Crime, Intimidation, and Whistleblowing:

A Theory of Inference from Unverifiable Reports*

Sylvain Chassang — New York University

Gerard Padró i Miquel† — Yale University

November 15, 2018.

Abstract

We consider a game between a principal, an agent, and a monitor in which the principal would like to rely on messages by the monitor (the potential whistleblower) to target intervention against a misbehaving agent. The difficulty is that the agent can credibly threaten to retaliate against the monitor in the event of an intervention. In this setting, intervention policies that are responsive to the monitor's message provide informative signals to the agent, which can be used to target threats efficiently. Principals that are too responsive to information shut down communication channels. Successful intervention policies must therefore garble the information provided by monitors and cannot be fully responsive. We show that policy evaluation on the basis of non-verifiable whistleblower messages is feasible under arbitrary incomplete information provided policy design takes into account that messages are endogenous.

KEYWORDS: crime, intimidation, whistleblowing, plausible deniability, inference, policy evaluation, structural experiment design.

*We are grateful to Johannes Hörner for a very helpful discussion. We are indebted to Nageeb Ali, Abhijit Banerjee, Michael Callen, Yeon Koo Che, Hans Christensen, Ray Fisman, Matt Gentzkow, Bob Gibbons, Navin Kartik, David Martimort, Marco Ottaviani, Andrea Prat, Jesse Shapiro, as well as seminar audiences at the 2013 Winter Meeting of the Econometric Society, the 2015 Minnesota-Chicago Accounting Theory Conference, Berkeley, Bocconi, Columbia, Essex, Hebrew University, the Institute for Advanced Study, MIT, MIT Sloan, the Nemmers Prize Conference, Norwegian Business School, NYU, NYU IO day, the Paris School of Economics, Pompeu Fabra, ThReD, Yale, the University of Chicago, UCSD, and the UCSD workshop on Cellular Technology, Security and Governance for helpful conversations. Chassang gratefully acknowledges the hospitality of the University of Chicago Booth School of Business, as well as support from the Alfred P. Sloan Foundation and the National Science Foundation under grant SES-1156154. Padró i Miquel acknowledges financial support from the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Starting Grant Agreement no. 283837.

†Chassang: chassang@nyu.edu, Padró i Miquel: gerard.padro@yale.edu.

1 Introduction

Organizations and regulatory agencies often attempt to protect informants and whistleblowers to improve information transmission.[1] Anonymity guarantees are widely regarded as one of the primary means to achieve this goal: the 2002 Sarbanes-Oxley act, for instance, requires public companies to establish anonymous reporting channels. However, work by Kaplan et al. (2007, 2009) shows that greater anonymity guarantees seem to have little effect on information flows in practice.[2] This can be explained by the fact that in many cases the set of people informed about misbehavior is small, so that formal anonymity offers little actual protection. Police officers on patrol are a particularly salient example in which anonymity becomes meaningless: a pattern of misbehavior by one officer is only observed by his or her patrol partner.[3] In such cases, whistleblowing is easily deterred with explicit or implicit threats of retaliation from misbehaving individuals. The primary objective of this paper is to better understand the effectiveness of intervention policies when potential whistleblowers are subjected to intimidation.

We formalize the problem using a principal-agent-monitor framework in which the principal relies on messages from a single informed monitor to target intervention against a potentially criminal agent.[4] The difficulty is that the agent can credibly threaten to retaliate against the whistleblower as a function of available observables — including the principal's intervention behavior. Our modeling approach emphasizes three issues that are important in practical applications. First, we take seriously the idea that misbehaving agents can un-

[1] For a review of the whistleblowing literature across social sciences, see Near and Miceli (1995).

[2] Kaplan and Schultz (2007) argue that anonymous reporting channels fail to increase intention-to-report rates relative to non-anonymous ones. Similarly, in Kaplan et al. (2009) external hotlines with stronger safeguards do not elicit a higher propensity to report than internal hotlines with weaker safeguards. In a special report on Wells Fargo, the Financial Times describes multiple situations of retaliation against internal whistleblowers in the period between 2005 and 2015 (accessible at `https://ig.ft.com/special-reports/whistleblowers/`). The Sarbanes-Oxley act clearly failed to protect these workers.

[3] Other examples include judges and courtroom officials, fraudulent firms and their external accountants, as well as bullying and harassment in small teams.

[4] Throughout the paper we refer to the misbehaving agent as "criminal" and to the misbehavior as "crime." This is shorthand for any decision that the principal finds undesirable or harmful. Following convention, we refer to the principal and monitor as she, and to the agent as he.

dermine the effectiveness of institutions by side-contracting with parties able to inform on them. In our model, this side-contracting takes the form of contingent retaliation.

Second, departing from much of the literature on collusion, we do not assume that messages are verifiable, reflecting the fact that hard measures of corruption, fraud, and crime are often difficult to obtain, and unreliable.[5] The outcomes of policies designed to improve the agent's behavior need not be directly measurable, and policies may have to be evaluated using soft, non-verifiable information. A consequence of non-verifiability is that malicious monitors may try to convey false information about honest agents. This makes policy design particularly challenging.

Third, we do not assume that the principal has precise control over the payoffs of either the agent or the monitor. Very often, rewards and punishments to agents and monitors are determined by imperfect and stochastic institutional processes. For instance, whistleblower protection schemes may not fully shield the monitor against ostracism, or harassment; supposedly anonymous information may be leaked; the judiciary may fail to act against criminal agents, and so on.[6] Therefore, our analysis takes payoffs as given and focuses on intervention strategies: how should the principal commit to react to reports of misconduct? Note that we do not dismiss payoffs as a key dimension of policy design. Rather, we show that regardless of the principal's ability to adjust payoffs, it is necessary to limit the information content of intervention policies to craft effective whistleblowing policies under the threat of intimidation.

We provide two sets of results. First, we establish a novel trade-off between eliciting information and using that information efficiently: effective intervention strategies must protect the content of the monitor's message by reducing the responsiveness of intervention to reports. Imagine a principal who is fully responsive: she launches an intervention if

[5] See Bertrand et al. (2007) or Olken (2007) for innovative approaches to measuring corruption.

[6] As an example of a principal failing to fully protect whistleblowers, Miceli et al. (1999) show that reported rates of retaliation against federal employees increased between 1980 and 1992 despite the tightening of whistleblower protection laws.

and only if she receives a report of crime. Intervention becomes a very precise signal of the monitor's report. The agent can then threaten to retaliate heavily in the event of intervention and thus ensure that there are no reports of crime. As a result there is no intervention, and threats are costless in equilibrium. Consider instead an intervention strategy with a positive baseline rate of intervention, so that intervention occurs with positive probability even when no crime is reported. This prevents the agent from inferring the monitor's intended report, making threats costly on the equilibrium path. To allow for whistleblowing in equilibrium, a successful intervention policy must guarantee the monitor sufficient plausible deniability.

We delineate the mechanics of crime, intimidation and reporting, as well as characterize the optimal intervention strategy, under perfect information. We highlight the impact of malicious monitors on comparative statics. Because malicious monitors potentially report honest agents, policies that increase both plausible deniability and intervention rates may reduce the welfare of honest agents, and increase the overall crime rate. Finally, we establish that under reasonable payoff assumptions, optimal intervention is interior. It may be optimal for the principal to intervene with probability less than one even after reports of crime.[7] This lets the principal reduce overall equilibrium intervention rates under responsiveness constraints.

Our second set of results provides guidelines for experimental policy evaluation on the basis of unverifiable reports in an environment with arbitrary uncertainty over the payoffs of agents and monitors. The difficulty is that threats by the agent, or a malicious monitor's own preferences, can lead to false reporting. Imagine that no misbehavior is reported. Does this imply that there is no underlying crime, or does it mean that would-be whistleblowers are being silenced by threats and intimidation? Alternatively, if greater plausible deniability increases complaints, are these complaints credible indication of crime? Could they be submitted by a malicious monitor who benefits from intervention against the agent? We

[7]Such policies may be untenable in practice: an anticorruption agency that does not follow every lead would likely be accused of undue favoritism.

show that for any *single* intervention strategy, monitors' messages are an unreliable estimate of the true prevalence of crime. However, testing specific *pairs* of intervention policies, chosen to keep responsiveness constant, does provide reliable bounds on the true prevalence of crime. These estimates can be exploited to craft effective intervention policies, robust to any true distribution of agents' and monitors' types.

