreated. Our first observation is that the overall informativeness of reports equals to the minimal informativeness of the two individual reports:

**Lemma 4.1.** *For any given transfer scheme, in every monotone-responsive equilibrium where* $q(0,0) = q(1,0) = q(0,1) = 0$*, we have* $\mathcal{I}_m = \min\{\mathcal{I}_1, \mathcal{I}_2\}$.

Intuitively, this is because the principal's marginal cost of abusing an agent is strictly lower when the latter's report is strictly less informative. As a result, when $\mathcal{I}_1 < \mathcal{I}_2$, the principal will abuse the agent 2 with zero probability and in equilibrium, all the variations in $\theta_1\theta_2$ are driven by the variations in $\theta_1$ instead of the variations in $\theta_2$. As a result, $\mathcal{I}_2$ carries zero weight in the formula of $\mathcal{I}_m$.

*Proof of Lemma 4.1:* The ratio between the principal's cost of mistreating agent 1 and that of mistreating agent 2 equals to:

$$\frac{\Psi_2^{**}(\Psi_1^* - \Psi_1^{**})}{\Psi_1^{**}(\Psi_2^* - \Psi_2^{**})}$$

which is strictly greater than 1 if and only if $\mathcal{I}_1 > \mathcal{I}_2$. In this case, $p_1 = 0$ and according to (4.1), we have $\mathcal{I}_m = \mathcal{I}_2$. Similarly, if $\mathcal{I}_1 < \mathcal{I}_2$, then $p_2 = 0$ and $\mathcal{I}_m = \mathcal{I}_1$. If $\mathcal{I}_1 = \mathcal{I}_2$, since $\mathcal{I}_m$ is a convex combination of $\mathcal{I}_1$ and $\mathcal{I}_2$, we have $\mathcal{I}_m = \mathcal{I}_1 = \mathcal{I}_2$ for all values of $p_1$ and $p_2$. □

Next, we construct a transfer scheme that eliminates the coordination inefficiencies, in the sense that the aggregate informativeness of reports converge to infinity and the probability of crime vanishes to 0 in the $L \to \infty$ limit.

**Proposition 5.** *For every* $c > 0$ *and* $\epsilon > 0$*, there exists* $\overline{L} > 0$ *such that when* $L > \overline{L}$ *and under the following transfer scheme:*

$$t_1(a_1,a_2) = \begin{cases} c & \text{if } (a_1,a_2) = (1,0) \\ 0 & \text{otherwise} \end{cases} \qquad t_2(a_1,a_2) = \begin{cases} c & \text{if } (a_1,a_2) = (0,1) \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{I}_m$ *exceeds* $1/\epsilon$ *and the probability of crime is less than* $\epsilon$ *in all monotone-responsive equilibria.*

The proof is in Appendix G.1. According to Proposition 5, a designer can overcome the coordination inefficiencies and restore reporting informativeness by compensating an agent when he is the

only one that reports. The amount to be transferred exactly offsets an agent's cost of reporting. This transfer scheme eliminates the agents' incentives to coordinate, as in every equilibrium, the distance between each agent's reporting cutoffs is exactly $b$. However, this is not to say that under this transfer scheme, the equilibrium outcome coincides with that of the single-agent benchmark. Due to the difference in the principal's incentive constraint, the cutoffs are strictly higher compared to the single-agent benchmark. As a result, the informativeness of reports is strictly lower and the probability of crime is strictly higher under a fixed $L$. Nevertheless, as stated in Proposition 4, the informativeness of reports and the equilibrium probability of crime coincide with those in the single-agent benchmark in the $L \to \infty$ limit.

Despite the above transfer scheme can effectively reduce the probability of crime when $L$ is large, two potential drawbacks emerge. First, the designer needs to incur a budget deficit with positive probability. This deficit is large when the agent's loss from the principal's retaliation is large. Second, it encourages collusion between the principal and the two agents, in the sense that the principal can commit not to retaliate when only one agent reports. After the reporting agent obtains the designer's transfer $c$, he can then share it with the other agent and the principal. Motivated by these observations, we explore the possibility of restoring reporting informativeness via budget balanced transfer schemes, which addresses the above concerns. However, it turns out that the informativeness ratio is uniformly bounded from above, which reveals the tension between improving the informativeness of reports, eliminating the budget deficit and deterring collusion.

**Proposition 6.** *There exist $\overline{\mathcal{I}} > 1$ and $\underline{\pi} \in (0, \pi^*)$ such that for all monotone-responsive equilibria under all budget balanced transfer schemes for all $L > \overline{L}(\delta, c)$, the informativeness ratio $\mathcal{I}_m$ is less than $\overline{\mathcal{I}}$ and the equilibrium probability of crime $\widetilde{\pi}_m$ is greater than $\underline{\pi}$.*

The proof is in Appendix G.2. Intuitively, this is driven by the tension between offsetting the harmful coordination motives, which can only be achieved via increasing $\Delta_1 \equiv t_1(1,0) - t_1(0,0)$ and $\Delta_2 \equiv t_2(0,1) - t_2(0,0)$, while at the same time, deterring false positives, i.e. decreasing $\omega_1^{**}$ and $\omega_2^{**}$. The budget balance requirement forbids the principal from achieving two goals at the same time, leading to bounded informativeness and non-vanishing probability of crime.

The key challenge to prove this result arises from flexibility to choose transfers and therefore, the symmetry part of Theorem 1 no longer applies. As a result, the previous argument to bound informativeness, based on the ratio condition (3.17), is no longer applicable. This is because in a general asymmetric equilibrium, if an agent is abused with probability close to 0, the RHS of (3.17)

approaches 1 even when $\mathcal{I}_m$ is large. Moreover, the denominator and numerator of the LHS of (3.17) also need to be adjusted according to the promised monetary transfers. As will become clear in the proof, whether the denominator and numerator need to plus or minus a common term depends on the sign of

$$X \equiv \frac{1}{q}\Big(t_1(1,0) + t_1(0,1) - t_1(1,1) - t_1(0,0)\Big). \tag{4.2}$$

When the absolute value of $X$ is large enough, two complications arise. First,

$$\frac{\omega_i^* - c - |X|}{\omega_i^{**} - c + b - |X|} \to 1$$

does not imply that $\mathcal{I}_m \to 1$ since the probability that agent $i$ being abused approaches 0. Second,

$$\frac{\omega_j^* - c + |X|}{\omega_j^{**} - c + b + |X|}$$

may not converge to 1 when $\omega_j^* \to -\infty$ due to the effect of a large $|X|$.

To address these issues, we use the observation that informativeness can be unbounded only when both $\omega_1^{**}$ and $\omega_2^{**}$ are arbitrarily small and moreover, $X$ needs to be arbitrarily large. To derive a contradiction, we exploit the comparisons between the two agents' reporting cutoffs in asymmetric equilibria to derive uniform upper bounds on $|\omega_1^{**}|$, $|\omega_2^{**}|$ or $|X|$ for all large enough $\mathcal{I}_m$. The existence of a uniform upper bound on either $|\omega_1^{**}|$, $|\omega_2^{**}|$ or $|X|$ will contradict the previous assumption that $\mathcal{I}_m$ can grow without bound, which implies that the informativeness of reports is uniformly bounded from above under every budget balanced transfer scheme.

# 5 Extensions

In this section, we examine the robustness of our main insights under two variations of the baseline model. In subsection 5.1, we analyze games with more than two agents. In subsection 5.2, we examine the game's equilibrium outcome when the principal's payoff is unknown to the agents and the evaluator.

## 5.1 More Than Two Agents

We generalize our findings in section 3 to environments with more than two agents and perform comparative statics on the number of agents. We focus on symmetric equilibria in which $q(\mathbf{a}) > 0$ if and only if $\mathbf{a} = (1, 1, ..., 1)$, which we call *unanimous equilibria*. Using similar arguments as those in

the proof of Theorem 1, one can show that unanimous equilibria exist and focusing on them is without loss of generality when $L$ is large enough.

Abusing notation, we use subscript $n \in \mathbb{N}$ to denote the value of variables in an environment with $n$ agents. For every $i \in \{1, 2, ..., n\}$, agent $i$'s reporting cutoff is

$$\omega_n^* = c - \frac{c}{q_n Q_{0,n}} \tag{5.1}$$

when $\theta_i = 0$, and is

$$\omega_n^{**} = -b + c - \frac{c}{q_n Q_{1,n}}, \tag{5.2}$$

when $\theta_i = 1$, where

$$Q_{0,n} \equiv \Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1} \tag{5.3}$$

and

$$Q_{1,n} \equiv \frac{n\mathcal{I}_n}{(n-1)l^* + n\mathcal{I}_n}\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}$$
$$+ \frac{(n-1)l^*}{(n-1)l^* + n\mathcal{I}_n}\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-2}\Big(\delta\Phi(\omega_n^{*}) + (1-\delta)\alpha\Big). \tag{5.4}$$

Since $n$ reports are required to convict the principal, the aggregate informativeness of agents' reports is given by the ratio between the probability with which $n$ agents report conditional on $\Pi_{i=1}^n \theta_i = 0$ and the probability with which $n$ agents report conditional on $\Pi_{i=1}^n \theta_i = 1$. This is denoted by $\mathcal{I}_n$ and equals to:

$$\mathcal{I}_n \equiv \frac{\Pr(\text{there are } n \text{ reports} \mid \Pi_{i=1}^n \theta_i = 0)}{\Pr(\text{there are } n \text{ reports} \mid \Pi_{i=1}^n \theta_i = 1)} = \frac{\delta\Phi(\omega_n^*) + (1-\delta)\alpha}{\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha}.$$

As in the two-agent economy, there exists a one-to-one mapping between the informativeness ratio $\mathcal{I}_n$ and the equilibrium probability of crime $\widetilde{\pi}_n$, which is given by:

$$\mathcal{I}_n = \frac{\pi^*}{1-\pi^*} \Big/ \frac{\widetilde{\pi}_n}{1-\widetilde{\pi}_n}. \tag{5.5}$$

In every equilibrium where $L$ is large enough, the principal is indifferent between abusing one agent and abusing no agent, which gives the following indifference condition:

$$\frac{1}{\delta L} = \delta q_n \Big(\Phi(\omega_n^*) - \Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha\Big)^{n-1}. \tag{5.6}$$

We have the following result that generalizes Theorems 1 and 2 to settings with $n$ ($\geq 2$) agents:

**Proposition 7.** *There exists $\overline{L}_n : \mathbb{R}_+ \times (0,1) \in \mathbb{R}_+$ such that for every $(L, c, \delta)$ with $L > \overline{L}_n(c, \delta)$, there exists an unanimous equilibrium satisfying (5.1), (5.2), (5.5) and (5.6).*
*Moreover, for every $\epsilon > 0$, there exists $\overline{L}_{n,\epsilon}(c, \delta) \geq \overline{L}_n(c, \delta)$ such that in every unanimous equilibrium when $L > \overline{L}_{n,\epsilon}(c, \delta)$, we have:*

$$\omega_n^*, \omega_n^{**} < -1/\epsilon, \quad \mathcal{I}_n < 1 + \epsilon \quad \textit{and} \quad \widetilde{\pi}_n \geq \pi^* - \epsilon.$$

According to Proposition 7, the main insights from the two-agent scenario extend to settings with more than two agents. In particular, coordination motives among agents and negative correlations between their private informative arise endogenously, which will undermine the credibility of their reports and increase the probability of crime. The proof of equilibrium existence uses a similar argument as that of Proposition 3, which we will briefly discuss in Appendix B.2. In what follows, we will show that $\mathcal{I}_n \to 1$ and $\widetilde{\pi}_n \to \pi^*$ once $L \to \infty$, which is similar to what we have seen in subsection 3.4.

*Proof of Proposition 7:* First, we show that $\omega_n^* - \omega_n^{**} \in (0, b)$. Suppose towards a contradiction that $\omega_n^* - \omega_n^{**} \leq 0$, then the comparison between (5.3) and (5.4) suggests that $Q_{0,n} \geq Q_{1,n}$. Plugging this into (5.1) and (5.2), it implies that $\omega_m^* \geq \omega_m^{**} + b$. On the other hand, since $\omega_n^* - \omega_n^{**} > 0$, we know that $Q_{0,n} < Q_{1,n}$. The expressions for the cutoffs the imply that $\omega_n^* - \omega_n^{**} < b$, leading to a contradiction.

Next, we show that $\mathcal{I}_n \to 1$ as $\omega_n^* \to -\infty$. To see this, apply the expressions of $\omega_n^*$ and $\omega_n^{**}$ in (5.1) and (5.2), we have:

$$\frac{|\omega_n^* - c|}{|\omega_n^{**} + b - c|} = \frac{Q_{1,n}}{Q_{0,n}} = \frac{(n-1)l^*}{(n-1)l^* + n\mathcal{I}_n}\mathcal{I}_n + \frac{n\mathcal{I}_n}{(n-1)l^* + n\mathcal{I}_n}. \tag{5.7}$$

Since $\omega_n^* - \omega_n^{**} \in (0, b)$, the LHS converges to 1 as $\omega_n^* \to -\infty$, which implies that the RHS also converges to 1. This can only be the case when $\mathcal{I}_n \to 1$.

In the last step, we show that $\omega_n^* \to -\infty$ as $L \to \infty$. Suppose towards a contradiction that there exists an interior accumulation point $\omega^* \in \mathbb{R}_-$, then as the LHS of (5.6) converges to 0 when $L \to \infty$, we know that either $q_n \to 0$ or $\Phi(\omega_n^*) - \Phi(\omega_n^{**}) \to 0$ or both. The latter implies that $\omega_n^* - \omega_n^{**} \to 0$ as $\omega_n^* \to \omega^*$ and $\omega^*$ is interior.

Suppose $q_n \to 0$, then $\omega_n^* \to -\infty$ according to (5.1). Next, suppose towards a contradiction that $q_n$ is bounded away from 0 along some subsequence, i.e strictly greater than some $\underline{q} > 0$, then we know

that $\omega_n^* - \omega_n^{**} \to 0$. Subtracting the expression of $\omega_n^*$ from that of $\omega_n^{**}$, we obtain:

$$\frac{q_n}{c}\Big(\omega_n^* - (\omega_n^{**} + b)\Big) = \frac{(n-1)l^*}{(n-1)l^* + n}\Big(\frac{1}{\delta\Phi(\omega_n^*) + (1-\delta)\alpha} - \frac{1}{\delta\Phi(\omega_n^{**}) + (1-\delta)\alpha}\Big). \tag{5.8}$$

The absolute value of the LHS is no less than $\underline{q}b/c$ in the limit as $\omega_n^* - \omega_n^{**} \to 0$. The absolute value of the RHS converges to 0 as $\Phi(\omega_n^*) - \Phi(\omega_n^{**}) \to 0$, leading to a contradiction. This suggests that $\omega_n^* \to -\infty$ in every equilibrium as $L \to \infty$.

The three parts together imply that as $L \to \infty$, $\omega_n^*$ and $\omega_n^{**}$ go to $-\infty$, the aggregate informativeness of reports, $\mathcal{I}_n$, will converge to 1 and the equilibrium probability of crime $\widetilde{\pi}_n$ will converge to $\pi^*$. □

Next, we state a result that establishes some comparative statics on the number of agents.

**Proposition 8.** *For every $k, n \in \mathbb{N}$ with $k > n$, when $L$ is large enough such that unanimous equilibria exist under $k$ and $n$, then compare any unanimous equilibrium under $k$ to any unanimous equilibrium under $n$, we have: $\omega_k^* > \omega_n^*$, $\omega_k^{**} > \omega_n^{**}$, $\omega_k^* - \omega_k^{**} < \omega_n^* - \omega_n^{**}$, $\widetilde{\pi}_k > \widetilde{\pi}_n$ and $\mathcal{I}_n > \mathcal{I}_k$.*

The proof is in Appendix H. According to Proposition 8, as the number of agents increases, each individual agent is more likely to report no matter whether he has been abused or not, the distance between the reporting cutoffs decrease and the aggregate informativeness of their reports also decreases. As a result, the equilibrium probability of crime increases. The driving force behind such comparative statics is still the interaction between the coordination motives among agents and the negative correlation between their private information. Such effects are more pronounced when information is more dispersed among the agents.

## 5.2 Uncertainty about the Principal's Payoff

In this subsection, we examine the robustness of our insights when there is uncertainty about the principal's payoff. The primary motivation is to accommodate the presence of *saints* (i.e. morally upright people who hates committing crimes) and *serial assaulters*. This is captured by the heterogeneity in the principal's $L$, which measures his loss from conviction relative to his benefit from committing crimes.