The economic analysis of punishment structures designed to deter crime goes back to Becker (1968). This classic approach takes the probability of detecting crime as exogenous.[8] A smaller strand of the literature has explicitly considered the incentives of supervisors or monitors to collect information on crime and act on it.[9] Mookherjee and Png (1995) explicitly considers moral hazard in supervisory effort and reporting in a Principal-Agent-Monitor model. A rich contract theory literature (see for instance Tirole (1986), Laffont and Martimort (1997, 2000), Prendergast (2000), or Faure-Grimaud et al. (2003)) uses the Principal-Agent-Monitor framework to tackle collusion in organizations. It emphasizes the role of incomplete information frictions as a limitation on colluding parties. We contribute to this literature in two ways.

First, while existing work has focused on bribes to keep the monitor from informing the principal, collusion in our model comes in the form of punishments which, as opposed to payments, take place off the equilibrium path. The nascent economic literature on whistle-blowing emphasizes that informants are likely targets of retribution.[10] In such contexts, addressing explicit or implicit threats of retaliation is essential to ensure proper information flows.[11] We show that even without real anonymity (i.e. when there are few potential

[8] The numerous insights generated by this approach are reviewed in Polinsky and Shavell (2000)

[9] See for instance Becker and Stigler (1974) or Polinsky and Shavell (2001)

[10] Similar to us, Heyes and Kapur (2009) and Bac (2009) provide models where informants suffer costs. However, they do not consider how the principal's strategy can protect whistleblowers, do not consider the presence of malicious informants, and do not explore robust policy design. Makowsky and Wang (2018) provides experimental evidence on willingness to report in the presence of organization-wide punishments.

[11] See for instance Ensminger (2013) who emphasizes the role of threats and failed information channels in recent corruption scandals affecting community-driven development projects. Also, in a discussion of why citizens fail to complain about poor public service, Banerjee and Duflo (2006) suggest that "the beneficiaries of education and health services are likely to be socially inferior to the teacher or healthcare worker, and a government worker may have some power to retaliate against them."

whistleblowers), careful policy design can keep information channels open in spite of threats.

Second, we assume that information from the monitor is soft and unverifiable.[12] Hence, we do not focus on effort by the monitor and allow for malicious monitors that may want to inflict costs on innocent agents. In addition to solving for the optimal intervention strategy in a Bayesian environment, we rely on our structural model to tackle robust, data-driven policy evaluation based on unverifiable messages. This constitutes a novel direction for the literature on collusion in organizations.[13] In this respect, we contribute to a growing literature which takes a structural approach to experiment design in order to make inferences about unobservables.[14]

Ortner and Chassang (2018), share the insight that collusion may be addressed by creating endogenous contracting frictions between the agent and the monitor. We consider different frictions and emphasize different policy channels.[15] We show here that the principal can make the agent's own incentive provision problem more difficult by garbling the content of the monitor's messages.[16] This creates a novel practical rationale for the use of random mechanisms.[17] Our work also shares much of its motivation with the seminal work

[12]This stands in contrast with the related corporate leniency literature, which focuses on dynamic settings where monitors can defect from a cartel and trigger an investigation by providing hard information to prosecutors. Investigation typically prevents agents from punishing monitors. Spagnolo (2008) provides a review of this literature.

[13]This non-Bayesian perspective is shared with a growing body of work on mechanism design. See for instance Hurwicz and Shapiro (1978), Segal (2003), Hartline and Roughgarden (2008), Madarász and Prat (2010), Chassang (2013), Frankel (2014), Carroll (2013).

[14]See for instance Karlan and Zinman (2009) or Chassang et al. (2012).

[15]We focus on moral hazard and emphasize endogenously imperfect monitoring by the agent. Ortner and Chassang (2018) focuses on asymmetric information between the monitor and the agent, and emphasizes endogenous bargaining failures.

[16]This echoes the point, made by Dal Bó (2007) in a legislative context, that anonymous voting helps prevent influence activities and vote-buying.

[17]Myerson (1986) and Rahman (2012) consider mechanism design problems with non-verifiable reports, and emphasize the value of random recommendation-based incentives to jointly incentivize multiple agents, and in particular to incentivize both effort provision and the costly monitoring of effort. This strand of literature excludes the possibility of side contracting between players. As a result, the role of mixed strategies in our work is entirely different: monitoring itself is costless and randomization occurs to complicate the agent's own agency problem vis a vis the monitor. Eeckhout et al. (2010) propose a different theory of optimal random intervention based on budget constraints, and non-linear responses of criminal behavior to the likelihood of enforcement. Finally, Ederer et al. (2017) show that opacity that results in agents facing random incentive schemes can be useful to a principal that wants to minimize gaming.

of Warner (1965) on the role of plausible deniability in survey design, and the recent work of Izmalkov et al. (2011), Ghosh and Roth (2010), Nissim et al. (2011), or Gradwohl (2012) on privacy in mechanism design.

The paper is structured as follows. Section 2 describes relevant contexts for our analysis, taking the US Army's Counterinsurgency Policy as a lead example. Section 3 lays out our model. Section 4 delineates the tradeoff between exploiting messages efficiently and the need to protect singular sources. Section 5 shows how to design experiments permitting unambiguous policy evaluation on the basis of unverifiable reports. Section 6 discusses implementation challenges. Appendix A presents additional results, including an extension to the case of multiple monitors. Proofs are contained in Appendix B.

2 Relevant Contexts

2.1 Key Features

Our model is tailored to capture the mechanics of crime and whistleblowing in settings with three specific characteristics: (1) there must be significant information about criminal agents which the principal wants to obtain; (2) anonymity does not fully protect the set of individuals who have this information and are able to pass it on to the principal; (3) the agent is able to retaliate (at least with some probability) even after the principal's intervention.

Crime in our setting can encompass any misbehavior the principal seeks to prevent for which a small set of monitors is informed. This includes fostering and supporting terrorism, arrangements between police officers or judges and organized crime, bribe collection by state officials, fraud by sub-contractors in public-goods projects, breach of fiduciary duty by a firm's top executives, harassment within teams. We emphasize that here "crime" really covers any behavior that the principal finds undesirable, such as "shirking" in a typical principal-agent model. Retaliation takes many forms: an honest bureaucrat may be socially

excluded by his colleagues and denied promotion; whistleblowers may be harrassed, see their careers derailed, or get sued for defamation; police officers suspected of collaborating with Internal Affairs may have their life threatened by lack of prompt support.[18] Finally, the interventions available to the principal will also vary across contexts. A police department might launch a targeted Internal Affairs investigation. The board of a company can demand additional checks on the company's books, or investigate allegations of misconduct.

In all of these cases only a few colleagues, subordinates, or frequent associates are informed about the agent's misbehavior, making anonymity ineffective. Note that even when several monitors have information, group punishments may be used. For instance, entire communities may be denied access to public services following complaints to authorities.[19] In addition, monitors may fear that anonymity is not properly ensured and that imperfect institutions may leak the source of complaints to the agent or one of his associates. In hierarchical 360° evaluations, subordinates may not be willing to complain about their superior to their superior's boss if they worry that the two may share information.

2.2 Lead Example: Human Intelligence in the US Army

The US Army's human intelligence operations exemplify the strategic environments we are interested in. It incorporates all key elements of our model, and illustrates the steps taken by a successful organization to support informants and whistleblowers in a high-stakes environment.

The US invasion of Afghanistan and Iraq in 2001 and 2003 forced the US Army into protracted conflicts against multiple insurgencies. Building on the classical work of Galula

[18]See Punch (2009) for examples of punishment of informants in a study of police crime. In the National Business Ethics Survey 2013, 21% of whistleblowers report suffering several forms of retribution despite the legal and institutional protection available.

[19]For instance, Ensminger (2013) suggests that egregious crime affecting the World Bank's arid land program were not reported by the local Kenyan communities that suffered from it for fear of being cut off from subsequent projects. Indeed, in recent episodes UK aid to Uganda and Tanzania has been frozen amid allegations of impropriety.

(1964), as well as later developments, the US Army was able to formulate a systematic approach to the specific challenges of counterinsurgency.[20] The key strength of insurgents is their ability to mix with the civilian population. To counter this strength, it became essential for the Army to obtain tips and information regarding insurgent networks, weapon depots, and modes of operation from local populations.

The Counterinsurgency Field Manual, US Army and Marine Corps (2006), known as FM3-24, provides unique insight into the Army's efforts to support and encourage informants. It recognizes the importance of public goods, such as security and local development projects, as inputs to change the population's "hearts and minds" (Nagl et al., 2008).[21] Importantly for our purpose, FM3-24 also emphasizes fear of retaliation as a key determinant of human intelligence collection:

> "the lives of people offering information on insurgents are often in danger [...] careless handling of human sources by untrained personnel can result in murder or intimidation of these sources. When this occurs [tips] can be dramatically reduced due to the word spreading that US forces are careless or callous about protecting their sources."

The importance of informant protection is reflected in the singular responsibilities and decision-rights given to human intelligence (HUMINT) personnel. They have extensive control over the process of collecting and exploiting information, and bear the responsibility of trading-off short-term military gains against the safety of their informant network:

> "[HUMINT personnel] may sometimes choose not to share information because acting on intelligence can compromise its sources."
>
> "Actions that undermine trust or disrupt these networks – even those that provide a short-term military advantage – help the enemy."

[20] See Nagl (2002) and Kilcullen (2009) for modern takes. Ricks (2006, 2009) describe the situation in Iraq as well as the counterinsurgency-oriented changes associated with the surge of American troops in 2007 and 2008.

[21] See also Berman et al. (2011) for empirical and theoretical analysis of these links.

FM3-24, as well as the Human Intelligence Field Manual (US Army, 2006) known as FM2-22.3, specify steps that HUMINT officers should take to ensure the safety of their sources. Both when collecting tips from a known source, and when acquiring military targets based on such information, HUMINT officers take costly actions that make it hard to identify whether there was an informant behind a specific operation.

For instance, when collecting information, HUMINT officers avoid directly going to their known informants. Instead, FM3-24 encourages a proactive patrolling policy under which "one to two-thirds of the force should be on patrol at any time, day or night." Proactive patrolling creates constant contact between the military and the population, which makes it difficult to elucidate whether or when information regarding insurgent assets is being passed on. To support this objective, HUMINT officers are specifically trained not to spend more time with their sources than they would with any other civilian that approaches, or is approached by, the patrol. In addition to proactive patrolling, FM2-22.3 encourages the use of "screening operations" in which the local commander attempts to create a local census. These steps allow HUMINT officers to interact with sources in an inconspicuous way.

When acting on information provided by sources to acquire a target, similar steps are taken to shield likely informants from retaliation. In the words of FM3-24 "using the targeting process to synchronize targeting decisions is usually a good way to protect sources."[22] This is done in two ways. First, synchronized targeting creates a natural and significant delay between the date of a tip and the time an operation is put in place. Officers are advised to avoid hasty actions even at the cost of short-term military setbacks. Second, operations put in place as a result of tips are folded within the normal patrolling operations so that they are indistinguishable from business as usual. Officers are encouraged to routinely conduct "cordon and search" operations in which a section of the area (the size might vary

[22]The "targeting process" is the process by which specific operations are designed in terms of timing, target, type, etc.

from a street to a few contiguous blocks) is blocked off for a few hours while each household is searched. If a tip is received that a specific household is harboring insurgent assets, the local commander includes this information as part of the targeting process for subsequent "cordon and search" operations. If assets are found, insurgents cannot be certain that an informant was involved: "cordon and search" operations occur even in the absence of tips.

The US Army's counterinsurgency operations exemplify the sort of environment we are interested in: retaliation is a very real concern for potential informants, monetary incentives alone need not keep information channels open, and protecting sources is made difficult by the fact that few parties have actionable information. The rest of this paper models and analyzes the challenges faced by a principal in such a context. We show that protecting sources creates a novel rationale for random intervention policies, and provide tools for experimental policy evaluation.

3 Model

We study the interaction between three players: a principal P, an agent A and a monitor M.[23] In Section 4, we characterize equilibrium play and optimal policy design under complete information. In Section 5 we study how an uninformed principal can robustly evaluate policies by running experiments on a population of agent-monitor pairs.

Actions and timing. The agent chooses whether to engage in crime ($c = 1$) or not ($c = 0$). The principal does not observe crime c, but the monitor does and privately sends a message $m \in \{0, 1\}$ to the principal. The principal commits ex ante to an intervention strategy σ that launches an intervention $i \in \{0, 1\}$ against the agent as a function of message $m \in \{0, 1\}$. The agent observes intervention, and can then punish the monitor with intensity

[23] See Appendix A for an extension to the case of multiple monitors.

$r \in [0, +\infty)$.

Formally, the timing of actions in the game is as follows.

1. The principal commits to an intervention policy $\sigma : m \in \{0, 1\} \mapsto \sigma_m \in [0, 1]$, where $\sigma_m \equiv \text{prob}(i = 1|m)$ is the likelihood of intervention given message m.

2. The agent observes σ and

   - chooses whether to engage in crime ($c = 1$) or not ($c = 0$);
   - commits to a retaliation strategy $r : i \in \{0, 1\} \mapsto r(i) \in [0, +\infty)$ as a function of whether or not he suffers intervention.

3. The monitor costlessly observes crime c and sends a message $m \in \{0, 1\}$ to the otherwise uninformed principal.

4. The principal observes message m and triggers an intervention ($i = 1$) or not ($i = 0$) against the agent according to σ. Intervention has payoff consequences for the principal, agent and monitor that are detailed below.

   The agent does not observe m, but observes whether the principal triggers an intervention $i \in \{0, 1\}$.[24]

5. The agent retaliates against the monitor according to strategy $r : i \mapsto r(i)$.

The order of moves reflects the fact that the principal can commit publicly, but the agent cannot do so. Since open intimidation by an agent would be directly punishable by the principal, the agent must threaten the monitor privately. As a result, the principal cannot condition her policy on the agent's threats, and the agent is effectively a second mover.

We assume throughout the paper that whenever the agent is indifferent, he chooses not to be criminal, and whenever the monitor is indifferent, she reveals the truth. This convention simplifies the exposition, but does not matter for our results.

[24] An extension described in Chassang and Padro i Miquel (2014) allows agents to observe informational leaks from the institutional intervention process itself.

Payoffs, information and types. As a function of crime $c \in \{0,1\}$, intervention $i \in \{0,1\}$ and retaliation intensity $r \geq 0$, payoffs u_M, u_A and u_P to the monitor, agent and principal take the form

$$u_M = \pi_M \times c + v_M(c,m) \times i - r$$
$$u_A = \pi_A \times c + v_A(c) \times i - k(r)$$
$$u_P = \pi_P \times c + v_P(c) \times i$$

where: π_M, π_A, and π_P capture the expected payoff consequences of crime; v_M, v_A, and v_P capture reduced-form expected payoff consequences associated with intervention. The level of retaliation imposed by the agent on the monitor is denoted by r, and $k(r)$ is the cost of such retaliation to the agent. Payoffs conditional on crime are such that $\pi_A \geq 0$ and $\pi_P < 0$. The cost of retaliation $k(r)$ is strictly increasing and convex in r, with $k(0) = 0$.

To fix ideas, in a corporate fraud setting, $v_A(c = 1)$ would be the expected punishment of a corrupt executive if an audit of his activities takes place. This expected value reflects the probability that the audit comes up with actionable evidence, the probability that a court convicts him, and the professional, financial, and penal consequences thereof. Similarly, $v_A(c = 0)$ would capture an honest executive's inconvenience for being subjected to an audit, as well as the possibility of being wrongfully accused. Regarding the monitor, $v_M(c = 1, m = 1)$ captures psychological, reputational, and material rewards associated with having tipped the principal about wrongdoing. In contrast, $v_M(c = 1, m = 0)$ may include some form of punishment for failing to report the executive's misbehavior. We take these reduced-form payoffs as given, and motivate this modeling choice below. Throughout the paper, we maintain the following assumption.

Assumption 1 (general payoffs)**.** *It is common-knowledge that payoffs satisfy*

$$\forall c \in \{0,1\}, \quad v_A(c) \leq 0 \qquad \textit{(costly intervention)}$$

$$\forall c \in \{0,1\}, \quad v_M(c, m = c) \geq v_M(c, m \neq c) \qquad \textit{(weak preference for the truth)}$$

Costly intervention implies that the agent weakly suffers from intervention, so that the principal can use the threat of intervention to discipline him. Weak preference for the truth implies that, when intervention occurs, the monitor is weakly better off if she has told the truth to the principal. This assumption gives an operational meaning to messages $m \in \{0,1\}$.

Note that weak preference for the truth does not imply that the monitor is aligned with the principal. In particular, we allow for the possibility of *malicious monitors* who benefit from intervention against an honest agent, i.e. $v_M(c = 0, m = 1) > 0$. Taking intervention as given, the monitor would weakly prefer to report the truth, but she may choose to misreport an honest agent if it triggers an intervention. For example, a worker may benefit from discrediting an office-mate with whom she is competing for a promotion, or an informant may tip the police against an innocent citizen in order to settle a grudge. Since messages are not verifiable, this creates a (realistic) challenge for policy design.