To formalize these ideas, consider the following variation of the baseline model in which the principal's preference is his private information. This is modeled as a random variable $\widetilde{L} \in [0, +\infty]$, which replaces $L$ in the baseline model. For illustration purposes, we focus on environments where there are two agents and $\widetilde{L}$ can take two possible values, $L_l$ and $L_h$, with $0 \leq L_l < L_h \leq +\infty$. We consider

two cases separately, which together explain the robustness of our results to the presence of saints and serial assaulters, as long as both of these types occur with low enough probability.

**Saints:** Suppose $L_l$ is larger than the lower bound on $L$ required by Theorem 1 and the probability with which $L = L_h$ is strictly less than $1 - \pi^*$. Monotone-responsive equilibria in this environment take a similar form as those in Theorem 1, with the only difference being, type $L_h$ never commits any crime and type $L_l$ commits a crime with probability $\widetilde{\pi}/\Pr(\widetilde{L} = L_l)$, where $\widetilde{\pi}$ is the unconditional probability of crime. Intuitively, this is because type $L_l$ is indifferent between committing and not committing a crime. Given type $L_l$'s indifference, type $L_h$ will have a strict incentive not to commit any crimes according to the standard supermodularity argument. The same insight extends when we replace the presence of type $L_h$ with types that receive zero or even negative benefit from committing crimes.

**Serial Assaulters:** Suppose next that $L_h$ is larger than the lower bound on $L$ required by Theorem 1 while $L_l$ is strictly less than 1.[18] We assume that $\widetilde{L} = L_l$ occurs with small but strictly positive probability, denoted by $\epsilon$. In this environment, type $L_l$ principal is interpreted as the *bad apples* who have very high propensity to commit crimes and will be the serial assaulters in the equilibrium of our model. The following result demonstrates the robustness of the endogenous negative correlation between $\theta_1$ and $\theta_2$, which is the key driving force behind the coordination inefficiencies identified by our results:

**Proposition 9.** *For every $\epsilon > 0$ small enough, there exists $\overline{R} > 1$ such that for every $R \in (1, \overline{R})$, there exist $\overline{\delta} \in (0,1)$, $L_h : [\overline{\delta}, 1) \to \mathbb{R}_+$ and an equilibrium $\{\omega^*, \omega^{**}, \widetilde{\pi}, q\}$ under each $(\epsilon, \delta, L_h(\delta), L_l)$ such that:*

1. *The probability of crime is $\widetilde{\pi}$, the reporting thresholds are $\omega^*$ and $\omega^{**}$ and the principal is convicted with positive probability if and only if there are two reports.*

2. *The conviction probability following two reports is $q$.*

3. *Type $L_l$ assaults both agents with probability 1. Type $L_h$ assaults agent $i$ with probability*

$$\frac{\widetilde{\pi} - \epsilon}{2(1 - \epsilon)}$$

[18] Assuming $L_l < 1$ is a simplification that facilitates the exposition. For our result to hold, we only need $L_l$ to be relatively small compared to $L_h$ such that type $L_l$ has an incentive to commit two crimes.

*for every $i \in \{1,2\}$ and assaults no agent with probability $(\widetilde{\pi}-\epsilon)/(1-\epsilon)$.*

4. *The likelihood ratio*

$$\frac{\delta\Phi(\omega^*)+(1-\delta)\alpha}{\delta\Phi(\omega^{**})+(1-\delta)\alpha} \tag{5.9}$$

*equals to $R$, the informativeness of report is given by*

$$\mathcal{I} \equiv \frac{\epsilon}{\widetilde{\pi}}R^2 + \Big(1-\frac{\epsilon}{\widetilde{\pi}}\Big)R, \tag{5.10}$$

*and the equilibrium probability of crime $\widetilde{\pi}$ satisfies:*

$$\widetilde{\pi}/(1-\widetilde{\pi}) = l^*/\mathcal{I}. \tag{5.11}$$

According to Proposition 9, for every $R$ close enough to 1 and $\epsilon$ close enough to 0, there exists $L_h$ such that the likelihood ratio of a monotone-responsive equilibrium under $(\epsilon, L_h)$ equals to $R$. As indicated by (5.10) and (5.11), the aggregate informativeness of reports $\mathcal{I}$ is also close to 1 and the equilibrium probability of crime is close to $\pi^*$.

Compared to Theorems 1 and 2 that address the common properties of *all equilibria*, Proposition 9 only demonstrates the existence of *an equilibrium* in which $\theta_1$ and $\theta_2$ are negatively correlated, the informativeness of reports is close to 1 and the probability of crime is close to $\pi^*$. This is because the presence of serial assaulters leads to a larger variety of self-fulfilling beliefs with different qualitative features, even when those types occur with arbitrarily small probability.

To understand the intuition, let us start from an agent's equilibrium strategy, which are summarized by two cutoffs $(\omega^*, \omega^{**})$, in an environment without serial assaulters, i.e. when $\epsilon = 0$. Once $\epsilon$ becomes strictly positive, it undermines the negative correlation between $\theta_1$ and $\theta_2$, which encourages agent $i$ to report when $\theta_i = 0$ and discourages him from reporting when $\theta_i = 1$. The increase in the distance between $\omega^*$ and $\omega^{**}$ will then increase the informativeness of the agent's report and in equilibrium, will decrease the net probability of crime. Since the probability of the principal being a serial assaulter is fixed to be $\epsilon$, a decrease in the total probability of crime increases

$$\Pr(\theta_1 = \theta_2 = 0 | \theta_1\theta_2 = 0). \tag{5.12}$$

This further weakens the negative correlation between $\theta_1$ and $\theta_2$, which will in turn increase the informativeness of report and decrease the probability of crime. Therefore, one can iterate the above

argument until it reaches a new fixed point. The probability of crime could be close to $\epsilon$, or close to $\pi^*$, or somewhere in between depending on the starting point.

# 6 Conclusion

We analyze the interactions between a potential assaulter's incentives to commit crimes and the potential victims' incentives to report crimes. In our model, the conviction process, the credibility of reports and the probability of crime are all endogenous. We find that when the punishment to the convicted is sufficiently large, the assaulter's strategic restraint induces negative correlation between the potential victims' private information. This together with their endogenous incentives to coordinate reduce the credibility of reports and increase the probability of crime.

On the other hand, one can restore the credibility of reports and reduce the probability of crime by rewarding an agent when he is the only one that reports, or by mitigating the punishment to the convicted principal. The first solution offsets agents' coordination motives. The second solution induces positive correlation between the agents' private information and as a result, their incentives to coordinate will enhance their reporting credibility.

We conclude with several remarks on the applicability and robustness of our main results. First, our results are derived under an *equilibrium analysis*. That being said, they address people's behaviors when they understand the rules of the game, the payoff consequences of their actions and know how to play their equilibrium strategies. This is well-suited for settings with stable institutions, laws and norms such that play has converged to an equilibrium and the potential assaulters, victims and evaluators (e.g. judges, public opinions, headquarters of firms) are likely to follow their equilibrium strategies.[19] On the other hand, our results are less relevant for situations where there have been recent changes in the environment, such as a sudden crack down of crimes, the introduction of new laws and regulations, drastic shifts of norms and perceptions, etc.

Second, the insights from our baseline model are robust against a variety of alternative specifications on players' payoffs and information structures. Aside from the extensions discussed in section 5, all our results remain robust when the principal's marginal benefit from committing crimes is decreasing in the number of crimes he has already committed, or the principal receives noisy private information about the agents' $\omega$s. They are also robust when there exists ex post evidence against false positive claims. For example, if agent $i$ files a false positive report (i.e. choosing $a_i = 1$ when

[19] See Fudenberg and Levine (1995) for theories on how equilibrium emerges in the long-run when players adopt heuristic learning processes, such as smooth fictitious play, etc.

$\theta_i = 1$) that gets the principal convicted, then with probability $p^*$, some ex post evidence will arrive that can reveal the principal's innocence. In this case, every agent who files a false positive claim will receive a penalty $k > 0$. Our analysis still goes through and all our qualitative results remain to be true as the presence of ex post evidence is equivalent to an increase in $b$.[20]

[20] To see this, agent $i$'s indifference condition when $\theta_i = 1$ is now given by:

$$q_m Q_1(\omega_i + b) = -c(1 - q_m Q_1) - q_m Q_1 p^* k.$$

The expression for the cutoff is then given by

$$\omega_m^{**} \equiv -b - p^* k - c\frac{1 - q_m Q_1}{q_m Q_1} = -b - p^* k + c - \frac{c}{q_m Q_1}.$$

The above expression is qualitatively the same as (3.9) except one needs to replace $b$ with $b + p^* k$.

# A Proof of Proposition 2: Comparative Statics

Our proof will repeatedly use the observation that when $\omega_s^*$ and $\omega_s^{**}$ decrease while keeping the distance between them constant, $\Phi(\omega_s^*) - \Phi(\omega_s^{**})$ decreases and $\Phi(\omega_s^*)/\Phi(\omega_s^{**})$ increases.

**Proof of Statements 1 and 3:** When $c$ increases, according to (3.1) and (3.2), both $\omega_s^*$ and $\omega_s^{**}$ will decrease for fixed $q_s$. Therefore, $q_s$ increases in equilibrium so that the LHS of (3.5) equals to the RHS. Since the RHS of (3.5) remains unchanged and the LHS is increasing in $q_s$, we know that both $\omega_s^*$ and $\omega_s^{**}$ will decrease. Therefore, $\mathcal{I}_s$ increases and $\widetilde{\pi}_s$ decreases.

When $c \to \infty$ while holding $L$ constant, if $q_s$ converges to any number strictly below 1, then according to (3.1) and (3.2), both $\omega_s^*$ and $\omega_s^{**}$ will converge to $-\infty$ and the LHS of (3.5) will converge to 0, leading to a contradiction. Therefore, $q_s \to 1$ and the limit value of $\omega_s^*$ and $\omega_s^{**}$ are given in statement 3.

**Proof of Statements 2 and 4:** When $L$ increases, the RHS of (3.5) decreases. As a result, either $q_s$ decreases or $\omega_s^*$ and $\omega_s^{**}$ decrease or both. As $\omega_s^*$ is strictly increasing in $q_s$, we know that $q_s, \omega_s^*$ and $\omega_s^{**}$ will decrease. As a result, $\mathcal{I}_s$ increases and $\widetilde{\pi}_s$ decreases.

When $L \to \infty$ while holding $c$ constant, the RHS of (3.5) converges to 0, and as a result, $q_s \to 0$ and $\omega_s^*, \omega_s^{**} \to -\infty$. In the limiting economy where $\delta \to 1$, we have:

$$\lim_{\omega_s^* \to -\infty} \lim_{\delta \to 1} \frac{\delta\Phi(\omega_s^*) + (1-\delta)\alpha}{\delta\Phi(\omega_s^* - b) + (1-\delta)\alpha} = \infty.$$

As a result, the probability of crime $\widetilde{\pi}_s$ will vanish to 0.

# B Existence of Equilibrium

## B.1 Proof of Proposition 3

We show that when $L$ is large enough, there exists a symmetric equilibrium in which (i) $q(\mathbf{a}) > 0$ if and only if $\mathbf{a} = (1,1)$; (ii) the principal either abuses no agent or abuses one agent. The main step of the proof is to establish the following proposition:

**Proposition B.** *For every $(c, \delta) \in \mathbb{R}_+ \times (0,1)$, there exists $\overline{L} > 0$ such that for every $L > \overline{L}$, there*

*exists a triple* $(\omega_m^*, \omega_m^{**}, q_m) \in \mathbb{R}_- \times \mathbb{R}_- \times (0,1)$ *that solves the following three equations:*

$$\frac{q_m}{c}(\omega_m^* - c) = -\frac{1}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha} \tag{B.1}$$

$$\frac{q_m}{c}(\omega_m^{**} - c + b) = -\frac{l^*}{l^*+2} \cdot \frac{1}{\delta\Phi(\omega_m^*) + (1-\delta)\alpha} - \frac{2}{l^*+2} \cdot \frac{1}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha} \tag{B.2}$$

$$\frac{1}{\delta L} = q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big). \tag{B.3}$$

*Proof of Proposition B:.* The proof consists of two steps. In **Step 1**, we show that once fixing $q_m$ to be 1, the value of the following expression:

$$A \equiv \inf_{(\omega_m^*, \omega_m^{**}) \text{ that solves (B.1) and (B.2) when } q_m = 1} \delta\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big)\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big) \tag{B.4}$$

is strictly bounded away from 0. We establish this bound by putting lower bounds on $\Phi(\omega_m^{**})$ and $\Phi(\omega_m^*) - \Phi(\omega_m^{**})$, respectively. To see this, first,

$$\omega_m^{**} \geq -b + c - \frac{c}{(1-\delta)\alpha}$$

and therefore,

$$\Phi(\omega_m^{**}) \geq \Phi\Big(-b + c - \frac{c}{(1-\delta)\alpha}\Big). \tag{B.5}$$

Next, let $\Delta \equiv \omega_m^* - \omega_m^{**}$, which has to be strictly between 0 and $b$. Deducting equation (B.2) from (B.1) and plugging in $q_m = 1$, we have:

$$\frac{b-\Delta}{c} = \frac{\delta l^*}{l^*+2}\Big(\delta\Phi(\omega_m^*) + (1-\delta)\alpha\Big)^{-1}\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)^{-1}\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big). \tag{B.6}$$

Consider two cases:

1. When $\Delta \geq b/2$, then

$$\delta\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big) \geq \frac{b\delta}{2}\phi(\omega_m^{**}) \geq \frac{b\delta}{2}\phi\Big(-b + c - \frac{c}{(1-\delta)\alpha}\Big), \tag{B.7}$$

which uses the assumption that the density of $\omega$ is increasing when $\omega < 0$.

2. When $\Delta < b/2$, then (B.6) implies that:

$$\delta\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big) \geq \frac{b(l^*+2)}{2l^*c}\Big(\delta\Phi(\omega_m^*) + (1-\delta)\alpha\Big)\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big).$$

$$\geq \frac{b(l^*+2)}{2l^*c}\Big(\delta\Phi\big(-b+c-\frac{c}{(1-\delta)\alpha}\big) + (1-\delta)\alpha\Big)^2. \tag{B.8}$$

Taking the minimum of the right-hand sides of (B.7) and (B.8), we obtain a lower bound for $\delta\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big)$. This together with (B.5) implies a lower bound for (B.4), which is strictly bounded above 0.

We use $\underline{A}$ to denote the lower bound we obtained in Step 1. In **Step 2**, we show that when $L > \underline{A}^{-1}$, there exists a solution to (B.1), (B.2) and (B.3) using a fixed point argument. For every $(\Phi^*, \Phi^{**}, q) \in [0,1]^2 \times [1/L, 1]$, let $f \equiv (f_1, f_2, f_3) : [0,1]^2 \times [1/L, 1] \to [0,1]^2 \times [1/L, 1]$ be the following mapping:

$$f_1(\Phi^*, \Phi^{**}, q) = \Phi\Big(c - \frac{c}{q(\delta\Phi^{**} + (1-\delta)\alpha)}\Big), \tag{B.9}$$

$$f_2(\Phi^*, \Phi^{**}, q) = \Phi\Big(-b + c - \frac{cl^*}{q(l^*+2)}\frac{1}{\delta\Phi^* + (1-\delta)\alpha} - \frac{2c}{q(l^*+2)}\frac{1}{\delta\Phi^{**} + (1-\delta)\alpha}\Big), \tag{B.10}$$

$$f_3(\Phi^*, \Phi^{**}, q) = \min\Big\{1, \frac{1}{\delta L}\frac{1}{\Big(\delta\Phi^{**} + (1-\delta)\alpha\Big)\Big(\Phi^* - \Phi^{**}\Big)}\Big\}. \tag{B.11}$$

Since $f$ is continuous, the Brouwer's fixed point theorem implies the existence of a fixed point.