Modeling choices. In setting up the model, we have made a few non-standard decisions which deserve a brief discussion. First, we assume that retaliation, rather than payments, is the incentive lever available to the agent, and this plays a role in our analysis. Appendix A provides sufficient conditions for this to be optimal even if the agent can commit to rewards as well as punishments. The intuition is that rewards to the monitor must be paid on the equilibrium path, whereas successful threats need only be implemented off of the equilibrium path. This point is particularly clear in bribe extortion cases where the monitor is the victim, and the point of crime is to extract money from the monitor in the first place.

Second, we allow the agent to commit to retaliate and the principal to commit to a mixed

intervention strategy. In Section 6 we justify these assumptions, explore the robustness of our results to weakening them, and describe implementation strategies that remove the need for mixing by the principal.

In a population setting, with many agent-monitor pairs, commitment to randomization can be replaced by commitment to a coarse targeting process. The principal intervenes against a set of agents that includes a target share of agents reported as criminal. In the case of Human Intelligence gathering by the US Army, intervention rates σ_0 and σ_1 directly correspond to the probability with which a household is included in a "cordon and search" operation depending on whether a tip was sent or not. This lets the Army commit to investigate households that have not been tipped off with positive probability by expanding the area of investigation around targets.

Finally, we treat payoffs upon intervention v_A, v_M, v_P as given rather than endogenize them. We do this for several reasons. First, it lets us focus on the novel aspects of our model: how the information content of intervention policies affects the agent's ability to discipline the monitor. Second, it reflects what we perceive as great heterogeneity in the ability of principals to reliably affect the payoffs of involved parties. For instance, international organizations, such as the World Bank, must go through local bureaucracies, and judicial systems to target misbehaving agents. This severely constrains their ability to tailor rewards and punishments. Similarly, given the informal nature of tips in many contexts, it is not clear that principals can fine-tune payoffs for all potential monitors. Third, and most importantly, even in the many contexts where the principal can affect payoffs directly, our analysis applies conditional on endogenous payoffs, provided they satisfy Assumption 1.

It is useful to state the last point formally. Let $\mathcal{V}$ denote the set of feasible payoff structures $v \equiv (v_A, v_M)$, Σ the set of possible intervention policies σ, and $c^*(v, \sigma)$, $m^*(v, \sigma)$ an appropriate selection of the agent and monitor's equilibrium behavior under payoff structure v and policy σ. The principal can be thought of as solving

$$\max_{v\in\mathcal{V},\sigma\in\Sigma}\mathbb{E}[u_P|\sigma,c^*(v,\sigma),m^*(v,\sigma)]=\max_{v\in\mathcal{V}}\max_{\sigma\in\Sigma}\mathbb{E}[u_P|\sigma,c^*(v,\sigma),m^*(v,\sigma)]. \tag{1}$$

If payoffs in $\mathcal{V}$ satisfy the weak requirements of Assumption 1, our analysis applies as a second stage within the broader mechanism-design problem in which payoffs are endogenously determined by the principal. For instance, our reduced-form payoffs capture schemes under which the monitor receives a reward ($v_M(c=1,m=1)>0$) for correctly informing the principal that the agent is criminal, and is instead punished for erroneous statements ($v_M(c,m\neq c)\leq 0$).

In the example of the US Army Counterinsurgency Policy, payoffs are endogenous (through the provision of private rewards, or public goods).[25] However, conditional on payoffs, keeping sources safe through a carefully crafted intervention policy remains essential.

4 Intervention and Intimidation under Complete Information

This section clarifies the joint mechanics of intervention and intimidation. It shows that as intervention policy σ varies, there is an important trade-off between efficiently exploiting reports from the monitor and limiting the agent's ability to silence such reports.

The first take-away is that effective policies must limit how responsive intervention can be to the reports, i.e. successful policies must garble the information content of intervention. The second take-away is that the presence of malicious monitors complicates the relationship between plausible deniability, information and crime. Finally, we show how these forces shape the optimal intervention strategy.

We maintain the following assumption throughout this section, except when explicitly mentioned otherwise.

[25] See also Basu et al. (2014) for a practical proposal on how to structure payoffs to maximize the flow of information, and an account of the real life limits that political controversy can put on policy-makers.

Assumption 2. *Payoffs are common knowledge and satisfy,* $\pi_A > 0$*, as well as*

$$\pi_A + v_A(c=1) < v_A(c=0) < 0, \text{ and} \qquad \text{(deterrent intervention)}$$

$$\pi_P \leq v_P(c=0) < 0. \qquad \text{(optimality of intervention)}$$

This assumption implies that: (1) when intervention occurs with probability 1, it is optimal for the agent to refrain from criminal behavior; (2) the principal prefers to incur the cost of intervention v_P over the cost of crime π_P. Altogether, this implies that it is optimal for the principal to deter crime in equilibrium. The question is how to do so in the most efficient way.

4.1 The Need for Plausible Deniability

Our first result shows that intimidation has a dramatic impact on the shape of efficient intervention policies. We contrast equilibrium outcomes when the monitor is an automaton compelled to report the truth ($m(c) = c$), and when the monitor endogenously responds to threats by the agent.[26]

Proposition 1 (plausible deniability)**.** *(i) If messages are exogenously informative, i.e.* $\mathbf{m}(c) = c$*, setting* $\sigma_0 = 0$ *and* $\sigma_1 = 1$ *is an optimal policy. There is no crime and no retaliation in equilibrium.*

(ii) If messages are endogenous, there exists $\overline{\lambda} > 1$ *such that for any intervention policy* σ *satisfying* $\frac{\sigma_1}{\sigma_0} \geq \overline{\lambda}$*,*

- *the agent engages in crime* $c = 1$*, and commits to retaliate conditional on intervention;*
- *the monitor sends message* $m = 0$*.*

We refer to the likelihood ratio of intervention rates, $\lambda \equiv \frac{\sigma_1}{\sigma_0}$ as the responsiveness of

[26] Note that in this simple setting, a binary message space is without loss of efficiency: collecting messages from the agent, or richer messages from the monitor (for instance about threats of retaliation) is not helpful. See Appendix A, Lemma A.1 for details.

policy σ to reports. It captures the information content of intervention as a signal of the monitor's report.

When the monitor is an automaton compelled to tell the truth, the optimal intervention is fully responsive to the monitor's message and sets $\sigma_0 = 0$, inducing $\lambda = +\infty$. Doing so provides the strongest incentives for the agent to choose $c = 0$, and since $\sigma_0 = 0$, there are no interventions, and therefore no costs to the principal, on the equilibrium path.

This is no longer the case when messages are endogenous and the monitor can be deterred from truthful reporting. When responsiveness λ is high enough, intervention becomes a very informative signal of the message the monitor sent. By committing to a sufficiently high (but bounded) level of retaliation r conditional on intervention $i = 1$, the agent can induce the monitor to send message $m = 0$ for all values of $\lambda > \overline{\lambda}$. On the equilibrium path, the expected cost of retaliation is equal to $\sigma_0 k(r)$. This implies that as σ_0 approaches 0, effective threats against the monitor are costless to the agent in equilibrium. Since $\pi_A > 0$, it is optimal for the agent to engage in criminal behavior and threaten the monitor into silence when σ_0 is sufficiently low.

As we emphasized in Section 3 when discussing our choice to use reduced-form payoffs, this result holds regardless of payoffs upon intervention, provided they satisfy Assumption 1. This implies that organizations who can alter payoffs through rewards, fines or other means will still find it necessary to use a positive baseline rate of intervention to reduce responsiveness. This is true in the case of the US Army: although the Army uses material incentives to encourage informants to come forth, it also expends considerable resources on maintaining a positive baseline rate of intervention through the routine use of "cordon and search" operations.

We now delineate the mechanics of reporting, intimidation and crime. We proceed by backward induction: first we study reporting and intimidation, taking crime decision $c \in \{0, 1\}$ as given; second, we study the agent's criminal choices as a function of intervention policy σ; finally we characterize the optimal intervention policy.

4.2 Optimal Threats by the Agent

The assumption of weak preference for the truth, $v_M(c, m = c) \geq v_M(c, m \neq c)$, implies that in equilibrium: (1) the principal's policy $\sigma = (\sigma_0, \sigma_1)$ satisfies $\sigma_0 < \sigma_1$, and (2) the agent chooses not to retaliate conditional on no intervention (i.e. $r(i = 0) = 0$, see Appendix A for a proof). We denote by r the agent's choice of retaliation conditional on intervention $i = 1$.