Next, we show that if $(\Phi^*, \Phi^{**}, q)$ is a fixed point, then $q < 1$. This will imply that every solution to the fixed point problem solves the system of equations (B.1), (B.2) and (B.3) as (B.11) and (B.3) are the same when $q < 1$. Suppose towards a contradiction that $q = 1$, then $\Phi^{-1}(\Phi^*)$ and $\Phi^{-1}(\Phi^{**})$ is a solution to (B.1) and (B.2) once we fix $q$ to be 1. According to Part I of the proof, the assumption that $L > \underline{A}^{-1}$ implies that

$$\frac{1}{\delta L}\frac{1}{\Big(\delta\Phi^{**} + (1-\delta)\alpha\Big)\Big(\Phi^* - \Phi^{**}\Big)} < 1.$$

Therefore the RHS of (B.11) is strictly less than 1. This contradicts the claim that $(\Phi^*, \Phi^{**}, 1)$ is a fixed point of $f$, which implies that the value of $q$ at the fixed point is strictly less than 1. □

Given the triple $(\omega_m^*, \omega_m^{**}, q_m) \in \mathbb{R}_- \times \mathbb{R}_- \times (0,1)$, one can then uniquely pin down $\widetilde{\pi}_m \in (0,1)$ via:

$$\frac{\delta\Phi(\omega_m^*) + (1-\delta)\alpha}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha} = l^* \Big/ \frac{\widetilde{\pi}_m}{1-\widetilde{\pi}_m}. \tag{B.12}$$

According to the analysis in subsection 3.4, equations (B.1), (B.2), (B.3) and (B.12) are sufficient con-

ditions for a monotone-responsive equilibrium $\{\omega_m^*, \omega_m^{**}, q_m, \widetilde{\pi}_m\}$ satisfying $q(1,1) = q_m$ and $q(0,0) = q(1,0) = q(0,1) = 0$. The existence of monotone-responsive equilibrium in the two-agent case immediately follows.

### B.2 Generalizations

We generalize our existence proof by allowing for more than two agents and alternative specifications of the mechanical types' reporting strategies. Assume that when agent $i$ is mechanical, he will report with probability $p_0$ when $\theta_i = 0$ and with probability $p_1$ when $\theta_i = 1$, with $p_0, p_1 > 0$.[21] We show that for every $\{c, \delta, p_0, p_1\}$, there exists $\overline{L} > 0$ such that for every $L > \overline{L}$, there exists a monotone-responsive equilibrium in symmetric strategies satisfying: (i) $q(\mathbf{a}) > 0$ if and only if $\mathbf{a} = (1, 1, ..., 1)$; (ii) the principal either abuses no agent or abuses only one agent. Similar to the proof of Proposition 3, the key step is to establish the following result:

**Proposition B'.** *For every $(c, \delta, p_0, p_1) \in \mathbb{R}_+ \times (0,1) \times (0,1) \times (0,1)$, there exists $\overline{L} > 0$ such that for every $L > \overline{L}$, there exists a triple $(\omega^*, \omega^{**}, q) \in \mathbb{R}_- \times \mathbb{R}_- \times (0,1)$ that solves the following three equations:*

$$\frac{q}{c}(\omega^* - c) = -\frac{1}{(\Psi^{**})^{n-1}} \tag{B.13}$$

$$\frac{q}{c}(\omega^{**} - c + b) = -\frac{n}{n + (n-1)l^*}\frac{1}{(\Psi^{**})^{n-1}} - \frac{(n-1)l^*}{n + (n-1)l^*}\frac{1}{(\Psi^{**})^{n-2}\Psi^*} \tag{B.14}$$

$$\frac{1}{\delta L} = q(\Psi^{**})^{n-1}(\Psi^* - \Psi^{**}) \tag{B.15}$$

*where*

$$\Psi^* \equiv \delta\Phi(\omega^*) + (1-\delta)p_0 \text{ and } \Psi^{**} \equiv \delta\Phi(\omega^{**}) + (1-\delta)p_1.$$

The proof of Proposition B' follows from similar steps as that of Proposition B, which will be available upon request. Notice that (B.13), (B.14), (B.15) together with (B.12) are sufficient for a monotone-responsive equilibrium, where $\widetilde{\pi}$ can be computed via (B.12) after fixing $\{\omega^*, \omega^{**}, q\}$.

## C Proof of Theorem 1: Symmetry

In this part, we establish the symmetric properties of all monotone-responsive equilibria satisfying $q(0,0) = q(1,0) = q(0,1) = 0$. We will show that the conviction probabilities in all monotone-

[21] In principle, we can also allow the mechanical types of different agents to adopt different reporting probabilities. For notation simplicity, we focus on environments where agents are symmetric.

responsive equilibria has this property in Appendix D. The conclusion in this section is summarized by the following proposition:

**Proposition C.** *In every monotone-responsive equilibrium where* $q(0,0) = q(1,0) = q(0,1) = 0$*, the principal chooses* $(\theta_1, \theta_2) = (0,1)$ *and* $(\theta_1, \theta_2) = (1,0)$ *with the same probability and the two agents share the same pair of reporting thresholds.*

*Proof of Proposition C:.* For $i \in \{1,2\}$, let $\beta_i$ be the probability that $\theta_i = 1$ conditional on $\theta_j = 1$. We have the following expressions on the cutoff types:

$$\omega_i^* = -c\frac{1 - qQ_{0,j}}{qQ_{0,j}} \tag{C.1}$$

and

$$\omega_i^{**} = -b - c\frac{1 - qQ_{1,j}}{qQ_{1,j}} \tag{C.2}$$

with

$$Q_{0,j} \equiv \delta\Phi(\omega_j^{**}) + (1-\delta)\alpha$$

and

$$Q_{1,j} \equiv \delta\Big[\beta_j\Phi(\omega_j^{**}) + (1-\beta_j)\Phi(\omega_j^{*})\Big] + (1-\delta)\alpha.$$

Without loss of generality, suppose the probability with which $\theta_i = 0$ is weakly higher compared to the probability with which $\theta_j = 0$. then $\beta_i \leq \beta_j$ and moreover, given that the equilibrium probability of misbehavior is interior, the principal's incentive constraints imply that the cost of setting $\theta_i = 0$ conditional on $\theta_j = 1$ is no more compared to the cost of setting $\theta_j = 0$ conditional on $\theta_i = 1$:

$$\frac{\delta q\Phi(\omega_j^{**})\Big(\Phi(\omega_i^{*}) - \Phi(\omega_i^{**})\Big)}{\delta q\Phi(\omega_i^{**})\Big(\Phi(\omega_j^{*}) - \Phi(\omega_j^{**})\Big)} \leq 1,$$

which is equivalent to:

$$\frac{\Phi(\omega_i^{*})\Phi(\omega_j^{**})}{\Phi(\omega_j^{*})\Phi(\omega_i^{**})} \leq 1. \tag{C.3}$$

First, we show that $\omega_1^* = \omega_2^*$ and $\omega_1^{**} = \omega_2^{**}$ when the probability of $\theta_1 = 0$ and the probability of $\theta_2 = 0$ are equal, i.e. $\beta_1 = \beta_2$. In this case, both probabilities are interior, which implies that (C.3) holds with equality. Suppose towards a contradiction that $\omega_1^* < \omega_2^*$, then (C.1) implies that $\omega_1^{**} > \omega_2^{**}$. But then we have $\Phi(\omega_1^*)\Phi(\omega_2^{**}) < \Phi(\omega_2^*)\Phi(\omega_1^{**})$, contradicting the equality in (C.3).

Next, we show that $\beta_1 = \beta_2$ in every equilibrium. Suppose towards a contradiction that $\beta_1 < \beta_2$,

i.e. $\theta_1 = 0$ occurs with strictly higher probability. Consider the following three cases:

1. If $\omega_1^* > \omega_2^*$, then (C.1) implies that $\omega_1^{**} < \omega_2^{**}$. This contradicts the requirement in (B.3) that $\Phi(\omega_1^*)\Phi(\omega_2^{**}) \leq \Phi(\omega_2^*)\Phi(\omega_1^{**})$.

2. If $\omega_1^* = \omega_2^*$, then (C.1) implies that $\omega_1^{**} = \omega_2^{**}$. However, (C.2), $\omega_1^* = \omega_2^*$, $\omega_1^{**} = \omega_2^{**}$ and $\beta_1 < \beta_2$ together imply that $\omega_1^{**} \neq \omega_2^{**}$, leading to a contradiction.

3. If $\omega_1^* < \omega_2^*$, then $\omega_1^{**} > \omega_2^{**}$ and we have $\Phi(\omega_1^*)\Phi(\omega_2^{**}) < \Phi(\omega_2^*)\Phi(\omega_1^{**})$. Therefore, the principal faces strictly lower cost to set $\theta_1 = 0$. Therefore in equilibrium, he will set $\theta_1 = 0$ with positive probability while setting $\theta_2 = 0$ with zero probability. This implies that $\beta_2 = 1$ and therefore

$$\omega_1^{**} = -b - c\frac{1 - \delta q\Phi(\omega_2^{**}) - (1-\delta)\alpha}{\delta q\Phi(\omega_2^{**}) + (1-\delta)\alpha}.$$

   Therefore, $\omega_1^* - \omega_1^{**} = b$. On the other hand, $\beta_1 \in (0,1)$ implies that $\omega_2^* - \omega_2^{**} < b$. However, the previous conclusions that $\omega_1^* < \omega_2^*$ and $\omega_1^{**} > \omega_2^{**}$ imply that $\omega_2^* - \omega_2^{**} > \omega_1^* - \omega_1^{**}$, leading to a contradiction.

□

# D Proof of Theorem 1: Conviction Probabilities

In this part, we show that $q(0,0) = q(1,0) = q(0,1) = 0$ in all monotone-responsive equilibria when $L$ is large enough. The conclusion is summarized by the following proposition:

**Proposition D.** *For every $\delta \in (0,1)$ and $c > 0$, there exists $\overline{L}(\delta, c) > 0$ such that $q(0,0) = q(1,0) = q(0,1) = 0$ and $q(1,1) \in (0,1)$ in every responsive equilibrium when $L > \overline{L}(\delta, c)$.*

To prove Proposition D, we rule out the possibilities of other types of equilibria when $L$ is large enough. For notation simplicity, let $\Phi_i^* \equiv \Phi(\omega_i^*)$ and for $i \in \{1,2\}$, let

$$\Psi_i^* \equiv \delta\Phi_i^* + (1-\delta)\alpha \text{ and } \Psi_i^{**} \equiv \delta\Phi_i^{**} + (1-\delta)\alpha.$$

## D.1 Complementarity & Substitutability of Principal's Actions

Our analysis starts from Lemma D.1 that formally establishes the complementarity and substitutability between the principal's choices of $\theta_1$ and $\theta_2$.

**Lemma D.1.** *In every equilibrium where $\omega_i^* > \omega_i^{**}$ for $i \in \{1,2\}$, then the principal's choice of $\theta_1$ and $\theta_2$ are strategic substitutes if the value of (3.7) is strictly positive and are strategic complements if the value of (3.7) is strictly negative.*

*Proof of Lemma D.1:* Given that $\theta_2 = 0$, the principal increases the probability of losing power by:

$$(\Psi_1^* - \Psi_1^{**})\Big((1-\Psi_2^{**})\big(q(1,0)-q(0,0)\big) + \Psi_2^{**}\big(q(1,1)-q(0,1)\big)\Big)$$

if he changes $\theta_1$ from 0 to 1. Similarly, given that $\theta_2 = 1$, the principal increases the probability of losing power by:

$$(\Psi_1^* - \Psi_1^{**})\Big((1-\Psi_2^{*})\big(q(1,0)-q(0,0)\big) + \Psi_2^{*}\big(q(1,1)-q(0,1)\big)\Big)$$

if he changes $\theta_1$ from 0 to 1. The first expression is greater than the second one if and only if:

$$(\Psi_1^* - \Psi_1^{**})(\Psi_2^* - \Psi_2^{**})\Big(q(1,0)+q(0,1)-q(0,0)-q(1,1)\Big) > 0$$

and vice versa. Since $\omega_i^* > \omega_i^{**}$ for both $i$, we know that $(\Psi_1^* - \Psi_1^{**})(\Psi_2^* - \Psi_2^{**}) > 0$, and therefore, the above inequality is equivalent to

$$q(1,0)+q(0,1)-q(0,0)-q(1,1) > 0,$$

which concludes the proof of Lemma D.1. □

The rest of the proof is organized as follows. In subsection D.2, we examine equilibria in which $\theta_1$ and $\theta_2$ are strategic substitutes from the principal's perspective. In subsection D.3, we examine equilibria in which $\theta_1$ and $\theta_2$ are strategic complements. In subsection D.4, we examine equilibria in the knife-edge case where the value of (3.7) equals to 0. To maintain the flow of the analysis, we will relegate the proofs of one of the technical lemma to subsection D.5.

## D.2 Value of (3.7) is Positive

In this subsection, we focus on equilibria in which the principal's decisions are strategic substitutes, i.e. $q(1,0)+q(0,1) < q(0,0)+q(1,1)$.

First, we claim that if either $q(0,1)$ or $q(1,0)$ is strictly positive, then $q(1,1) = 1$. To understand why, suppose towards a contradiction that $q(1,0), q(1,1) \in (0,1)$. Then whether agent 2 reports or

not will lead to the same posterior belief about the variable of interest $\theta_1\theta_2$. This can only be the case where $a_2$ is uninformative about $\theta_2$, which implies that $\omega_2^* = \omega_2^{**}$ and hence $\Phi(\omega_2^*) = \Phi(\omega_2^{**})$. This implies that the principal's cost of abusing agent 2 is 0 as it is proportional to $\Phi(\omega_2^*) - \Phi(\omega_2^{**})$, contradicting the fact that his probability of abusing agent 2 is strictly less than 1.

Given that $q(1,1) = 1$ and $q(0,0) = 0$, we have the following expressions for player 1's reporting cutoffs when he has and has not been abused:

$$\omega_1^* \equiv -c\frac{(1-\Psi_2^{**})(1-q(1,0))}{q(1,0)+\Psi_2^{**}(1-q(1,0)-q(0,1))}, \tag{D.1}$$

$$\omega_1^{**} \equiv -b - c\frac{(1-X_2)(1-q(1,0))}{q(1,0)+X_2(1-q(1,0)-q(0,1))}, \tag{D.2}$$

where

$$X_2 \equiv \frac{1-p_1-p_2}{1-p_1}\Psi_2^{**} + \frac{p_2}{1-p_1}\Psi_2^{*} \tag{D.3}$$

and $p_i$ is the probability with which $\theta_i = 0$. One observation is that $\omega_1^*$ is increasing in $\Psi_2^{**}$ and $q(1,0)$, and is decreasing in $q(0,1)$; $\omega_1^{**}$ is increasing in $X_2$ and $q(1,0)$, and is decreasing in $q(0,1)$. The distance between the two cutoffs is given by:

$$\omega_1^* - \omega_1^{**} = b - (\Psi_2^* - \Psi_2^{**})C_1 \tag{D.4}$$

where

$$C_1 \equiv c(1-q(0,1))(1-q(1,0)) \cdot \frac{p_2}{1-p_1}$$
$$\cdot \frac{1}{q(1,0)+X_2(1-q(1,0)-q(0,1))} \cdot \frac{1}{q(1,0)+\Psi_2^{**}(1-q(1,0)-q(0,1))}. \tag{D.5}$$

Symmetrically, one can obtain the expressions for $\omega_2^*$ and $\omega_2^{**}$ as well as the distance between them. Conditional on $\theta_2 = 1$, the probability with which the principal loses power is increased by:

$$(\Psi_1^* - \Psi_1^{**})\Big(q(1,0)+\Psi_2^{**}(1-q(1,0)-q(0,1))\Big) \tag{D.6}$$

if he chooses to set $\theta_1 = 0$. Similarly, if he sets $\theta_2 = 0$ given that $\theta_1 = 1$, this probability is increased by:

$$(\Psi_2^* - \Psi_2^{**})\Big(q(0,1)+\Psi_1^{**}(1-q(1,0)-q(0,1))\Big) \tag{D.7}$$

In equilibrium, neither (D.6) nor (D.7) can exceed $1/L$. In what follows, we establish a lower bound

for the maximum of these two expressions, which does not depend on $L$. This will be sufficient to rule out equilibria of this form when $L$ is large enough. Throughout the proof, we assume that $\omega_1^* \geq \omega_2^*$, which is without loss of generality. This leads to the following lemma on the comparison between $q(1,0)$ and $q(0,1)$, the proof of which can be found in subsection D.5:

**Lemma D.2.** *In every equilibrium where $\omega_1^* \geq \omega_2^*$, we have $q(1,0) \geq q(0,1)$.*

**Lower Bound on $\omega_1^*$:** For every $\epsilon > 0$,

1. Suppose $q(1,0) \geq \epsilon$, then
$$\omega_1^{**} \geq -b - c\frac{1-\epsilon}{\epsilon}. \tag{D.8}$$

2. Suppose $q(1,0) < \epsilon$, then $q(0,1) \in (0,\epsilon)$ according to Lemma D.1. Therefore, we have:
$$\begin{aligned}
\omega_2^* &= -c\frac{(1-\Psi_1^{**})(1-q(0,1))}{q(0,1)+\Psi_1^{**}(1-q(1,0)-q(0,1))} \\
&\geq -c\frac{\Big(1-\delta\Phi(\omega_1^*-b)-(1-\delta)\alpha\Big)(1-q(0,1))}{q(0,1)+\Big(\delta\Phi(\omega_1^*-b)+(1-\delta)\alpha\Big)\Big(1-q(1,0)-q(0,1)\Big)} \\
&\geq -c\frac{\Big(1-\delta\Phi(\omega_2^*-b)-(1-\delta)\alpha\Big)(1-q(0,1))}{q(0,1)+\Big(\delta\Phi(\omega_2^*-b)+(1-\delta)\alpha\Big)\Big(1-q(1,0)-q(0,1)\Big)} \\
&\geq -c\frac{1-\delta\Phi(\omega_2^*-b)-(1-\delta)\alpha}{(1-\epsilon)\Big(\delta\Phi(\omega_2^*-b)+(1-\delta)\alpha\Big)}.
\end{aligned} \tag{D.9}$$

As have shown in Appendix A, there exists a solution to the following equation:

$$\omega_2^* = -c\frac{1-\delta\Phi(\omega_2^*-b)-(1-\delta)\alpha}{(1-\epsilon)\Big(\delta\Phi(\omega_2^*-b)+(1-\delta)\alpha\Big)},$$

which is denoted by $\underline{\omega}^*(\epsilon)$ such that (D.9) is satisfied only when $\omega_2^* \geq \underline{\omega}^*(\epsilon)$. Since $\underline{\omega}^*(\epsilon)$ is decreasing in $\epsilon$, a lower bound for $\omega_1^*$ is then given by:

$$\underline{\omega}_1^* \equiv \sup_{\epsilon\in[0,1]}\Big\{\min\Big\{-b-c\frac{1-\epsilon}{\epsilon}, \underline{\omega}^*(\epsilon)\Big\}\Big\}, \tag{D.10}$$

which is finite and moreover, does not depend on $L$.