Threats needed to keep the monitor silent. Given a choice $c \in \{0, 1\}$ by the agent, we identify the minimum retaliation r needed to induce message $m = 0$. To do so, note that the monitor sends message $m = 1$, if and only if

$$\sigma_1[v_M(c, m = 1) - r] \geq \sigma_0[v_M(c, m = 0) - r].$$

This holds whenever

$$r \leq r_\sigma^c \equiv \left[\frac{\sigma_1 v_M(c, m = 1) - \sigma_0 v_M(c, m = 0)}{\sigma_1 - \sigma_0}\right]^+ \tag{2}$$

where $x^+ \equiv \max\{x, 0\}$ by convention. By committing to retaliation at least r_σ^c, an agent making a crime decision $c \in \{0, 1\}$ can induce the monitor to send message $m = 0$. Observe that r_σ^c can be expressed as a function of crime c and responsiveness $\lambda = \frac{\sigma_1}{\sigma_0}$:

$$\begin{aligned} r_\sigma^c = r_\lambda^c &\equiv \left[\frac{\lambda v_M(c, m = 1) - v_M(c, m = 0)}{\lambda - 1}\right]^+ \\ &= \left[v_M(c, m = 1) + \frac{v_M(c, m = 1) - v_M(c, m = 0)}{\lambda - 1}\right]^+ \end{aligned} \tag{3}$$

The following lemma contrasts the impact of responsiveness λ on effective threats by a criminal agent (an agent that chooses $c = 1$) and an honest agent (an agent who chooses $c = 0$). We highlight the impact of malicious monitors.

Lemma 1 (responsiveness and retaliation)**.** *The following comparative statics hold.*

(i) The level of retaliation $r_\lambda^{c=1}$ that a criminal agent must commit to in order to induce message $m = 0$ is weakly decreasing in responsiveness λ.

If $v_M(c = 1, m = 1) < 0$, then for λ high enough, $r_\lambda^{c=1} = 0$.

(ii) The level of retaliation $r_\lambda^{c=0}$ that an honest agent must commit to in order to induce message $m = 0$ is weakly increasing in λ.

If $v_M(c = 0, m = 1) > 0$, then for λ high enough, $r_\lambda^{c=0} > 0$.

Point (i) shows that higher responsiveness facilitates intimidation by a criminal agent. The intuition derives from (3). Because we assume weak preferences for the truth, $v_M(c = 1, m = 1) - v_M(c = 1, m = 0) \geq 0$, if intervention occurs, a monitor who remained silent about a criminal agent experiences some loss. For high values of responsiveness, σ_0 must be low, which reduces the odds that a misreporting monitor experiences this loss. As a result, the need for retaliation diminishes. In fact if the monitor dislikes intervention ($v_M(c = 1, m = 1) < 0$) there will be no need for retaliation for responsiveness λ high enough. This possibility is a notable concern in the context of foreign aid: if corruption scandals are typically followed by suspension of aid programs, aid recipients may choose not to report administrative abuses (Ensminger, 2013).

Inversely, higher responsiveness increases necessary retaliation levels for an honest agent. Indeed, for an honest agent, $v_M(c = 0, m = 1) - v_M(c = 0, m = 0) < 0$. As a result, if λ is close to 1, the monitor may as well not lie since it doesn't change the odds of intervention, and saves her the loss from misreporting. Consider the case of a malicious monitor who benefits from triggering intervention, even if she is required to lie: $v_M(c = 0, m = 1) > 0$. As λ increases, and σ_0 becomes small, misreporting becomes the only way for the monitor to enjoy payoff $v_M(c = 0, m = 1)$. For sufficiently high responsiveness λ, even an honest agent needs to commit to positive levels of retaliation $r_\lambda^{c=0} > 0$ to ensure the monitor sends message $m = 0$.

Optimal intimidation. We now study when it is in the interest of the agent to induce message $m = 0$. An agent taking decision $c \in \{0, 1\}$ chooses to induce message $m = 0$ through the threat of retaliation if and only if:

$$\sigma_1 v_A(c) \leq \sigma_0[v_A(c) - k(r^c_\lambda)] \iff \lambda v_A(c) \leq v_A(c) - k(r^c_\lambda). \tag{4}$$

As Lemma 1(*ii*) highlights, the presence of malicious monitors has a large impact on the relationship between intervention policy and information transmission. Facing a malicious monitor, even honest agents have to commit to retaliation to ensure message $m = 0$ is reported, thus increasing the relative cost of honesty. The following lemma characterizes optimal commitment to retaliation.

Lemma 2 (optimal intimidation)**.** *Assume the agent is criminal ($c = 1$). The agent's decision to threaten the monitor into silence is monotonic in λ: there exists $\lambda_1 > 1$ such that the agent induces message $m = 0$ if and only if $\lambda > \lambda_1$.*

Assume the agent is honest ($c = 0$).

(i) If the monitor is not malicious, $v_M(c = 0, m = 1) \leq 0$, an honest agent sets $r = 0$ and the monitor sends message $m = 0$.

(ii) If the monitor is malicious, the agent's decision to induce message $m = 0$ can be non-monotonic in λ.

- *If $\lambda \to 1$, then $m = 0$ and $r = 0$.*
- *If $\lambda \to \infty$, then $m = 0$ and $r > 0$.*
- *Given (v_A, v_M), there exists a cost function k and values $\lambda \in (1, +\infty)$ such that $m = 1$ and $r = 0$.*

An implication of Lemma 2 is that malicious monitors greatly complicate the evaluation of intervention policies: if low responsiveness results in reports $m = 1$, is it because the agent is criminal, or because the monitor is malicious? We tackle this issue in Section 5.

4.3 Crime and Optimal Intervention

Crime. Taking into account the agent's optimal intimidation behavior, the agent chooses to engage in crime $c = 1$ if and only if

$$\max\{\sigma_1 v_A(c=0),\, \sigma_0[v_A(c=0) - k(r_\lambda^0)]\} \\ < \pi_A + \max\{\sigma_1 v_A(c=1),\ \sigma_0[v_A(c=1) - k(r_\lambda^1)]\}. \tag{5}$$

The following comparative static is at the heart of the policy design question: how do crime and reporting decisions vary with the principal's intervention policies? Specifically, we consider two policies σ^α and σ^β such that:

$$\sigma_0^\alpha < \sigma_0^\beta, \quad \sigma_1^\alpha < \sigma_1^\beta, \quad \text{and} \quad \frac{\sigma_1^\alpha}{\sigma_0^\alpha} \geq \frac{\sigma_1^\beta}{\sigma_0^\beta}. \tag{6}$$

Policy σ^β involves both higher intervention frequency and lower responsiveness. Let (c^α, m^α), (c^β, m^β) denote the equilibrium crime and reporting decisions under σ^α and σ^β. We also relax Assumption 2 and consider both the cases where $v_A(c=0) = 0$ and $v_A(c=0) < 0$.

Proposition 2 (comparative statics)**.** *(i) Assume that there are no malicious monitors and $v_A(c=0) = 0$. We have that $c^\alpha \geq c^\beta$. Furthermore, if $m^\alpha > m^\beta$ then $c^\alpha > c^\beta$.*

(ii) Neither of these properties needs to hold when there are malicious monitors or $v_A(c=0) < 0$.

Point (i) establishes an intuitive and encouraging result. In an ideal payoff environment, where monitors are non-malicious and intervention does not hurt honest agents, policies that increase intervention frequency and reduce responsiveness must reduce crime. In addition, drops in complaints across policies reliably indicate a reduction in crime.

Point (ii) establishes that neither result holds if the monitor is malicious or intervention

is costly to the agent: there exist environments under which crime grows despite higher frequency of interventions, $c^\beta > c^\alpha$; there may also exist environments for which reductions in complaints do not imply reductions in actual crime.

Optimal intervention. We now have the tools to characterize the principal's optimal policy. It follows from Assumption 2 that under any optimal policy the principal ensures that the agent does not engage in crime. The optimal policy implements $c = 0$ at the lowest equilibrium cost:

$$\min_{\sigma_0,\sigma_1} \sigma_0 \tag{7}$$

$$s.t. \quad \max\{\sigma_1 v_A(c=0),\, \sigma_0[v_A(c=0) - k(r^0_\lambda)]\}$$
$$\geq \pi_A + \max\{\sigma_1 v_A(c=1),\, \sigma_0[v_A(c=1) - k(r^1_\lambda)]\} \tag{8}$$

Constraint (8) ensures that the agent chooses not to be criminal. Note that it nests the agent's choice to intimidate the monitor or not.