**Upper Bound on $C_1$:** The key to bound $C_1$ is to bound the term

$$\frac{1}{q(1,0)+\Psi_2^{**}(1-q(1,0)-q(0,1))} \tag{D.11}$$

from above. For every $\epsilon>0$, consider the following two cases:

1. If $q(1,0)\geq\epsilon$, then (D.11) is no more than $1/\epsilon$.

2. If $q(1,0)<\epsilon$, then $q(0,1)<\epsilon$ according to Lemma D.1. Let $\underline{\omega}_2^{**}(\epsilon)$ be the smallest root of the following equation:
$$\omega\equiv -b-c\frac{1-\delta\Phi(\omega)-(1-\delta)\alpha}{\Big(\delta\Phi(\omega)+(1-\delta)\alpha\Big)(1-\epsilon)}, \tag{D.12}$$
which is a lower bound for $\omega_2^{**}$ given that $q(1,0),q(0,1)\in[0,\epsilon]$. The upper bound of (D.11) is then given by:
$$\frac{1}{q(1,0)+\Psi_2^{**}(1-q(1,0)-q(0,1))}\leq\frac{1}{\Phi(\underline{\omega}_2^{**}(\epsilon))(1-2\epsilon)}. \tag{D.13}$$

Summarizing these two cases, we have:

$$C_1\leq cY^2 \tag{D.14}$$

where

$$Y\equiv\inf_{\epsilon\in[0,1]}\Big\{\max\big\{1/\epsilon,\frac{1}{\Phi(\underline{\omega}_2^{**})(1-2\epsilon)}\big\}\Big\}.$$

**Lower Bound on the Maximum of (D.6) and (D.7):** In this last step, we establish a lower bound on the maximum of (D.6) and (D.7). A useful inequality that will be used is that for every $\omega',\omega''$ with $\omega'>\omega''$,

$$\Phi(\omega')-\Phi(\omega'')\geq(\omega'-\omega'')\min_{\omega\in[\omega',\omega'']}\phi(\omega). \tag{D.15}$$

We consider two cases. First, consider the case in which $\Phi(\omega_1^*)-\Phi(\omega_1^{**})\geq\Phi(\omega_2^*)-\Phi(\omega_2^{**})$. Using the fact that $\Psi_i^*-\Psi_i^{**}=\delta(\Phi(\omega_i^*)-\Phi(\omega_i^{**}))$, we have:

$$\frac{\delta}{\min_{\omega\in[\omega_1^{**},\omega_1^*]}\phi(\omega)}\Big(\Phi(\omega_1^*)-\Phi(\omega_1^{**})\Big)\geq\omega_1^*-\omega_1^{**}=b-C_1(\Psi_2^*-\Psi_2^{**})\geq b-C_1(\Psi_1^*-\Psi_1^{**}). \tag{D.16}$$

This together with (D.14) gives an lower bound for $\Psi_1^*-\Psi_1^{**}$. Moreover,

$$\begin{aligned}q(1,0)+\Psi_2^{**}(1-q(1,0)) &\geq q(1,0)+\Psi_2^{**}\big(1-q(1,0)-q(0,1)\big)\\ &\geq \frac{c\big(1-q(1,0)\big)(1-\Psi_2^{**})}{|\underline{\omega}_1^*|},\end{aligned} \tag{D.17}$$

where the last inequality uses (D.1) as well as the previous conclusion that $\underline{\omega}_1^*$ is a lower bound of $\omega_1^*$. This gives a lower bound on $q(1,0)$. The two parts together imply a lower bound on (D.6).

Second, consider the case in which $\Phi(\omega_1^*) - \Phi(\omega_1^{**}) < \Phi(\omega_2^*) - \Phi(\omega_2^{**})$. Let

$$\beta \equiv \frac{\omega_1^* - \omega_1^{**}}{b}. \tag{D.18}$$

Since $X_2 > \Psi_2^{**}$, we have $\beta \in (0,1)$. First, recall that $\underline{\omega}_1^*$ is the lower bound on $\omega_1$, we have:

$$\frac{1}{\delta}(\Psi_1^* - \Psi_1^{**}) = \Phi(\omega_1^*) - \Phi(\omega_1^{**}) \geq \beta b \phi(\underline{\omega}_1^* - b). \tag{D.19}$$

On the other hand, (D.4) and (D.14) imply that:

$$\Psi_2^* - \Psi_2^{**} = (1-\beta)b/C_1 \geq \frac{(1-\beta)bY^2}{c} \tag{D.20}$$

Since the pdf of normal distribution increases in $\omega$ when $\omega < 0$, (D.20) leads to a lower bound on $\omega_2^{**}$. We denote this lower bound by $\widetilde{\omega}(\beta)$. By definition, $\widetilde{\omega}(\beta)$ decreases with $\beta$.

1. When $\beta \geq 1/2$, (D.19) implies a lower bound for $\Phi(\omega_1^*) - \Phi(\omega_1^{**})$. Applying (D.17),[22] one can obtain a lower bound for $q(1,0)$. These together lead to a lower bound on (D.6).

2. When $\beta < 1/2$, we have $\omega_2^{**} \geq \widetilde{\omega}(1/2)$ and furthermore,

$$\Psi_2^* - \Psi_2^{**} \geq \frac{b}{2C_1}.$$

   The lower bound on $\omega_2^{**}$ also leads to a lower bound on $q(0,1) + \Psi_1^{**}(1 - q(1,0) - q(0,1))$, as (D.2) implies:

$$\widetilde{\omega}(1/2) \leq \omega_2^{**} \leq \omega_2^* = -c\frac{(1-\Psi_1^{**})(1-q(0,1))}{q(0,1) + \Psi_1^{**}(1-q(1,0)-q(0,1))},$$

   which leads to:

$$q(0,1) + \Psi_1^{**}(1-q(1,0)-q(0,1)) \geq \frac{(1-\Psi_1^{**})(1-q(0,1))}{-\widetilde{\omega}(1/2)/c}. \tag{D.21}$$

   Since $1-\Psi_1^{**} \geq \delta - \delta\Phi(0)$ and $1-q(0,1) \geq 1/2$, the lower bound on $q(0,1)+\Psi_1^{**}(1-q(1,0)-q(0,1))$ is strictly bounded away from 0. This leads to a uniform lower bound on (D.7).

[22] The validity of inequality (D.17) does not depend on the sign of $\Psi_1^* - \Psi_1^{**} - \Psi_2^* + \Psi_2^{**}$.

### D.3 Value of (3.7) is Negative

Next, we study the case where $q(1,0)+q(0,1) > q(0,0)+q(1,1)$, or in another word, the choice of $\theta_1$ and $\theta_2$ are strategic complements from the principal's perspective. Lemma D.1 implies that conditional on abusing one agent, the principal will have a strict incentive to abuse the other agent. Therefore in such equilibria, either both agents are abused or no agent is abused.

We start from two observations. First, $q(1,1) = 1$ in all such equilibria. This is because if $q(1,1) \in (0,1)$ and $q(1,0)+q(0,1) > q(0,0)+q(1,1)$, then one of the agent's report is not informative about the state, leading to a contradiction. Second, due to the strategic complementarity between $\theta_1$ and $\theta_2$, agent $i$'s belief about agent $j$'s probability of submitting a report is strictly higher when $\theta_i = 0$ compared to $\theta_i = 1$. This implies that:

$$\min\{\omega_1^* - \omega_1^{**}, \omega_2^* - \omega_2^{**}\} \geq b. \tag{D.22}$$

By setting $\theta_1 = \theta_2 = 1$, the principal's probability of losing power is increased by at least

$$(\Psi_1^* - \Psi_1^{**})\Big(\Psi_2^*(1-q(0,1)) + (1-\Psi_2^*)q(1,0)\Big) + (\Psi_2^* - \Psi_2^{**})\Big(\Psi_1^{**}(1-q(1,0)) + (1-\Psi_1^{**})q(0,1)\Big), \tag{D.23}$$

compared to the case in which he sets $\theta_1 = \theta_2 = 0$. Therefore, the value of (D.23) cannot exceed $2/L$. The rest of this proof establishes a lower bound on (D.23) that applies uniformly across all $L$. This in turn implies that when $L$ is large enough, equilibria that exhibit strategic complementarities between $\theta_1$ and $\theta_2$ do not exist.

First, $\max\{q(0,1), q(1,0)\} \geq 1/2$ since $q(0,1)+q(1,0) \geq 1$. Without loss of generality, we assume that $q(1,0) \geq 1/2$. Second, agent $i$ has a dominant strategy of not reporting when $\omega_i > 0$, so $1 - \Psi_i^* \geq \delta(1-\Phi(0))$. Third, player 1's reporting threshold when $\theta_1 = 0$ is:

$$\omega_1^* = -c\frac{(1-Q_2^H)(1-q(1,0))}{Q_2^H(1-q(0,1)) + (1-Q_2^H)q(1,0)} \tag{D.24}$$

where $Q_2^H$ is the probability with which player 2 submits a report conditional on $\theta_1 = 0$. One can verify that the RHS of (D.24) is strictly increasing in $Q_2^H$. Therefore,

$$\omega_1^* \geq -c\frac{1-q(1,0)}{q(1,0)} \geq -c.$$

According to (D.22), we have:

$$\frac{1}{\delta}(\Psi_1^* - \Psi_1^{**}) = \Phi(\omega_1^*) - \Phi(\omega_1^{**}) \geq b \min_{\omega \in [-b-c,0]} \phi(\omega). \tag{D.25}$$

The uniform lower bound on (D.23) is then given by:

$$\underbrace{(\Psi_1^* - \Psi_1^{**})}_{\text{use (D.25)}} \Big( \underbrace{\Psi_2^*(1 - q(0,1))}_{\geq 0} + \underbrace{(1 - \Psi_2^*)}_{\geq \delta(1-\Phi(0))} \underbrace{q(1,0)}_{\geq 1/2} \Big) + \underbrace{(\Psi_2^* - \Psi_2^{**}) \Big( \Psi_1^{**}(1 - q(1,0)) + (1 - \Psi_1^{**}) q(0,1) \Big)}_{\geq 0}$$

$$\geq \frac{\delta^2 b}{2}(1 - \Phi(0)) \min_{\omega \in [-b-c,0]} \phi(\omega), \tag{D.26}$$

which concludes the proof.

## D.4 Value of (3.7) equals to 0

**Part I:** We show that each agent is being abused with strictly positive probability and $q(1,1) = 1$. The implications of these conclusions are:

1. $q(1,0) + q(0,1) = 1$.

2. The marginal cost of abusing each agent is the same.

$$(\Psi_1^* - \Psi_1^{**}) q(1,0) = (\Psi_2^* - \Psi_2^{**}) q(0,1). \tag{D.27}$$

Suppose towards a contradiction that agent 1 is abused with probability 0, then whether agent 1's reports or not does not change the posterior belief about $\theta_1\theta_2$. Given that $c > 0$, agent 1 will never report, which implies that the principal will have a strict incentive to abuse him. This contradicts the responsiveness requirement.

Suppose towards a contradiction that $q(1,1) \in (0,1)$, then either $q(1,0) \in (0,1)$ or $q(0,1) \in (0,1)$ or both. Suppose $q(1,0) \in (0,1)$, then whether agent 2's reports or not does not change the posterior belief about $\theta_1\theta_2$. Given that $c > 0$, agent 2 will never report, which implies that the principal will have a strict incentive to abuse him. This contradicts the responsiveness requirement.

**Part II:** We place a lower bound on the value of (D.27) that uniformly applies across all $L$. Without loss of generality, we assume that $q(1,0) \geq q(0,1)$, and therefore, $q(1,0) \geq 1/2$. The expressions for

agent 1's reporting cutoffs are given by:

$$\omega_1^* = -c\frac{q(0,1)}{q(1,0)}\Big(1 - p_x\Psi_2^* - (1-p_x)\Psi_2^{**}\Big)$$

and

$$\omega_1^{**} = -b - c\frac{q(0,1)}{q(1,0)}\Big(1 - p_y\Psi_2^* - (1-p_y)\Psi_2^{**}\Big)$$

for some $p_x, p_y \in [0,1]$, which are agent 1's beliefs about $\theta_2$ conditional on the realization of $\theta_1$. The difference between them is then:

$$\omega_1^* - \omega_1^{**} = b - c\frac{q(0,1)}{q(1,0)}(p_x - p_y)(\Psi_2^* - \Psi_2^{**}). \tag{D.28}$$

where the absolute value of

$$c\frac{q(0,1)}{q(1,0)}(p_x - p_y)$$

is at most $c$. To bound the LHS of (D.27) from below, we proceed in two steps.

**Substep 1: Lower bound on $\omega_1^*$** According to the expression for $\omega_1^*$ and using the assumption that $q(1,0) \geq q(0,1)$, we have:

$$\omega_1^* \geq -c\Big(1 - p_x\Psi_2^* - (1-p_x)\Psi_2^{**}\Big) \geq -c\delta(1-\Phi(0)). \tag{D.29}$$

Let this lower bound be $\underline{\omega}_1^*$.

**Substep 2: Lower bound on (D.27)** This can be accomplished by establishing strictly positive lower bounds on either of the following expressions: $\Psi_1^* - \Psi_1^{**}$ or $q(0,1)(\Psi_2^* - \Psi_2^{**})$. The former is sufficient since $q(1,0) \geq 1/2$.

The case in which $p_x - p_y \leq 0$ is trivial, as $\omega_1^* - \omega_1^{**} \geq b$. The lower bound on $\omega_1^*$ then implies a strictly positive lower bound on $\Psi_1^* - \Psi_1^{**}$. The case in which $p_x - p_y > 0$ follows similarly from the last step of subsection C.2. To illustrate the details, we consider two cases separately.

First, suppose $\Psi_1^* - \Psi_1^{**} \geq \Psi_2^* - \Psi_2^{**}$, then we have:

$$\frac{\Psi_1^* - \Psi_1^{**}}{\phi(\underline{\omega}_1^* - b)} \geq \omega_1^* - \omega_1^{**} = b - c(\Psi_2^* - \Psi_2^{**}) \geq b - c(\Psi_1^* - \Psi_1^{**}). \tag{D.30}$$

which yields a strictly positive lower bound on $\Psi_1^* - \Psi_1^{**}$.

Second, suppose $\Psi_1^* - \Psi_1^{**} < \Psi_2^* - \Psi_2^{**}$, then let $\beta \equiv (\omega_1^* - \omega_1^{**})/b$ which is between 0 and 1 due to the assumption that $p_x - p_y > 0$. Equality (D.28) implies that:

$$\omega_1^* - \omega_1^{**} = b - c\frac{q(0,1)}{q(1,0)}(p_x - p_y)(\Psi_2^* - \Psi_2^{**}) \geq b - c(\Psi_2^* - \Psi_2^{**})$$

which yields

$$\Psi_2^* \geq \Psi_2^* - \Psi_2^{**} \geq (1-\beta)b/c. \tag{D.31}$$

This leads to a lower bound on the cutoff $\omega_2^*$ for each $\beta$. We denote this lower bound by $\widetilde{\omega}(\beta)$, which is a decreasing function of $\beta$. On the other hand, we also have:

$$\frac{1}{\delta}(\Psi_1^* - \Psi_1^{**}) = \Phi(\omega_1^*) - \Phi(\omega_1^{**}) \geq \beta b\phi(\underline{\omega}_1^* - b). \tag{D.32}$$

Now consider two subcases, depending on the comparison between $\beta$ and $1/2$.

1. If $\beta \geq 1/2$, then (D.32) implies that

$$\Psi_1^* - \Psi_1^{**} \geq b\delta\phi(\underline{\omega}_1^* - b)/2. \tag{D.33}$$

2. If $\beta < 1/2$, then (D.31) implies that:

$$\Psi_2^* - \Psi_2^{**} \geq b/2c \tag{D.34}$$

Since

$$\omega_2^* = -c(1-Q)\frac{q(1,0)}{q(0,1)} \geq \underline{\omega}_2(\beta) \tag{D.35}$$

where $Q$ is some number between 0 and $(1-\delta)\alpha + \delta\Phi(0)$. This yields the following lower bound on $q(0,1)$, namely

$$q(0,1) \geq \frac{-c(1-Q)q(1,0)}{\underline{\omega}_2(\beta)} \geq -\frac{c}{2\underline{\omega}_2(\beta)} \tag{D.36}$$

which is strictly bounded above 0 for all $\beta < 1/2$. This together with (D.34) lead to the following lower bound on the RHS of (D.27):

$$q(0,1)(\Psi_2^* - \Psi_2^{**}) \geq -\frac{b}{4\underline{\omega}_2(\beta)}. \tag{D.37}$$

### D.5 Proof of Lemma D.2

Suppose towards a contradiction that in some equilibrium with strategic substitutability, $\omega_1^* > \omega_2^*$ but $q(1,0) < q(0,1)$, then (D.1) implies that $\Phi(\omega_2^{**}) > \Phi(\omega_1^{**})$ or equivalently, $\omega_2^{**} > \omega_1^{**}$. This together with $\omega_i^* > \omega_i^{**}$ for both $i$ imply that:

$$\omega_1^{**} < \omega_2^{**} < \omega_2^* < \omega_1^*. \tag{D.38}$$

We start from showing that $p_1, p_2 > 0$. Suppose towards a contradiction that $p_1 = 0$ and $p_2 > 0$, then (D.3) implies that $X_1 = \Psi_1^{**}$. Therefore, $\omega_2^* - \omega_2^{**} = b > \omega_1^* - \omega_1^{**}$, contradicting (D.38). Suppose towards a contradiction that $p_1 > 0$ and $p_2 = 0$, then

$$p_1 \frac{\Psi_1^*}{\Psi_1^{**}} + p_2 \frac{1-\Psi_2^*}{1-\Psi_2^{**}} > p_2 \frac{\Psi_2^*}{\Psi_2^{**}} + p_1 \frac{1-\Psi_1^*}{1-\Psi_1^{**}}, \tag{D.39}$$

which is to say that the public's posterior belief attaching to $\theta_1\theta_2 = 0$ is strictly higher when agent 1 is the only one that reports compared to the case where agent 2 is the only one that reports. Therefore, $q(1,0) \geq q(0,1)$, leading to a contradiction.

Given that we have already shown that $p_1, p_2 > 0$, while the responsiveness requirement implies that $\theta_1\theta_2 = 0$ with probability less than 1, i.e. $p_1, p_2 < 1$, we know that both of them are interior so that (D.6) and (D.7) must be equal. Applying the expression for reporting threshold (D.1) to both agents, we have:

$$\begin{aligned} \left|\frac{\omega_1^*}{\omega_2^*}\right| &= \frac{1-\Psi_2^{**}}{1-\Psi_1^{**}} \cdot \frac{1-q(1,0)}{1-q(0,1)} \cdot \frac{q(0,1) + \Psi_1^{**}(1-q(1,0)-q(0,1))}{q(1,0) + \Psi_2^{**}(1-q(1,0)-q(0,1))}. \\ &= \frac{1-\Psi_2^{**}}{1-\Psi_1^{**}} \cdot \frac{1-q(1,0)}{1-q(0,1)} \cdot \frac{\Psi_1^* - \Psi_1^{**}}{\Psi_2^* - \Psi_2^{**}}. \end{aligned} \tag{D.40}$$

Since

$$\frac{1-\Psi_1^{**}}{1-\Psi_2^{**}} < \frac{\Psi_1^* - \Psi_1^{**}}{\Psi_1^* - \Psi_2^{**}} \leq \frac{\Psi_1^* - \Psi_1^{**}}{\Psi_2^* - \Psi_2^{**}}$$

Plugging this back in, we have:

$$1 \geq \left|\frac{\omega_1^*}{\omega_2^*}\right| > \frac{1-q(1,0)}{1-q(0,1)} > 1, \tag{D.41}$$

leading to a contradiction.

# E Proof of Theorem 1: Comparisons

First, we show that $\omega_m^* > \omega_s^*$. Suppose towards a contradiction that $\omega_m^* \leq \omega_s^*$, then the comparison between (3.1) and (3.8) implies that

$$q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big) \leq q_s.$$

Therefore,

$$\begin{aligned} q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) &\leq q_s\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) \\ = 1/\delta L = q_m\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big)\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big). & \qquad \text{(E.1)} \end{aligned}$$

On the other hand, since $\omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**}$ and $\omega_m^* < \omega_s^*$, we have:

$$\Phi(\omega_m^*) - \Phi(\omega_m^{**}) < \Phi(\omega_s^*) - \Phi(\omega_s^{**}). \qquad \text{(E.2)}$$

Inequality (E.2) implies that

$$\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_s^*) - \Phi(\omega_s^{**})\Big) > \Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big), \qquad \text{(E.3)}$$

which contradicts (E.1). This implies that $\omega_m^* > \omega_s^*$. Moreover, according to Lemma 3.2,

$$0 < \omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**},$$

we know that $\omega_m^* > \omega_s^*$ implies $\omega_m^{**} > \omega_s^{**}$. The comparison between $\mathcal{I}_s$ and $\mathcal{I}_m$ immediately follows, as $\omega_m^{**} > \omega_s^{**}$ and $\omega_m^* - \omega_m^{**} < \omega_s^* - \omega_s^{**}$ imply that $\mathcal{I}_s > \mathcal{I}_m$. The comparison between $\widetilde{\pi}_s$ and $\widetilde{\pi}_m$ can then be obtained by comparing (3.6) to (3.14), which yields $\widetilde{\pi}_s < \widetilde{\pi}_m$.

Next, we show $q_m > q_s$. Given that $\omega_m^* > \omega_s^*$, then the comparison between (3.1) and (3.8) implies that $q_m Q_0 > q_s$. That is $1 \geq Q_0 > q_s/q_m$, which implies that $q_m > q_s$.

# F Proof of Theorem 2

According to Lemma 3.3, we only need to show that $\omega_m^* \to -\infty$ as $L \to \infty$. Suppose towards a contradiction that for some $(c, \delta) \in \mathbb{R}_+ \times (0,1)$ and $\epsilon > 0$, there exists $\omega^* \in \mathbb{R}_-$ such that for every $\overline{L} > \overline{L}(c, \delta)$, there exists $(L, c, \delta)$ with $L \geq \overline{L}$ and a monotone-responsive equilibrium under $(L, c, \delta)$ in

which the first cutoff $\omega_m^* \in B(\omega^*, \epsilon)$.

Consider a sequence of such equilibria as $L \to \infty$. The principal's indifference condition:

$$\frac{1}{\delta L} = q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha\Big)\Big(\Phi(\omega_m^*) - \Phi(\omega_m^{**})\Big) \tag{F.1}$$

implies that as $L \to \infty$, the LHS converges to 0. Therefore, either $q_m \to 0$ or $\Phi(\omega_m^*) - \Phi(\omega_m^{**}) \to 0$. As $\omega_m^* \geq \omega^* - \epsilon$ and according to Lemma 3.2, $|\omega_m^* - \omega_m^{**}| \in (0, b)$, we know that $\omega_m^* - \omega_m^{**} \to 0$ once $\Phi(\omega_m^*) - \Phi(\omega_m^{**}) \to 0$.

Next, we show that $\omega_m^* \to -\infty$ as $q_m \to 0$. According to (3.8), suppose towards a contradiction that $\omega_m^*$ is finite, then the LHS converges to 0 while the RHS is strictly negative, leading to a contradiction. Lastly, we rule out the possibility that $\omega_m^* - \omega_m^{**} \to 0$ when $q_m$ is bounded away from 0, i.e. there exists $\underline{q} \in (0,1)$ such that $q_m \geq \underline{q}$ along this sequence. To see this, rewrite (3.8) and (3.9) as:

$$\frac{q_m}{c}(\omega_m^* - c) = -\frac{1}{\delta\Phi(\omega_m^*) + (1-\delta)\alpha} \tag{F.2}$$

and

$$\frac{q_m}{c}(\omega_m^{**} + b - c) = -\frac{2}{l^*+2}\frac{1}{\delta\Phi(\omega_m^*) + (1-\delta)\alpha} - \frac{l^*}{l^*+2}\frac{1}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha}. \tag{F.3}$$

Subtracting (F.3) from (F.2), we obtain:

$$\frac{q_m}{c}\Big(\omega_m^* - (\omega_m^{**} + b)\Big) = \frac{l^*}{l^*+2}\Big(\frac{1}{\delta\Phi(\omega_m^*) + (1-\delta)\alpha} - \frac{1}{\delta\Phi(\omega_m^{**}) + (1-\delta)\alpha}\Big). \tag{F.4}$$

The RHS of (F.4) converges to 0 as $\Phi(\omega_m^*) - \Phi(\omega_m^{**}) \to 0$ while the LHS is strictly less than $-\underline{q}b/c$ since $\omega_m^* - \omega_m^{**} \to 0$, which leads to a contradiction.

# G Proof of Propositions 5 & 6

According to Theorem 2, when $c > c^*$ and $L > \overline{L}(\delta, c)$, $q(0,0) = q(1,0) = q(0,1) = 0$, $q(1,1) \in (0,1)$ in every responsive equilibrium, which will be the focus of this proof.

## G.1 Proof of Proposition 5

We start from analyzing the agents' incentives. Agent 1's reporting cutoffs are given by:

$$\omega_1^* = c + \frac{1}{q\Psi_2^{**}}\Big\{\Psi_2^{**}\big(t_1(1,1) - t_1(0,1)\big) + (1-\Psi_2^{**})\big(t_1(1,0) - t_1(0,0)\big) - c\Big\}$$

and

$$\omega_1^{**} = -b + c + \frac{1}{qQ_2}\Big\{Q_2\big(t_1(1,1) - t_1(0,1)\big) + (1-Q_2)\big(t_1(1,0) - t_1(0,0)\big) - c\Big\}$$

where

$$Q_2 \equiv \frac{1-p_1-p_2}{1-p_1}\Psi_2^{**} + \frac{p_2}{1-p_1}\Psi_2^{*}.$$

Using the equilibrium condition that:

$$\mathcal{I}_m = \frac{\pi^*}{1-\pi^*}\Big/\frac{p_1+p_2}{1-p_1-p_2},$$

we know that:

$$p_1 + p_2 = \frac{l^*}{l^* + \mathcal{I}_m}.$$

One can then obtain:

$$Q_2 = \Psi_2^{**}\frac{\mathcal{I}_m + (1-\alpha)l^*\mathcal{I}_2}{(1-\alpha)l^* + \mathcal{I}_m},$$

where $\alpha \equiv p_1/(p_1+p_2)$. Let $\Delta_1 \equiv t_1(1,0) - t_1(0,0)$ and $\Delta_2 \equiv t_2(0,1) - t_2(0,0)$. Subtracting $\omega_1^{**}$ from $\omega_1^{*}$, we get:

$$\omega_1^* - \omega_1^{**} = b + \frac{1}{q}\Big(\frac{1}{\Psi_2^{**}} - \frac{1}{Q_2}\Big)(\Delta_1 - c). \tag{G.1}$$

Similarly, the distance between agent 2's reporting cutoffs is given by:

$$\omega_2^* - \omega_2^{**} = b + \frac{1}{q}\Big(\frac{1}{\Psi_1^{**}} - \frac{1}{Q_1}\Big)(\Delta_2 - c). \tag{G.2}$$

That is, whether $\omega_i^* - \omega_i^{**}$ is larger or smaller than $b$ only depends on the sign of $\Delta_i - c$.

Under the transfer scheme in Proposition 5, each agent's incentive to report does not depend on his belief about the other agent's strategy, i.e.

$$\omega_1^* = \omega_2^* = c - \frac{c}{q_m}, \quad \omega_1^{**} = \omega_2^{**} = -b + c - \frac{c}{q_m}.$$

The principal's incentive constraint is given by following indifference condition:

$$1/L = q_m(\Psi_1^* - \Psi_1^{**})\Psi_2^{**} = q_m(\Psi_2^* - \Psi_2^{**})\Psi_1^{**}.$$

As $q_m \to 0$ when $L \to \infty$, we know that $\omega^*, \omega^{**} \to -\infty$. The informativeness ratio $\mathcal{I}_m$ converges to $\infty$ and the equilibrium probability of crime, equals to $p_1 + p_2$, converges to 0.

## G.2 Proof of Proposition 6

Recall the definitions of $\Delta_1$ and $\Delta_2$. Without loss of generality, let $t_1(0,0) = t_2(0,0) = 0$. Then $t_1(1,0) = -t_2(1,0) = \Delta_1$, $-t_1(0,1) = t_2(0,1) = \Delta_2$. Let $t_1(1,1) = T$, then $t_2(1,1) = -T$. The two players' reporting cutoffs are then given by:

$$\omega_1^* = c + \frac{1}{q}(T + \Delta_2 - \Delta_1) + \frac{1}{q\Psi_2^{**}}(\Delta_1 - c), \tag{G.3}$$

$$\omega_1^{**} = -b + c + \frac{1}{q}(T + \Delta_2 - \Delta_1) + \frac{1}{qQ_2}(\Delta_1 - c), \tag{G.4}$$

$$\omega_2^* = c + \frac{1}{q}(\Delta_1 - \Delta_2 - T) + \frac{1}{q\Psi_1^{**}}(\Delta_2 - c), \tag{G.5}$$

$$\omega_2^{**} = -b + c + \frac{1}{q}(\Delta_1 - \Delta_2 - T) + \frac{1}{qQ_1}(\Delta_2 - c). \tag{G.6}$$

We consider three cases separately, depending on the signs of $\Delta_1 - c$ and $\Delta_2 - c$.

### G.2.1 Case 1: $\Delta_1, \Delta_2 \geq c$

Suppose $\Delta_1, \Delta_2 \geq c$, then

$$\omega_1^{**} \geq -b + c + \frac{1}{q}(T + \Delta_2 - \Delta_1) \text{ and } \omega_2^{**} \geq -b + c + \frac{1}{q}(\Delta_1 - \Delta_2 - T)$$

Therefore,

$$\omega_1^{**} + \omega_2^{**} \geq -2b + 2c, \tag{G.7}$$

which implies that $\max\{\omega_1^{**}, \omega_2^{**}\} \geq -b + c$. Since $\mathcal{I}_m = \min\{\mathcal{I}_1, \mathcal{I}_2\}$, we know that

$$\mathcal{I}_m \leq \frac{1}{\delta\Phi(-b+c) + (1-\delta)\alpha} \leq \frac{1}{\delta\Phi(-b+c)}, \tag{G.8}$$

which establishes the uniform upper bound.