The following proposition summarizes relevant characteristics of the optimal policy. Recall that λ_1 (defined in Proposition 2) is the unique responsiveness rate such that a criminal agent is indifferent between silencing the monitor or not. Similarly, let Λ_0 denote the set of responsiveness rates λ such that an honest agent is indifferent between inducing truthful reporting or not. Set Λ_0 may be empty or may include one or more values.

Proposition 3 (optimal intervention policy)**.** *Under Assumption 2, the optimal intervention strategy σ^* is such that:*

(i) If the monitor is not malicious, then $\frac{\sigma_1^*}{\sigma_0^*} = \lambda_1$;

(ii) If the monitor is malicious, then either $\frac{\sigma_1^*}{\sigma_0^*} = \lambda_1$ or $\frac{\sigma_1^*}{\sigma_0^*} \in \Lambda_0$;

(iii) σ^* is interior for generically every pair (v_A, v_M): $\sigma_0^* \in (0,1)$ and $\sigma_1^* \in (0,1)$.

Points (i) and (ii) state that at an optimal policy, either an honest, or a criminal agent is indifferent between silencing the monitor or not. This corresponds to either minimizing a criminal agent's payoff, or maximizing an honest agent's payoff, under responsiveness constraints.

Point (iii) establishes that optimal policy generically sets $\sigma_1^* < 1$. The logic of the argument is worth clarifying. Assumption 2 implies that full intervention ($\sigma = (1, 1)$) strictly deters crime. It is therefore possible to bring σ_1 below 1 while still deterring crime. In addition, lowering σ_1 allows the principal to reduce σ_0 while satisfying responsiveness constraints needed to maintain information flows.[27]

We now provide comparative statics with respect to payoffs, showing that payoff design is indeed an important policy dimension.

Proposition 4 (the role of payoffs)**.** *The following comparative statics hold:*

(i) σ_0^* *is decreasing in* $v_A(c = 0)$ *and increasing* $v_A(c = 1)$.

(ii) σ_0^* *is weakly decreasing in* $v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$.

(iii) σ_0^* *is weakly decreasing in* $v_M(c = 0, m = 0) - v_M(c = 0, m = 1)$.

This proposition highlights that principals who can better punish criminal agents or reward truthful monitors can afford to reduce the baseline rate of intervention. However, the ability to affect payoffs does not dispense the principal from maintaining plausible deniability, since Proposition 1 still applies. The correct insight here is that better material incentives make it easier to reach the necessary level of plausible deniability. These results are aligned with the findings of Berman et al. (2011): material rewards help reduce insurgent violence in Iraq, but only if counterinsurgent presence is high enough to allow for sufficient plausible deniability.

[27] Due to political or technological constraints, it might be difficult for a principal to commit to $\sigma_1^* < 1$. An immediate corollary of Proposition 3 is that a principal required to set $\sigma_1 = 1$ will incur higher costs along the equilibrium path as she will need to set a higher σ_0 in order to keep responsiveness low enough.

5 Experiment Design and Policy Evaluation

In this section we depart from the objective of setting intervention policies that are optimal under perfect information on payoffs, or even under a more sophisticated Bayesian prior. Instead, we seek to provide conditions under which the principal is able evaluate whether one policy is effectively controlling crime in a prior-free way. However, policy evaluation is difficult in this setting because crime c is typically difficult to measure, and messages m are unverifiable. We need a theory of inference from unverifiable messages.

5.1 Inference from Experiments

To allow for meaningful inference and policy experimentation, consider a situation in which a single principal faces a large set of agent-monitor pairs and can test intervention policies on subsets of this population. These are demanding requirements, but they are plausibly satisfied in several settings of interest: managers and subordinates in large organizations; police officers and their partners; police officers and the citizens they are supposed to serve; bureaucrats and the citizens who avail themselves of public services.

We denote by $\tau = (\tau_M, \tau_A) \in T_M \times T_A = T$ the types associated to agent-monitor pairs. The monitor's type τ_M determines her payoffs (π_M, v_M), while the agent's type τ_A determines both his payoffs (π_A, v_A, k), and his belief over the type τ_M of the monitor, which we denote by $\Phi(\tau_M|\tau_A) \in \Delta(T_M)$. Pairs of types (τ_M, τ_A) are drawn i.i.d. from the population distribution $\mu_T \in \Delta(T)$.

While agents and monitors know their own type, the true distribution μ_T is unknown to the players and may exhibit arbitrary correlation between the types of monitors and agents. We assume that Assumption 1 holds and is common knowledge. Crucially, we do not impose Assumption 2: a positive mass of agents may engage in crime regardless of intervention. As in Banerjee et al. (2017), the role of policy evaluation here is to help craft effective policies under all possible type distributions μ_T.

Bounds on crime. We now show that although messages are unverifiable, it is possible to place useful bounds on the true frequency of crime provided policy experiments are suitably chosen.

Given policy σ, the aggregate mass of reports $m = 1$ takes the form

$$\int_T m^*(\sigma, \tau) \mathrm{d}\mu_T(\tau),$$

where $m^*(\sigma, \tau)$ is the message induced by intervention policy σ on an agent-monitor pair of type τ. This sample mass of complaints is observable and, since messages are binary, it is a sufficient statistic for the sample of messages $(m_\tau)_{\tau \in T}$ sent by monitors participating in the experiment, i.e. it is the only data that the principal can rely on for inference.

The next proposition states that unverifiable messages at a single policy profile σ imply no restrictions on the true frequency of crime.

Proposition 5 (single experiments are uninformative)**.** *Take as given a policy profile σ, and a true distribution μ_T yielding aggregate complaint rate $\int_T m^*(\sigma, \tau) d\mu_T(\tau)$. We have that*

$$\left\{ \int_T c^*(\sigma, \tau_A) d\widehat{\mu}_T(\tau), \quad \textit{for} \quad \widehat{\mu}_T \ \textit{s.t} \ \int_T m^*(\sigma, \tau) d\widehat{\mu}_T(\tau) = \int_T m^*(\sigma, \tau) d\mu_T(\tau) \right\} = [0, 1].$$

In words, reports of crime at a single policy profile are compatible with any frequency of crime. This negative result follows from the possible existence of malicious monitors. A high number of complaints might be a sign that crime is high, or that malicious monitors are misreporting honest agents. We now show that suitably chosen *pairs* of intervention policies imply useful restrictions on underlying levels of crime.

Consider two intervention policies σ^α and σ^β implemented on two different samples of agent-monitor pairs, drawn i.i.d. from the same underlying distribution μ_T.

Proposition 6 (bounds on crime)**.** *Pick policies σ^α, σ^β such that $\sigma^\alpha = \rho \sigma^\beta$ with $\rho < 1$. For*

all underlying distributions $\mu_T \in \Delta(T)$,

$$\int_T \left[c^*(\sigma^\alpha, \tau) - c^*(\sigma^\beta, \tau)\right] d\mu_T(\tau) \geq \left| \int_T [m^*(\sigma^\alpha, \tau) - m^*(\sigma^\beta, \tau)] d\mu_T(\tau) \right|. \tag{9}$$

The practical implications of Proposition 6 are best illustrated by a numerical example: if, say, 20% of monitors send message $m = 1$ under policy σ^β, while 35% do so under policy σ^α, then the principal knows that crime must be at least 15% higher under policy σ^α than σ^β. This inference holds independently of the underlying distribution of types μ_T.

To understand why this result is true, recall expression (4). Given payoffs and conditional on $c \in \{0, 1\}$, the decision to threaten the monitor to ensure $m = 0$ depends only on λ. Proposition 6 compares two policies that share the same responsiveness λ. Hence, if the messages are different across the two policies, it must be because the underlying crime decisions $c \in \{0, 1\}$ are different.

Proposition 6 provides a bound for the difference in the frequency of crime between two intervention policies. An immediate corollary provides bounds on crime at both policies.

Corollary 1. *The difference in messages* $\left| \int_T [m^*(\sigma^\alpha, \tau) - m^*(\sigma^\beta, \tau)] d\mu_T(\tau) \right|$ *is a lower bound for the mass* $\int_T c^*(\sigma^\alpha, \tau) d\mu_T(\tau)$ *of criminal agents at policy* σ^α*, and a lower bound for the mass* $\int_T [1 - c^*(\sigma^\beta, \tau)] d\mu_T(\tau)$ *of honest agents at policy* σ^β*.*

Indeed, if there is 15% less crime under policy σ^β than σ^α, it must be that the crime rate is at least 15% under σ^α and at least 15% of agents are honest under σ_B.