### G.2.2 Case 2: $\Delta_1, \Delta_2 < c$

Let

$$X \equiv \frac{1}{q}(\Delta_1 - \Delta_2 - T).$$

Without loss of generality, assume $X \geq 0$. Let $\beta \in (0,1)$ be the probability with which agent 1 is abused conditional on the principal being guilty. The expressions for the two cutoffs imply that:

$$\frac{\omega_2^* - c - X}{\omega_2^{**} + b - c - X} = \frac{(1-\beta)l^*\mathcal{I}_1 + \mathcal{I}_m}{(1-\beta)l^* + \mathcal{I}_m} \quad \text{and} \quad \frac{\omega_1^* - c + X}{\omega_1^{**} + b - c + X} = \frac{\beta l^*\mathcal{I}_2 + \mathcal{I}_m}{\beta l^* + \mathcal{I}_m}. \tag{G.9}$$

We start with the following Lemma:

**Lemma G.1.** *There exists a function $\epsilon : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$ such that for every $\eta \in (0,1)$, if $\beta \leq 1-\eta$ and $\omega_2^* < -M$, then $\mathcal{I}_m < 1 + \epsilon(M,\eta)$.*

*Proof of Lemma G.1:* Since $\Delta_2 < c$, we know that $\omega_2^* - \omega_2^{**} < b$. Since $X \geq 0$, $\omega_2^* - c - X < 0$ and $\omega_2^{**} + b - c - X < 0$,

$$\frac{\omega_2^* - c - X}{\omega_2^{**} + b - c - X} < \frac{\omega_2^* - c}{\omega_2^{**} + b - c} \leq \frac{M + c}{M + c - b} = 1 + \frac{b}{M + c - b}$$

On the other hand, since $\mathcal{I}_1 \geq \mathcal{I}_m \geq 1$, we know that:

$$1 + \frac{b}{M + c - b} \geq \frac{(1-\beta)l^*\mathcal{I}_1 + \mathcal{I}_m}{(1-\beta)l^* + \mathcal{I}_m} \geq \frac{(1-\beta)l^*\mathcal{I}_m + \mathcal{I}_m}{(1-\beta)l^* + \mathcal{I}_m} \geq \frac{\eta l^*\mathcal{I}_m + \mathcal{I}_m}{\eta l^* + \mathcal{I}_m}$$

This places an upper bound on $\mathcal{I}_m$, which converges to 1 as $M \to -\infty$. □

Lemma G.1 implies that for every $\eta \in (0,1)$, if $\beta \leq 1-\eta$, the informativeness of report is bounded from above by:

$$\max_{M \in \mathbb{R}_+} \Big\{ \min\{1 + \epsilon(M,\eta), \frac{1}{\Phi(-M-b)}\} \Big\}, \tag{G.10}$$

which is bounded from above for every given $\eta$. Therefore, in order to establish a uniform upper bound on $\mathcal{I}_m$, we only need to show that unbounded informativeness cannot arise when $\alpha$ is close to or equals to 1. That is to say, it is without loss to consider cases in which $\beta \geq 1/2$. Therefore, agent 1 is abused with strictly positive probability, which implies that $\mathcal{I}_m = \mathcal{I}_1 \leq \mathcal{I}_2$.

Suppose towards a contradiction that for every $\overline{\mathcal{I}} > 0$, there exists $\{\Delta_1, \Delta_2, X\}$ under which there exists an equilibrium in which $\mathcal{I}_m > \overline{\mathcal{I}}$. In what follows, we consider two subcases separately.

**Subcase 1: $\boldsymbol{\omega_1^* - \omega_1^{**} \geq \omega_2^* - \omega_2^{**}}$** Since $\mathcal{I}_1 \leq \mathcal{I}_2$, we know that $\omega_1^{**} \geq \omega_2^{**}$. According to the assumption that $\omega_1^* - \omega_1^{**} \geq \omega_2^* - \omega_2^{**}$, we know that $\omega_1^* \geq \omega_2^*$. Therefore:

$$X + \frac{\Delta_2 - c}{q\Psi_1^{**}} \leq -X + \frac{\Delta_1 - c}{q\Psi_2^{**}}$$

or equivalently,

$$X \leq \frac{1}{2}\Big(\frac{|c-\Delta_1|}{q\Psi_2^{**}} - \frac{|c-\Delta_2|}{q\Psi_1^{**}}\Big) \tag{G.11}$$

On the other hand,

$$\omega_2^* - \omega_2^{**} = b - \frac{|c-\Delta_2|}{q\Psi_1^{**}} \cdot \frac{\beta(\mathcal{I}_1 - 1)l^*}{\mathcal{I}_m + \beta l^* \mathcal{I}_1} > 0,$$

which implies that for every $\varepsilon > 0$, there exists $\mathcal{I}^*$ such that whenever $\mathcal{I}_m > \mathcal{I}^*$,

$$\frac{|c-\Delta_2|}{q\Psi_1^{**}} \leq b\frac{1+\beta l^*}{\beta l^*} + \varepsilon.$$

Since $\beta \geq 1/2$ and the RHS is decreasing in $\beta$, we know that when $\mathcal{I}_m$ is sufficiently large,

$$X \leq \frac{b}{2} \cdot \frac{1 + l^*/3}{l^*/3}. \tag{G.12}$$

Given this uniform upper bound on $X$, we know that as $\omega_1^* \to -\infty$,

$$\frac{\omega_1^* - c + X}{\omega_1^{**} + b - c + X} \to 1.$$

The second part of (G.9) together with $\beta \geq 1/2$ implies that $\mathcal{I}_2$ is uniformly bounded from above as $\omega_1^* \to -\infty$, which contradicts the assumption that $\mathcal{I}$ is unbounded.

**Subcase 2: $\boldsymbol{\omega_1^* - \omega_1^{**} < \omega_2^* - \omega_2^{**}}$** Since $\beta \geq 1/2$, the distance between $\omega_1^*$ and $\omega_1^{**}$ is at most $b$ and

$$\frac{\omega_1^* - c + X}{\omega_1^{**} + b - c + X} = \frac{\beta l^* \mathcal{I}_2 + \mathcal{I}_m}{\beta l^* + \mathcal{I}_m},$$

if $\mathcal{I}_m$ is unbounded, then $\omega_1^* - c + X$ is bounded from below. That is, there exists $A \in \mathbb{R}_+$ such that

$$|\omega_1^* - c + X| = \frac{|c-\Delta_1|}{q\Psi_2^{**}} \leq A. \tag{G.13}$$

Since $\omega_1^* - \omega_1^{**} < \omega_2^* - \omega_2^{**}$, we know that when $\mathcal{I}_m$ is sufficiently large,

$$\frac{|c-\Delta_1|}{q\Psi_2^{**}} \cdot \frac{1-\beta}{1+(1-\beta)l^*} \leq \frac{|c-\Delta_2|}{q\Psi_1^{**}} \cdot \frac{\beta}{1+\beta l^*}. \tag{G.14}$$

Therefore,

$$\frac{|c-\Delta_2|}{q\Psi_1^{**}} \leq \frac{|c-\Delta_1|}{q\Psi_2^{**}} \cdot \frac{1-\beta}{\beta} \cdot (1+l^*) \leq A(1+l^*)\frac{1-\beta}{\beta}. \tag{G.15}$$

According to Lemma G.1, $\beta \to 1$ and $\omega_1^* \to -\infty$ are required when $\mathcal{I}_m \to \infty$. Therefore, $X \to \infty$ and

$$\frac{|c - \Delta_2|}{q\Psi_1^{**}} \to 0.$$

But according to the expression that

$$\omega_2^* = c + X + \frac{|c - \Delta_2|}{q\Psi_1^{**}},$$

we know that $\omega_2^*$ is strictly positive when $\mathcal{I}_m$ is sufficiently large. Therefore $\omega_2^{**} \geq \omega_2^* - b \geq -b$ and therefore, $\mathcal{I}_m \leq \mathcal{I}_2 \leq 1/\Phi(-b)$, leading to a contradiction.

### G.2.3 Case 3: $\Delta_1 \geq c$ and $\Delta_2 < c$

Define $X$ in the same way as in the previous subsection. If $X \leq 0$, then

$$\omega_1^{**} \geq -b + c$$

which implies that $\mathcal{I} \leq 1/\Phi(-b + c)$.

If $X > 0$, then

$$\frac{\omega_2^* - c - X}{\omega_2^{**} + b - c - X} \to 1$$

as $\omega_2^* \to -\infty$. Since

$$\frac{\omega_2^* - c - X}{\omega_2^{**} + b - c - X} = \frac{(1-\beta)l^*\mathcal{I}_1 + \mathcal{I}_m}{(1-\beta)l^* + \mathcal{I}_m},$$

we know that in order for $\mathcal{I}_m \to \infty$, we need $\omega_2^* \to -\infty$ and $\beta \to 1$. Therefore, it is without loss to consider situations in which

$$\beta \geq \overline{\beta} \equiv \max\{1 - 1/l^*, 1/2\}.$$

When $\beta \geq \overline{\beta}$, we know that $\mathcal{I}_m = \mathcal{I}_1 \leq \mathcal{I}_2$. Since $\omega_1^* - \omega_1^{**} \geq b > \omega_2^* - \omega_2^{**}$, we know that $\omega_2^{**} < \omega_1^{**}$, which further implies that $\omega_2^* < \omega_1^*$. This implies that

$$X + \frac{\Delta_2 - c}{q\Psi_1^{**}} \leq -X + \frac{\Delta_1 - c}{q\Psi_2^{**}}$$

which is equivalent to:

$$X \leq \frac{1}{2}\Big(\frac{|\Delta_1 - c|}{q\Psi_2^{**}} + \frac{|c - \Delta_2|}{q\Psi_1^{**}}\Big).$$

Since $\omega_2^* - \omega_2^{**} > 0$, we know that for every $\mathcal{I}_m$ above some threshold,

$$\frac{|c-\Delta_2|}{q\Psi_1^{**}} \leq b\frac{1+\widetilde{\beta}l^*}{\widetilde{\beta}l^*},$$

where $\widetilde{\beta} \equiv \overline{\beta}/2$. Therefore

$$\begin{aligned}
\omega_1^{**} &= -b+c-X+\frac{\Delta_1-c}{q\Psi_2^{**}}\cdot\frac{(1-\beta)l^*+\mathcal{I}_m}{\mathcal{I}_m+(1-\beta)l^*\mathcal{I}_2} \\
&\geq -b+c-\frac{|c-\Delta_2|}{2q\Psi_1^{**}}-\frac{|\Delta_1-c|}{2q\Psi_2^{**}}+\frac{|\Delta_1-c|}{q\Psi_2^{**}}\cdot\frac{(1-\beta)l^*+\mathcal{I}_m}{\mathcal{I}_m+(1-\beta)l^*\mathcal{I}_2} \\
&\geq -b+c-b\frac{1+\widetilde{\beta}l^*}{2\widetilde{\beta}l^*}+\frac{|\Delta_1-c|}{q\Psi_2^{**}}\Big(\frac{(1-\beta)l^*+\mathcal{I}_m}{\mathcal{I}_m+(1-\beta)l^*\mathcal{I}_2}-\frac{1}{2}\Big) \qquad \text{(G.16)}
\end{aligned}$$

The coefficient

$$\frac{(1-\beta)l^*+\mathcal{I}_m}{\mathcal{I}_m+(1-\beta)l^*\mathcal{I}_2}-\frac{1}{2}$$

is strictly positive when $\beta \geq \overline{\beta}$ and $\mathcal{I}_m$ is sufficiently large. Therefore (G.16) implies that

$$\omega_1^{**} \geq \overline{\omega}_1^{**} \equiv -b+c-b\frac{1+\widetilde{\beta}l^*}{2\widetilde{\beta}l^*} \qquad \text{(G.17)}$$

which further implies that

$$\mathcal{I}_m = \mathcal{I}_1 \leq \Phi(\overline{\omega}_1^{**})^{-1}.$$

# H Proof of Proposition 8

The proof consists of two parts, which studies the comparative statics on the reporting cutoffs (subsection H.1) and the informativeness of reports as well as the equilibrium probability of crime (subsection H.2), respectively.

## H.1 Reporting Cutoffs & Distance Between Cutoffs

In this subsection, we show that $\omega_k^* > \omega_n^*$. Suppose towards a contradiction that $\omega_k^* \leq \omega_n^*$, then according to (5.1), we have:

$$q_k\Big(\delta\Phi(\omega_k^{**})+(1-\delta)\alpha\Big)^{k-1} \leq q_n\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)^{n-1}. \qquad \text{(H.1)}$$

Therefore, $q_k Q_{0,k} \leq q_n Q_{0,n}$ which is equivalent to:

$$q_k\Big(\delta\Phi(\omega_k^{**})+(1-\delta)\alpha\Big)^{k-1}\Big(\Phi(\omega_n^*)-\Phi(\omega_n^{**})\Big) \leq q_n\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)^{n-1}\Big(\Phi(\omega_n^*)-\Phi(\omega_n^{**})\Big)$$

$$= q_k\Big(\delta\Phi(\omega_k^{**})+(1-\delta)\alpha\Big)^{k-1}\Big(\Phi(\omega_k^*)-\Phi(\omega_k^{**})\Big).$$

This implies that

$$\Phi(\omega_n^*)-\Phi(\omega_n^{**}) \leq \Phi(\omega_k^*)-\Phi(\omega_k^{**}). \tag{H.2}$$

Since $\omega_k^* \leq \omega_n^*$, (H.2) can only be true when

$$\omega_n^*-\omega_n^{**} \leq \omega_k^*-\omega_k^{**}, \tag{H.3}$$

which in turn implies that $\omega_k^{**} \leq \omega_n^{**}$ and therefore $q_k Q_{1,k} \leq q_n Q_{1,n}$. Computing the two sides of (H.3) by subtracting (5.2) from (5.1), we have:

$$\omega_n^*-\omega_n^{**} = b-\frac{c}{q_n}\frac{Q_{1,n}-Q_{0,n}}{Q_{1,n}Q_{0,n}} \text{ and } \omega_k^*-\omega_k^{**} = b-\frac{c}{q_k}\frac{Q_{1,k}-Q_{0,k}}{Q_{1,k}Q_{0,k}}.$$

Due to the previous conclusion that $q_k Q_{0,k} \leq q_n Q_{0,n}$ and $q_k Q_{1,k} \leq q_n Q_{1,n}$, (H.3) is true only when

$$q_n(Q_{1,n}-Q_{0,n}) \geq q_k(Q_{1,k}-Q_{0,k}). \tag{H.4}$$

Since

$$Q_{1,n}-Q_{0,n} = \frac{(n-1)l^*}{(n-1)l^*+n\mathcal{I}_n}\delta\Big(\Phi(\omega_n^*)-\Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)^{n-2}$$

and the term

$$\delta\Big(\Phi(\omega_n^*)-\Phi(\omega_n^{**})\Big)\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)^{n-2} = L^{-1}q_n^{-1}\frac{1}{\delta\Phi(\omega_n^{**})+(1-\delta)\alpha}$$

according to (5.6), we know that (H.4) is equivalent to:

$$\frac{(n-1)l^*}{(n-1)l^*\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)+n\Big(\delta\Phi(\omega_n^*)+(1-\delta)\alpha\Big)}$$

$$\geq \frac{(k-1)l^*}{(k-1)l^*\Big(\delta\Phi(\omega_k^{**})+(1-\delta)\alpha\Big)+k\Big(\delta\Phi(\omega_k^*)+(1-\delta)\alpha\Big)}$$

which in turn reduces to:

$$(n-1)(k-1)l^*\Big(\delta\Phi(\omega_k^{**})+(1-\delta)\alpha\Big)+(n-1)k\Big(\delta\Phi(\omega_k^{*})+(1-\delta)\alpha\Big)$$

$$\geq (n-1)(k-1)l^*\Big(\delta\Phi(\omega_n^{**})+(1-\delta)\alpha\Big)+(k-1)n\Big(\delta\Phi(\omega_n^{*})+(1-\delta)\alpha\Big)$$

The above inequality cannot be true as $\delta\Phi(\omega_k^{**})+(1-\delta)\alpha < \delta\Phi(\omega_n^{**})+(1-\delta)\alpha$, $\delta\Phi(\omega_k^{*})+(1-\delta)\alpha < \delta\Phi(\omega_n^{*})+(1-\delta)\alpha$ and moreover, as $k > n$, we know that $(n-1)k < (k-1)n$. This leads to a contradiction which shows that $\omega_k^* > \omega_n^*$ whenever $k > n$.

Notice that up until the last step, we did not use the fact that $k > n$. Given the previous conclusion that $\omega_k^* > \omega_n^*$ and repeat the same reasoning up until (H.3), we know that

$$\omega_n^* - \omega_n^{**} > \omega_k^* - \omega_k^{**}, \tag{H.5}$$

and this further implies that $\omega_k^{**} > \omega_n^{**}$.

## H.2 Informativeness & Probability of Crime

In this subsection, we establish the comparison between informativeness by showing that $\mathcal{I}_n > \mathcal{I}_k$, i.e. having more agents decreases the net informativeness of reports. Due to the one-to-one mapping between net informativeness and the probability of at least one assault taking place, this will also imply that $\widetilde{\pi}_k > \widetilde{\pi}_n$, i.e. the probability of crime increases.