5.2 Policy Implications

We conclude this section by proposing a heuristic decision-rule that exploits Proposition 6 to inform policy design.

Imagine that some set of intervention strategies $\sigma \in \Sigma$ is tested on subsets of the same population μ_T of types (τ_M, τ_A), where Σ is a set of feasible policy profiles. Denote by

$\widehat{C} : [0,1]^2 \to [0,1]$ the function defined by

$$\forall \sigma \in [0,1]^2, \quad \widehat{C}(\sigma) \equiv 1 - \max \left\{ \left| \int_T [m^*(\sigma,\tau) - m^*(\sigma',\tau)] \mathrm{d}\mu_T(\tau) \right| \middle| \sigma' \in \Sigma \cap \{\rho\sigma | \rho \in [0,1]\} \right\}.^{28}$$

In words, for each policy σ we take the largest difference in messages between this policy and other tested policies that have the same responsiveness λ but feature lower intervention rates. From Corollary 1, we know that $\widehat{C}(\sigma)$ is an upper bound to the amount of underlying crime at σ. Let $\underline{v}_P = v_P(c=0) \leq v_P(c=1)$. Noting that for a given intervention profile σ, the principal's payoff is

$$\mathbb{E}_{\mu_T}[u_P | c^*, m^*, \sigma] = \pi_P \int_T c^*(\sigma,\tau_A) \mathrm{d}\mu_T(\tau) + \int_T v_P(c^*(\sigma,\tau_A)) \left[\sigma_0 + (\sigma_1 - \sigma_0) m^*(\sigma,\tau)\right] \mathrm{d}\mu_T(\tau),$$

we obtain the following corollary.

Corollary 2. *For any intervention profile σ, we have that*

$$\mathbb{E}_{\mu_T}[u_P | c^*, m^*, \sigma] \geq \pi_P \widehat{C}(\sigma) + \underline{v}_P \left[\sigma_0 + (\sigma_1 - \sigma_0) \int_T m^*(\sigma,\tau) d\mu_T(\tau)\right].$$

Furthermore, if $\Sigma = \{\sigma \in [0,1]^2 \text{ s.t. } \sigma_1 \geq \sigma_0\}$, then the data-driven heuristic policy $\widehat{\sigma}(\mu_T)$ defined by

$$\widehat{\sigma}(\mu_T) \in \arg\max_{\sigma \in \Sigma} \pi_P \widehat{C}(\sigma) + \underline{v}_P \left[\sigma_0 + (\sigma_1 - \sigma_0) \int_T m^*(\sigma,\tau) d\mu_T(\tau)\right]$$

is a weakly undominated policy with respect to the unknown true distribution μ_T.

By running pairs of policy experiments that share the same responsiveness, it is possible to obtain a tight upper bound $\widehat{C}(\sigma)$ on the amount of underlying crime. This upper bound can be computed using only unverifiable reports at each tested policy σ. Policy $\widehat{\sigma}(\mu_T)$ depends on the true distribution μ_T only through these unverifiable reports. It balances

[28] In the event that $\Sigma \cap \{\rho\sigma | \rho \in [0,1]\}$ is empty, we let $\widehat{C}(\sigma) = 1$.

estimates of crime against the cost of intervention. When the set of policy experiments Σ is large enough, policy $\widehat{\sigma}(\mu_T)$ is weakly undominated.

6 Discussion

6.1 Summary

We model the problem of a principal who relies on messages from an informed monitor to target intervention against a possibly criminal agent. The difficulty is that the agent can dissuade the monitor from informing the principal by threatening to retaliate conditional on intervention. In this setting, intervention becomes a signal which the agent can exploit to effectively dissuade the monitor from complaining. As a consequence, effective intervention strategies must garble the information content of messages. In particular, there needs to be a positive baseline rate of intervention following messages indicating no criminal behavior. This creates an imperfect monitoring problem between the agent and the monitor which limits the agent's effectiveness at silencing the monitor.

Because hard evidence of crime is hard to come by, we explore the extent to which one can make inferences about unobservable crime, as well as evaluate policies, on the basis of unverifiable messages alone. We consider a general framework which allows for near arbitrary incomplete information and heterogeneity across agents and monitors. We establish general properties of reporting and crime patterns in equilibrium that imply bounds on underlying crime as a function of unverifiable reports. These bounds permit policy evaluation on the basis of unverifiable messages.

A strength of our analysis is that it does not presume that the principal has extensive control over the payoffs of the agent and the monitor. This accommodates environments in which the relevant principal has to rely on existing institutional channels to carry out interventions. On the other hand our policy suggestions raise practical concerns.

Commitment by the principal. Our analysis assumes the principal is able to commit to mixed strategies, which is admittedly more demanding than committing to pure strategies. A standard way to justify this assumption would be to invoke reputational concerns in an unmodelled continuation game. After all, principals can commit publicly to a specific policy, such as the US Army's "cordon and search" practice. Informants would likely stop providing information to the US Army if the latter acted recklessly, allowing insurgents to infer that tips had been provided. Indeed, this dynamic consideration is explicit in the quotes provided in Section 2.2. Committing to mixed strategies is then essentially equivalent to forming a reputation under imperfect public monitoring (Fudenberg and Levine, 1992).

An additional, more practical, observation is that commitment to mixed strategies can be achieved through hard-wired garbling of messages at the surveying stage. Specifically, instead of recording messages directly, the principal may record the outcomes of two Bernoulli lotteries l_0 and l_1 such that

$$l_0 = \begin{cases} 1 & \text{with proba} \quad \sigma_0 \\ 0 & \text{with proba} \quad 1-\sigma_0 \end{cases} \quad \text{and} \quad l_1 = \begin{cases} 1 & \text{with proba} \quad \sigma_1 \\ 0 & \text{with proba} \quad 1-\sigma_1. \end{cases}$$

The monitor's message $m \in \{0,1\}$ corresponds to a choice of lottery. The lottery yields an outcome $\widehat{m} \in \{0,1\}$ observed by the principal who then intervenes according to pure strategy $i(\widehat{m}) = \widehat{m}$. This approach has the benefit of making plausible deniability manifest to participating monitors, and no repeated-game reputational incentives are needed. Crucially for the results of Section 5, one can recover aggregate submitted reports from outcome data $\widehat{m}$ alone. For any mapping $m : T \to \{0,1\}$,

$$\int_T m(\tau)\mathrm{d}\mu_T(\tau) = \frac{\int_T \widehat{m}(\tau)\mathrm{d}\mu_T(\tau) - \sigma_0}{\sigma_1 - \sigma_0}.$$

Note that this implementation of mixed strategies is closely related to the randomized re-

sponse techniques introduced by Warner (1965).[29]

Commitment by the agent. Throughout the paper we assume that the agent has commitment power: the agent can commit to arbitrary levels of retaliation in the event of intervention. Intuitively, repeated games incentives seem a plausible motivation for this assumption. Agents are long-run players who want to maintain a reputation vis-à-vis the monitor, and other potential monitors. Failing to punish one monitor could lead to further information being passed on to the principal. However this repeated game interpretation suggests some refinements of the commitment assumption.

First, it seems reasonable that the agent need not be able to retaliate for sure in the event of intervention. If the intervention process is sufficiently effective, the agent may sometimes be prevented from retaliating altogether. This could be modeled by assuming that the agent can only retaliate with some probability q. Provided that $q > 0$, this would not change our analysis. What matters for incentives is the expected retaliation conditional on intervention. Probabilistic retaliation is then equivalent to a change in the cost of expected retaliation.

Second, if commitment derives from repeated game incentives, the maximum amount of retaliation r the agent can commit to will be bounded above by the discounted surplus he expects from maintaining his reputation. This means that there will be an upper bound $\overline{r}$ on credible retaliation levels. This can be captured within the existing framework by assuming that the cost of retaliation $k(r)$ becomes arbitrarily large when $r \geq \overline{r}$. Note that as long as $\overline{r} \geq v_M(c, m = 1)$, all of the results in the paper continue to hold. This is intuitive: to silence the monitor, the agent must be able to commit to a retaliation as large as the benefit the monitor obtains from sending message $m = 1$. If $\overline{r} \leq v_M(c = 1, m = 1)$ results that do not consider the case where λ approaches $+\infty$ continue to hold. This is the case of Propositions

[29]The main difference is that typical randomized response techniques simply enjoin the monitor to garble her response, but the monitor can always submit her preferred message. Hence, in our fully rational framework, traditional randomized response techniques do not guarantee plausible deniability in equilibrium. This difference is important when messages are used for equilibrium incentive design, rather than for one-shot surveys.