Applying (5.1) and (5.2) to both $n$ and $k$, we obtain the following expression for the ratios:

$$\frac{\omega_n^* - c}{\omega_k^* - c} = \frac{q_k Q_{0,k}}{q_n Q_{0,n}} \quad \text{and} \quad \frac{\omega_n^{**} + b - c}{\omega_k^{**} + b - c} = \frac{q_k Q_{0,k}(\beta_k + (1-\beta_k)\mathcal{I}_k)}{q_n Q_{0,n}(\beta_n + (1-\beta_n)\mathcal{I}_n)}. \tag{H.6}$$

First, we show that

$$\frac{\omega_n^* - c}{\omega_k^* - c} > \frac{\omega_n^{**} + b - c}{\omega_k^{**} + b - c}. \tag{H.7}$$

Suppose towards a contradiction that the opposite of (H.7) is true, then

$$\frac{\omega_n^{**} + b - c - (\omega_n^* - c)}{\omega_k^{**} + b - c - (\omega_k^* - c)} \geq \frac{\omega_n^* - c}{\omega_k^* - c}. \tag{H.8}$$

The RHS of (H.8) is strictly greater than 1 as $0 > \omega_k^* > \omega_n^*$. The LHS of (H.8) being greater than 1

is equivalent to

$$b - (\omega_n^* - \omega_n^{**}) > b - (\omega_k^* - \omega_k^{**})$$

which contradicts the previous conclusion in (H.5). This establishes (H.7). This together with (H.6) imply that

$$\beta_k + (1 - \beta_k)\mathcal{I}_k < \beta_n + (1 - \beta_n)\mathcal{I}_n.$$

Plugging in the expressions of $\mathcal{I}_n$ and $\mathcal{I}_k$ in (5.5), we have:

$$\mathcal{I}_k\big(k + (k-1)l^*\big)\big(n\mathcal{I}_n + (n-1)l^*\big) < \mathcal{I}_n\big(n + (n-1)l^*\big)\big(k\mathcal{I}_k + (k-1)l^*\big).$$

Let $\Delta \equiv \mathcal{I}_k - \mathcal{I}_n$, the above inequality reduces to:

$$(k-n)\mathcal{I}_n(\mathcal{I}_n + \Delta - 1) < k\Delta - \Big(l^*(k-1)(n-1) + nk\Big)\Delta.$$

Suppose towards a contradiction that $\Delta \geq 0$, then the LHS is strictly positive since $\mathcal{I} > 1$ and $k > n$. The RHS is negative as $l^*(k-1)(n-1) + nk > k$. This leads to a contradiction which implies that $\Delta < 0$ and therefore, $\mathcal{I}_n > \mathcal{I}_k$.

# I Mitigating Punishment to the Convicted

In Appendix I.1, we prove Proposition 4 by constructing equilibria in which the principal's decisions are strategic complements when $L$ belongs to an open interval. In Appendix I.2, we show there exist values of $L$ such that focusing on equilibria where the principal's decisions are strategic complements is without loss of generality. Namely, the value of (3.7) is non-positive in all equilibria.

## I.1 Proof of Proposition 4

Consider equilibria where $q(1,1) = 1$, $q(1,0) = q(0,1) = q$ and $q(0,0) = 0$ with $q \geq 1/2$. According to Lemma D.1, the value of (3.7) is strictly negative, which implies that the principal's decisions to abuse agents are strategic complements. Therefore in equilibrium, the principal either chooses $\theta_1 = \theta_2 = 1$ or chooses $\theta_1 = \theta_2 = 0$ but he will never abuse only one agent. Agent $i$ will report if $\omega_i$ is below

$$\omega^* \equiv -\frac{c(1-q)(1-\Psi^*)}{q + \Psi^*(1-2q)} \tag{I.1}$$

and $\theta_i = 0$, or if $\omega_i$ is below

$$\omega^{**} \equiv -b - \frac{c(1-q)(1-\Psi^{**})}{q + \Psi^{**}(1-2q)}. \tag{I.2}$$

when $\theta_i = 1$. The principal's indifference condition is given by:

$$2/L = (\Psi^* - \Psi^{**})\Big((1-2q)(\Psi^* + \Psi^{**}) + 2q\Big), \tag{I.3}$$

where

$$\Psi^* \equiv \delta\Phi(\omega^*) + (1-\delta) \quad \text{and} \quad \Psi^{**} \equiv \delta\Phi(\omega^{**}) + (1-\delta).$$

Moreover, the equilibrium probability of crime, denoted by $\widetilde{\pi}_m$, is pinned down by:

$$\frac{\Psi^*(1-\Psi^*)}{\Psi^{**}(1-\Psi^{**})} = \frac{\pi^*}{1-\pi^*} \Big/ \frac{\widetilde{\pi}_m}{1-\widetilde{\pi}_m}, \tag{I.4}$$

where

$$\mathcal{I} \equiv \frac{\Psi^*(1-\Psi^*)}{\Psi^{**}(1-\Psi^{**})}$$

measures the aggregate informativeness of reports. This is because in such equilibria, one report is sufficient to convict the principal, and therefore, the evaluator is indifferent between $s = 1$ and $s = 0$ when there is exactly one report.

Comparing (I.1) to (I.2), we know that $\omega^* - \omega^{**} > b$. Rewrite (I.1) and (I.2) as:

$$\frac{\omega^*}{c} = -\frac{1-\Psi^*}{\Psi^* + (1-\Psi^*)\frac{q}{1-q}} \tag{I.5}$$

and

$$\frac{\omega^{**} + b}{c} = -\frac{1-\Psi^{**}}{\Psi^{**} + (1-\Psi^{**})\frac{q}{1-q}}, \tag{I.6}$$

notice that the RHS is bounded within $[-\frac{1-q}{q}, 0]$ and is continuous with respect to $q$. For every $c > 0$, both (I.4) and (I.5) admit a unique solution. Moreover, for every $\overline{q} < 1$ and $A \in \mathbb{R}_+$, there exists $\overline{c} > 0$ such that for every $c > \overline{c}$ and $q \in [1/2, \overline{q}]$, the solution satisfies $|\omega^*| > A$. Since $\omega^* - \omega^{**} > b$, we know that the aggregate reporting informativeness in the $\delta \to 1$ limit also goes to infinity, that is,

$$\lim_{c\to\infty} \lim_{\delta\to 1} \frac{\delta\Phi(\omega^*) + (1-\delta)\alpha}{\delta\Phi(\omega^{**}) + (1-\delta)\alpha} = \infty.$$

Therefore, when $c$ is large enough and by setting $L$ to be in an open set consisting of the values of

(I.3) when $q \in [1/2, \overline{q}]$, the informativeness ratio $\mathcal{I}$ goes to infinity and the equilibrium probability of crime converges to 0.

## I.2 Complementarity in Principal's Actions

For every $(c, \delta)$, let $(\omega_0^*, \omega_0^{**})$ be the unique solution to

$$\frac{\omega_0^*}{c} = \Psi_0^* - 1 \quad \text{and} \quad \frac{\omega_0^{**} + b}{c} = \Psi_0^{**} - 1,$$

where $\Psi_0^* \equiv \delta\Phi(\omega_0^*) + (1-\delta)$ and $\Psi_0^{**} \equiv \delta\Phi(\omega_0^{**}) + (1-\delta)$. Let

$$L_0 \equiv \frac{2}{\Psi_0^* - \Psi_0^{**}}. \tag{I.7}$$

According to the analysis in subsection I.1, we know that $(\omega_0^*, \omega_0^{**})$ are the agents' reporting cutoffs when $L_0$ is the punishment to the convicted and the conviction probabilities are given by $q(1,1) = 1$, $q(0,1) = q(1,0) = 1/2$ and $q(0,0) = 0$. We show the following proposition:

**Proposition 10.** *There exists an open neighborhood of $L_0$ such that for every $L$ belonging to this neighborhood, the value of (3.7) is non-positive in every equilibrium.*

The proof of Proposition 10, which can be found in subsections I.3 and I.4, shows that whenever there exists a monotone-responsive equilibrium in which the principal's choices of $\theta_1$ and $\theta_2$ are strategic substitutes, then $L$ needs to be strictly larger than $L_0$. This together with the existence of equilibrium implies that the principal's decisions are strategic complements in all equilibria. The proof considers two cases separately, depending on the conviction probabilities.

## I.3 Case 1: q(1, 0) or q(0, 1) is Strictly Positive

Suppose towards a contradiction that there exists a monotone-responsive equilibrium in which either $q(0,1)$ or $q(1,0)$ is strictly positive or both. For notation simplicity, let $q_1 \equiv q(1,0)$ and $q_2 \equiv q(0,1)$. For $i \in \{1,2\}$, let $p_i$ be the probability with which $\theta_i = 0$. Let $\omega_i^*$ and $\omega_i^{**}$ be the agent $i$'s reporting cutoffs, with expressions given by:

$$\omega_i^* = -c\frac{(1-\Psi_j^{**})(1-q_i)}{q_i + \Psi_j^{**}(1-q_1-q_2)} \tag{I.8}$$

and

$$\omega_i^{**} = -b - c\frac{(1-X_j)(1-q_i)}{q_i + X_j(1-q_1-q_2)} \tag{I.9}$$

where $j \equiv 3-i$ and

$$X_i \equiv \frac{1-p_1-p_2}{1-p_i}\Psi_i^{**} + \frac{p_j}{1-p_i}\Psi_i^{*}.$$

According to the conclusions in Appendix D, it is without loss to focus on equilibria in which the principal's cost of abusing the two agents are equal. This leads to the indifference condition:

$$L = \frac{1}{(\Psi_1^* - \Psi_1^{**})\Big(\Psi_2^{**}(1-q_1-q_2)+q_1\Big)} = \frac{1}{(\Psi_2^* - \Psi_2^{**})\Big(\Psi_1^{**}(1-q_1-q_2)+q_2\Big)}. \tag{I.10}$$

Without loss of generality, we assume $q_1 \leq q_2$. Since $q_1 + q_2 \leq 1$, we know that

$$L = \frac{1}{(\Psi_1^* - \Psi_1^{**})\Big(\Psi_2^{**}(1-q_1-q_2)+q_1\Big)} \geq \frac{2}{\Psi_1^* - \Psi_1^{**}}.$$

In what follows, we will show that

$$\frac{2}{\Psi_1^* - \Psi_1^{**}} > L_0$$

or equivalently,

$$\Psi_0^* - \Psi_0^{**} > \Psi_1^* - \Psi_1^{**}. \tag{I.11}$$

According to the expression of $\omega_1^*$ in (I.8), we know that:

$$\omega_1^* = -c\frac{(1-\Psi_2^{**})(1-q)}{q + \Psi_2^{**}(1-2q)} \leq -c(1-\Psi_1^{**}) \leq c(\Psi_1^* - 1).$$

Therefore, $\omega_1^*$ is strictly below the unique solution of the equation:

$$\omega = c\Big(\underbrace{\delta\Phi(\omega) + (1-\delta)\alpha}_{\equiv\Psi(\omega)} - 1\Big)$$

which equals to $\omega_0^*$. Furthermore, since $\omega_0^* - \omega_0^{**} > b > \omega_1^* - \omega_1^{**}$, one can obtain (I.11).

## I.4 Case 2: $\mathbf{q(1,0) = q(0,1) = 0}$

Given that we have already shown in Appendix C that symmetric equilibria is without loss of generality when $q(0,0) = q(1,0) = q(0,1) = 0$, we will be focusing on symmetric equilibria. Let $q \equiv q(1,1) \in$

$(0,1]$. Let $\omega_1^*$ and $\omega_1^{**}$ be the agents' reporting cutoffs, which are the same across agents. The expressions for the cutoffs are the same as those for $\omega_m^*$ and $\omega_m^{**}$, which are given by (3.8) and (3.9), respectively. The principal's indifference condition is given by:

$$\frac{1}{L} = \Psi_1^{**}(\Psi_1^* - \Psi_1^{**}).$$

To show $L > L_0$, one only needs to show that:

$$\Psi_0^* - \Psi_0^{**} > \Psi_1^{**}(\Psi_1^* - \Psi_1^{**}). \tag{I.12}$$

According to (3.8), we have:

$$\omega_1^* = c - \frac{c}{q\Psi_1^*} \leq c - \frac{c}{\Psi_1^*} \leq c(\Psi_1^* - 1).$$

Similar to the previous case, we know that $\omega_1^*$ is strictly below $\omega_0^*$. Since $\omega_0^* - \omega_0^{**} > b > \omega_1^* - \omega_1^{**}$, we know that

$$\Psi_0^* - \Psi_0^{**} > \Psi_1^* - \Psi_1^{**}. \tag{I.13}$$

This in turn implies (I.12).

# J Proof of Proposition 9

We start from listing the sufficient conditions for equilibria in which $q(0,0) = q(1,0) = q(0,1) = 0$ and $q(1,1) \in (0,1)$. Since the posterior attaches to $\theta_1\theta_2 = 0$ reaches $\pi^*$ after observing two reports, the relationship between the equilibrium probability of crime $\widetilde{\pi}$ and the informativeness of reports $\mathcal{I}$ is given by (5.11), which according to (5.10) can be rewritten as:

$$\widetilde{\pi} = \frac{l^* + \epsilon R - \epsilon R^2}{R + l^*}. \tag{J.1}$$

The expressions for the cutoffs are given by:

$$\omega^* = -c\frac{1 - qQ_0}{qQ_0} \quad \text{and} \quad \omega^{**} = -b - c\frac{1 - qQ_1}{qQ_1} \tag{J.2}$$

where $Q_i = \beta_i \Psi^* + (1-\beta_i)\Psi^{**} = (\beta_i + \frac{1-\beta_i}{R})\Psi^*$ for $i \in \{0,1\}$ with

$$\beta_0 \equiv \frac{2\epsilon(R+l^*)}{2R + l^* + \epsilon(l^* + R^2)} \tag{J.3}$$

and

$$\beta_1 \equiv \frac{l^* - \epsilon(R^2 + l^*)}{l^* + 2R - \epsilon(2R - R^2 + l^*)}. \tag{J.4}$$

Rewrite (J.2) by plugging into the definition of $R$, we have:

$$-\frac{\omega^* - c}{c} = \frac{1}{\xi_0(R,\epsilon)\Psi^* q} \text{ and } -\frac{\omega^{**} + b - c}{c} = \frac{1}{\xi_1(R,\epsilon)\Psi^* q} \tag{J.5}$$

where

$$\xi_0(R,\epsilon) = \beta_0 + (1-\beta_0)\frac{1}{R} \text{ and } \xi_1(R,\epsilon) = \beta_1 + (1-\beta_1)\frac{1}{R}. \tag{J.6}$$

Notice that both $\xi_0$ and $\xi_1$ are continuous functions with values no more than 1. The values of both functions equal to 1 when $R = 1$ and $\epsilon = 0$. More importantly, the values of $\xi_0$ and $\xi_1$ are fixed once we fix $R$ and $\epsilon$.

Next, consider the following mapping $f \equiv (f_1, f_2) : [0,1] \times [0,1] \to [0,1] \times [0,1]$:

$$f_1(\Psi^{**}, q) \equiv \Psi\Big(-b + c - \frac{c}{q\xi_1 R \Psi^{**}}\Big) \tag{J.7}$$

$$f_2(\Psi^{**}, q) \equiv \min\Big\{1, \frac{c}{\xi_0 \Psi^*(c - \omega^*)}\Big\}, \tag{J.8}$$

where for given $\Psi^{**}$, $\Psi^*$ (and hence $\omega^*$) is pinned down via $\Psi^*/\Psi^{**} = R$. Since $f$ is continuous, the Brouwer's fixed point theorem implies the existence of a fixed point. In what follows, we show that $q = 1$ cannot be part of any fixed point of $f$ when $\epsilon$ is close to 0 and $R$ is close to 1. For this purpose, we need to show that:

$$\frac{c}{\xi_0 \Psi^*(c - \omega^*)} < 1 \tag{J.9}$$

for every $\Psi^{**}$ solving the equation:

$$-\frac{\omega^{**} + b - c}{c} = \frac{1}{R\xi_1 \Psi^{**}}. \tag{J.10}$$

To see this, first, (J.10) admits at least one solution as $\Psi^{**} \geq (1-\delta)\alpha$. Second, (J.9) is equivalent to:

$$\frac{\xi_1}{\xi_0}\frac{|\omega^{**}|+|c|-b}{|\omega^{*}|+|c|} < 1. \tag{J.11}$$

The above inequality holds as first, $\frac{\xi_1}{\xi_0} \to 1$ as $R \to 1$ and $\epsilon \to 0$; and second, $|\omega^*| > |\omega^{**}| - b$ whenever $\beta_0 < \beta_1$, the latter is true when $\epsilon$ is small enough.

Since every fixed point features $q \in (0,1)$, the fixed point of $f$ is also the level of $(\Psi^*, \Psi^{**}, q)$ in one of the monotone-responsive equilibrium. One can then pin down $L_h$ via:

$$\frac{1}{\delta L_h} = \delta(\Phi(\omega^*) - \Phi(\omega^{**}))(\delta\Phi(\omega^{**}) + 1 - \delta). \tag{J.12}$$

Let $\widetilde{\pi}$ be given by (J.1). We know that $(\omega^*, \omega^{**}, q, \widetilde{\pi})$ is an equilibrium under $(L_h, L_l, \epsilon)$.

# K Alternative Commitment Types

In this Appendix, we examine the robustness of our findings against alternative specifications of mechanical types. In particular, the mechanical types' reports can be informative about the principal's innocence. We show that when commitment types are rare and the principal's loss from being convicted is sufficiently large, the informativeness of reports vanishes to 1 and the probability of crime converges to $\pi^*$ as in the baseline model. This confirms the robustness of our findings. For illustration purposes, we will again focus on the comparison between one and two agents.

## K.1 Model & Result

Consider the following modification of the baseline model. With probability $\delta$, the agent is a strategic type maximizes payoff function given by (2.5). With probability $1-\delta$, the agent is a mechanical type whose reporting cutoff is $\overline{\omega}$ when $\theta_i = 0$ and $\underline{\omega}$ when $\theta_i = 1$. We assume that both $\overline{\omega}$ and $\underline{\omega}$ are finite with $\overline{\omega} \geq \underline{\omega}$, that is, the mechanical type's report could be informative about $\theta$.[23] An example of such mechanical types are agents who are immune to retaliation, that is, they maximize:

$$(\omega_i + b\theta_i)s. \tag{K.1}$$

[23]Our analysis also applies when mechanical types are using arbitrary strategies contingent on $(\theta_i, \omega_i)$, as long as conditional on each realization of $\theta_i$, the probability with which the mechanical type reports is interior, and moreover, this conditional probability is weakly higher when $\theta_i = 0$ compared to $\theta_i = 1$.

In this example, $\overline{\omega} = 0$ and $\underline{\omega} = -b$.

When there is only one agent, his reporting cutoffs $\omega_s^*$ and $\omega_s^{**}$ are given by (3.1) and (3.2). The probability with which the principal is convicted after one report is $q_s$, with $(q_s, \omega_s^*, \omega_s^{**})$ satisfying:

$$q_s\Big(\delta(\Phi(\omega_s^*) - \Phi(\omega_s^{**})) + (1-\delta)(\Phi(\overline{\omega}) - \Phi(\underline{\omega}))\Big) = 1/L. \tag{K.2}$$

One can show that when $\delta \to 1$ and $L$ is larger than some cutoff $L(\delta)$, the informativeness of report:

$$\mathcal{I}_s \equiv \frac{\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\overline{\omega})}{\delta\Phi(\omega_s^{**}) + (1-\delta)\Phi(\underline{\omega})}$$

converges to $\infty$, namely, the agent's report becomes arbitrarily informative in the limit.

In the two-agent case, for every $i \in \{1,2\}$, agent $i$'s probability of filing a report is $\Psi^* \equiv \delta\Phi(\omega_m^*) + (1-\delta)\Phi(\overline{\omega})$ conditional on $\theta_i = 0$; his probability of filing a report is $\Psi^{**} \equiv \delta\Phi(\omega_m^{**}) + (1-\delta)\Phi(\underline{\omega})$ conditional on $\theta_i = 1$. The strategic agent's reporting cutoffs are given by:

$$\omega_m^* \equiv c - \frac{c}{q_m\Psi^{**}} \quad \text{and} \quad \omega_m^{**} \equiv -b + c - \frac{c}{q_m\Big(\beta\Psi^{**} + (1-\beta)\Psi^*\Big)}. \tag{K.3}$$

Let $\mathcal{I}_m \equiv \Psi^*/\Psi^{**}$. When $L$ is large enough, the conviction probabilities in every monotone-responsive equilibrium satisfies $q(0,0) = q(0,1) = q(1,0) = 0$ and $q(1,1) \in (0,1)$. Therefore, the expressions for $\beta$ and $1-\beta$ remain the same as in (3.15). The distance between the two cutoffs is given by:

$$\omega_m^* - \omega_m^{**} = b - \frac{c}{q_m}\frac{(1-\beta)(\mathcal{I}_m - 1)}{\Psi^{**}(\beta + (1-\beta)\mathcal{I}_m)} = b - \frac{c}{q_m\Psi^{**}}\frac{l^*}{2+l^*}\frac{\mathcal{I}_m - 1}{\mathcal{I}_m}. \tag{K.4}$$

One can then show that $\omega_m^* - \omega_m^{**} < b$. This is because for $\omega_m^* - \omega_m^{**}$ to be greater or equal to $b$, we need $\mathcal{I}_m \leq 1$ which can only be true when $\omega_m^* \leq \omega_m^{**}$, leading to a contradiction.

Different from the baseline model, when mechanical types' reports are informative about the principal's innocence, the strategic types' coordination motives can *reverse* the ordering between the two cutoffs. That is to say, $\omega_m^*$ can be strictly smaller than $\omega_m^{**}$ in equilibrium. As a result, the argument that shows $\mathcal{I}_m \to 1$ when $\omega_m^* \to -\infty$ in Lemma 3.3 no longer applies. This is because in principle, $\omega_m^*$ could be much smaller than $\omega_m^{**}$, so the ratio between the absolute values in (3.17) can converge to something strictly above 1 as $\omega_m^*$ and $\omega_m^{**}$ converge to $-\infty$. To circumvent this problem, we take an alternative approach based on the comparison between $\omega_m^*$ and $\omega_s^*$. The result in this subsection is the following proposition:

**Proposition 11.** *There exists $\overline{L} : \mathbb{R}_+ \times (0,1) \to \mathbb{R}_+$ such that when $L > \overline{L}(c,\delta)$, an equilibrium exists. Compared to the single-agent benchmark, $q_m > q_s$, $\omega_m^* > \omega_s^*$ and $\omega_m^{**} > \omega_s^{**}$.*
*Moreover, as $\delta \to 1$ and $L \to \infty$ with the relative speed of convergence satisfying $L \geq \overline{L}(c,\delta)$, we have $\omega_m^*, \omega_m^{**} \to -\infty$, $\mathcal{I}_m \to 1$ and $\widetilde{\pi}_m \to \pi^*$.*

The proof is in the next subsection that treats two cases separately. Intuitively, in the *regular case* where $\omega_m^* \geq \omega_m^{**}$, one can still apply the ratio condition (3.17) to show that as $\omega_m^* \to -\infty$, the LHS converges to 1 which implies that $\mathcal{I}_m \to 1$. In the *irregular case* where $\omega_m^* < \omega_m^{**}$, the distance between $|\omega_m^* - c|$ and $|\omega_m^{**} + b - c|$ can be strictly larger than $b$ and can explode as $\omega_m^* \to -\infty$. However, since $\omega_m^* > \omega_s^*$ and the informativeness in the single-agent benchmark grows without bound as $L \to \infty$, it places an upper bound on the informativeness of reports in the two-agent scenario. Since informativeness is entirely contributed by the mechanical types in the irregular case, the value of the aforementioned upper bound will converge to 1 as $\mathcal{I}_s \to \infty$. Summing up the two cases together, we know that the agents' reports are arbitrarily uninformative in the limit even when the mechanical types' reports are informative.

## K.2 Proof of Proposition 11

We start from establishing the comparisons between the single-agent benchmark and the two-agent scenario when mechanical types' reports can be informative about $\theta$, captured by the two exogenous reporting cutoffs $\overline{\omega}$ and $\underline{\omega}$ with $\overline{\omega} \geq \underline{\omega}$.

Suppose towards a contradiction that $\omega_m^* \leq \omega_s^*$, the expressions for these cutoffs imply:

$$q_m\Big(\delta\Phi(\omega_m^{**}) + (1-\delta)\Phi(\underline{\omega})\Big) \leq q_s.$$

Therefore,

$$\begin{aligned}
&q_m\Psi^{**}\Big(\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\overline{\omega}) - \delta\Phi(\omega_s^{**}) - (1-\delta)\Phi(\underline{\omega})\Big) \\
\leq\ &q_s\Big(\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\overline{\omega}) - \delta\Phi(\omega_s^{**}) - (1-\delta)\Phi(\underline{\omega})\Big) = 1/L \\
=\ &q_m\Psi^{**}\Big(\delta\Phi(\omega_m^*) + (1-\delta)\Phi(\overline{\omega}) - \delta\Phi(\omega_m^{**}) - (1-\delta)\Phi(\underline{\omega})\Big)
\end{aligned}$$

or equivalently

$$\Phi(\omega_m^*) - \Phi(\omega_m^{**}) \geq \Phi(\omega_s^*) - \Phi(\omega_s^{**}). \tag{K.5}$$

On the other hand, since $\omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**}$ and $\omega_m^* < \omega_s^*$, we have:

$$\Phi(\omega_m^*) - \Phi(\omega_m^{**}) < \Phi(\omega_s^*) - \Phi(\omega_s^{**}). \tag{K.6}$$

which contradicts (K.5). This contradiction implies that $\omega_m^* > \omega_s^*$. Since $\omega_m^* - \omega_m^{**} < b = \omega_s^* - \omega_s^{**}$, we know that $\omega_m^{**} > \omega_s^{**}$. Moreover, $\omega_m^* > \omega_s^*$ implies that $q_m \Psi^{**} > q_s$. That is $1 \geq \Psi^{**} > q_s/q_m$, which implies that $q_m > q_s$.

Next, we establish the informativeness of the agents' reports when there are two agents and $\delta$ and $L$ being sufficiently large. First, for every $X \in \mathbb{R}_+$, there exists $\overline{\delta} \in (0,1)$ and $L^* : (\overline{\delta}, 1) \to \mathbb{R}_+$ such that when $\delta > \overline{\delta}$ and $L > L^*(\delta)$, the resulting cutoffs in the single-agent case satisfies:

$$\frac{\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\overline{\omega})}{\delta\Phi(\omega_s^* - b) + (1-\delta)\Phi(\underline{\omega})} > X, \tag{K.7}$$

which implies that

$$\delta\Phi(\omega_s^*) > (1-\delta)\Big(X\Phi(\underline{\omega}) - \Phi(\overline{\omega})\Big). \tag{K.8}$$

Next, we establish an upper bound on the informativeness of reports in the limit of the two-agent case. Consider a two-agent economy under parameter values $(L, c, \delta)$ such that $L \geq \overline{L}(c, \delta)$, i.e. monotone-responsive equilibria exist In equilibria where $\omega_m^* \geq \omega_m^{**}$, the expressions for $\omega_m^*$ and $\omega_m^{**}$ imply that:

$$\frac{|\omega_m^* - c|}{|\omega_m^{**} - c + b|} = \frac{(l^* + 2)\mathcal{I}_m}{l^* + 2\mathcal{I}_m}. \tag{K.9}$$

The LHS converges to 1 as $\omega_m^* \to -\infty$ so the RHS also converges to 1, which implies that $\mathcal{I}_m \to 1$.

In equilibria where $\omega_m^* < \omega_m^{**}$, since $\omega_s^* < \omega_m^*$, we have:

$$\mathcal{I}_m \leq \frac{\delta\Phi(\omega_m^*) + (1-\delta)\Phi(\overline{\omega})}{\delta\Phi(\omega_m^*) + (1-\delta)\Phi(\underline{\omega})} \underbrace{\leq}_{\text{since } \mathcal{I}_m > 1 \text{ and } \omega_m^* > \omega_s^*} \frac{\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\overline{\omega})}{\delta\Phi(\omega_s^*) + (1-\delta)\Phi(\underline{\omega})}$$

$$\leq \frac{(1-\delta)\Big(X\Phi(\underline{\omega}) - \Phi(\overline{\omega})\Big) + (1-\delta)\Phi(\overline{\omega})}{(1-\delta)\Big(X\Phi(\underline{\omega}) - \Phi(\overline{\omega})\Big) + (1-\delta)\Phi(\underline{\omega})} = \frac{X\Phi(\underline{\omega})}{X\Phi(\underline{\omega}) - \Phi(\overline{\omega}) + \Phi(\underline{\omega})} \tag{K.10}$$

which also converges to 1 as $X \to \infty$.

To summarize, since $\omega_m^* \to -\infty$ and $X \to \infty$ as $\delta \to 1$ and $L \to \infty$, we know that the informativeness ratio $\mathcal{I}_m$ converges to 1 no matter whether $\omega_m^* \geq \omega_m^{**}$ or $\omega_m^* < \omega_m^{**}$.

## References

[1] Ali, Nageeb, Maximilian Mihm and Lucas Siga (2018) "Adverse Selection in Distributive Politics," Working Paper.

[2] Austen-Smith, David and Jeffrey Banks (1996) "Information Aggregation, Rationality, and the Condorcet Jury Theorem," *The American Political Science Review*, 90(1), 34-45.

[3] Baliga, Sandeep, Ethan Bueno de Mesquita and Alexander Wolitzky (2018) "Deterrence with Imperfect Attribution," Working Paper.

[4] Baliga, Sandeep and Tomas Sjöström (2004) "Arms Races and Negotiations," *Review of Economic Studies*, 71(2), 351-369.

[5] Banerjee, Abhijit (1992) "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), 797-817.

[6] Bhattacharya, Sourav (2013) "Preference Monotonicity and Information Aggregation in Elections," *Econometrica*, 81(3), 1229-1247.

[7] Bikhchandani, Sushil, David Hirshleifer, Ivo Welch (1992) "A Theory of Fads , Fashion , Custom , and Cultural Change as Informational Cascades," *Journal of Political Economy*, 100(5), 992-1026.

[8] Carlson, Hans and Eric Van Damme (1993) "Global Games and Equilibrium Selection," *Econometrica*, 61(5), 989-1018.

[9] Chassang, Sylvain and Gerard Padró i Miquel (2010) "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, 125(4), 1821-1858.

[10] Chassang, Sylvain and Gerard Padró i Miquel (2018) "Corruption, Intimidation and Whistle-Blowing: A Theory of Inference from Unverifiable Reports," NBER working paper.

[11] Crémer, Jacques and Richard McLean (1985) "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist when Demands are Interdependent," *Econometrica*, 53(2), 345-361.

[12] Crémer, Jacques and Richard McLean (1988) "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56(6), 1247-1257.

[13] Dresher, Melvin (1962) "A Sampling Inspection Problem in Arms Control Agreements – A Game-Theoretic Analysis," Memorandum, No. RM-2972-ARPA, The RAND Corporation, Santa Monica, California.

[14] Feddersen, Timothy and Wolfgang Pesendorfer (1996) "The Swing Voter's Curse," *American Economic Review*, 86(3), 408-424.

[15] Feddersen, Timothy and Wolfgang Pesendorfer (1997) "Voting Behavior and Information Aggregation in Elections with Private Information," *Econometrica*, 65(5), 1029-1058.

[16] Fedderson, Timothy and Wolfgang Pesendorfer (1998) "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting," *The American Political Science Review*, 92(1), 23-35.

[17] Fudenberg, Drew and David Levine (1995) "The Theory of Learning in Games," MIT Press.

[18] Lee, Frances Xu and Wing Suen (2018) "Credibility of Crime Allegations," Working Paper.

[19] Morgan, John and Phillip Stocken (2008) "Information Aggregation in Polls," *American Economic Review*, 98(3), 864-896.

[20] Morris, Stephen and Hyun Song Shin (1998) "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 88(3), 587-597.

[21] Ottaviani, Marco and Peter Norman Sørensen (2000) "Herd Behavior and Investment: Comment," *American Economic Review*, 90(3), 695-704.

[22] Scharfstein, David and Jeremy Stein (1990) "Herd Behavior and Investment," *Amercian Economic Review*, 80(3), 465-479.

[23] Silva, Francesco (2018) "If We Confess Our Sins," Working Paper.

[24] Smith, Lones and Peter Norman Sørensen (2000) "Pathological Outcomes of Observational Learning," *Econometrica*, 68(2), 371-398.

[25] Strulovici, Bruno (2018) "Can Society Learn from Anethical Agents? A Theory of Mediated Learning with Information Attrition," Working Paper.