1, 2, 5 and 6. Other results need not extend. For instance, if $\overline{r} < v_M(c = 1, m = 1)$ Proposition 1 does not hold: although $\sigma_0 = 0$ the agent cannot induce message $m = 0$.

Ethics of experimentation. Proposition 6 and Corollary 2 suggest that a principal may use experimental variation in intervention rates to evaluate the effectiveness of intervention policies. This requires the principal to experiment, at least temporarily, with relatively low intervention rates. Such experiments are subject to the usual trade off between experimentation and exploitation, with the added ethical demands that come from the fact that situations of interest likely involve high stakes. Ex post inefficient use of resources may be unfeasible.

One acceptable way for a principal to create necessary variation in intervention rates may be to redistribute investigative resources unevenly across different selections of the population. Keeping the ratio of intervention rates constant, one set of agent-monitor pairs would be exposed to low intervention rates, while an other set of agent-monitor pairs would experience high intervention rates. This is not an obvious misuse of available resources.

Costly messaging. Our analysis relies on the assumption that the monitor's report has an impact on payoffs only conditional on intervention. This assumption would fail if some messages had an intrinsic cost, but not others. For instance, the analysis of Section 5 fails if filing a complaint involves a costly administrative process, while not complaining is free. Our results apply in environments where the monitor is already being surveyed, so that the cost of sending a report, positive or not, is sunk.

Endogenous response by the judiciary. Our analysis does not consider the possibility that payoffs upon intervention (v_A, v_M) could depend on the intervention profile σ. For instance, the judiciary could put lower effort into investigating agents if it is known that the baseline rate of intervention σ_0 is high. While a full fledged analysis of such an environment is beyond the scope of this paper, we believe that many of our results would continue to hold.

A baseline rate of intervention would remain necessary to prevent the agent from completely silencing the monitor. In addition, Proposition 6 considers policies such that the information content of intervention remains constant. As a result, there is no reason for the endogenous response of the judiciary to change across these policy choices.

Appendix – For Online Publication

A Additional Results

A.1 Simple Policies

This section motivates the class of policies studied in the paper. We work under the complete information setting of Section 4. Assumption 2 holds throughout the section.

Messages about threats. We first show that the principal need only elicit binary messages from the monitor. Because the agent has commitment power of his own, the principal is unable to leverage the usual cross validation techniques to extract surplus.

Lemma A.1. *It is without loss of efficiency for the principal to:* (i) *not elicit messages from the agent;* (ii) *offer the monitor only binary messages* $0, 1$*;* (iii) *use an intervention policy satisfying* $\sigma_0 \leq \sigma_1$*.*

Proof. We begin by showing point (i): it is without loss of efficiency not to elicit messages from the agent. The agent has commitment power and therefore can commit to the messages he sends. When the agent sends a message, we can think of him as choosing the intervention profile σ he will be facing, as well as the messages sent by the monitor, at some implementation cost. If a non-criminal agent chooses intervention profile σ, then giving additional choices can only increase the payoffs of a criminal agent. Hence the principal can implement the same outcome by offering only the profile σ chosen by a non-criminal agent.

We now turn to point (ii) and consider enlarging the set of messages submitted by the monitor. The monitor observes only two pieces of information: the crime status $c \in \{0, 1\}$ of the agent, and the level of retaliation $r \in \mathbb{R}$ that he is threatened with in the event of intervention. A priori, the principal may elicit messages $(m, \rho) \in \{0, 1\} \times [0, +\infty)$ about both the crime status of the agent and the retaliation level she has been threatened with. This means that intervention rates now take the form $\sigma_{m,\rho} \in [0, 1]$.

Take as given an intervention profile $\sigma = (\sigma_{m,\rho})_{m\in\{0,1\},\rho\in[0,+\infty)}$. First, note that we can focus on the case where the agent's optimal decision is to be non-criminal, otherwise no-intervention is the optimal policy. Second, noting that the value of ρ submitted by the monitor must solve $\max_{\rho\in[0,+\infty)} \sigma_{m,\rho}(v_M(c,m) - r)$ it follows that without loss of generality one can focus on binary values of $\rho \in \{-,+\}$ such that $\sigma_{m,-} = \inf_{\rho\in[0,+\infty)} \sigma_{m,\rho}$ and $\sigma_{m,+} = \sup_{\rho\in[0,+\infty)} \sigma_{m,\rho}$. When the monitor is indifferent, she must be inducing the lowest possible intervention rate, otherwise the agent would increase retaliation by an arbitrarily small amount.

Finally, without loss of efficiency, one can consider intervention profiles such that for all $\rho \in \{-,+\}$, $\sigma_{0,\rho} \leq \sigma_{1,\rho}$. Indeed, given ρ, define $\overline{\sigma} = \max_{m\in\{0,1\}} \sigma_{m,\rho}$ and $\underline{\sigma} = \min_{m\in\{0,1\}} \sigma_{m,\rho}$, as well as $\underline{m}$ and $\overline{m}$ the corresponding messages. Given ρ, the level of retaliation r needed to induce $\underline{\sigma}$ rather than $\overline{\sigma}$ must satisfy

$$\overline{\sigma}(v_M(c,\overline{m}) - r) \leq \underline{\sigma}(v_M(c,\underline{m}) - r) \iff r \geq \left[\frac{\overline{\sigma}v_M(c,\overline{m}) - \underline{\sigma}v_M(c,\underline{m})}{\overline{\sigma} - \underline{\sigma}}\right]^+.$$

Since, v_M satisfies *weak preferences for the truth*, setting $\overline{m} = 1$ and $\underline{m} = 0$ maximizes the cost of inducing $\underline{\sigma}$ for the criminal agent and minimizes the cost of inducing $\underline{\sigma}$ for the non criminal agent. In addition weak preferences for the truth imply that whenever a criminal agent induces message $m = 0$, then he also induces $\rho = -$.

Given a profile σ satisfying the properties established above, we now establish the existence of a binary intervention profile $\widehat{\sigma} = (\widehat{\sigma}_m)_{m\in\{0,1\}}$ which can only increase the payoff of a non-criminal agent and can only decrease the payoff of a criminal agent. Specifically set $\widehat{\sigma}_0$ to be the equilibrium intervention rate against a non-criminal agent, and set $\widehat{\sigma}_1$ solving

$$\min_{\sigma_1\in\{\sigma_{m,\rho}|m\in\{0,1\},\ \rho\in\{-,+\}\}} \left\{\widehat{\sigma}_0(v_A(c=1) - k(r^1_\lambda));\ \sigma_1 v_A(c=1)\right\}$$

where r^1_λ is defined by (3).

By reducing the set of possible deviations from the monitor, the welfare of the non-criminal agent must increase. In addition, it reduces the equilibrium welfare of a criminal agent. □

Retaliation. The paper assumes that retaliation is equal to 0 following no intervention. This is an equilibrium result.

Lemma A.2. *For any crime decision c, it is optimal for the agent to retaliate only condi-*

tional on intervention: for any intervention policy σ, the agent's optimal retaliation policy is such that $r(i=0)=0$.

Proof. Taking a crime decision c as given, the agent's expected payoff under any retaliation profile $r:\{0,1\}\to[0,+\infty)$ is

$$\begin{aligned}\pi_A\times c+&[\text{prob}(m=0|r,c,\sigma)\sigma_0+\text{prob}(m=1|r,c,\sigma)\sigma_1][v_A(c)-k(r(1))]\\&-[1-\text{prob}(m=0|r,c,\sigma)\sigma_0-\text{prob}(m=1|r,c,\sigma)\sigma_1]k(r(0)).\end{aligned}$$

By committing not to retaliate (i.e. $\forall i,\ r(i)=0$), the agent can guarantee herself a payoff at least equal to $\sigma_1 v_A(c)$. Hence, if the agent engages in a positive amount of retaliation, it must be that

$$\begin{aligned}\sigma_1 v_A(c)&<\sigma_1[v_A(c)-k(r(1))]-\text{prob}(m=0)(\sigma_1-\sigma_0)[v_A(c)-k(r(1)))]\\&\quad-[1-\sigma_1+\text{prob}(m=0)(\sigma_1-\sigma_0)]k(r(0))\\&<\sigma_0(v_A(c)-k(r(1)).\end{aligned}\tag{10}$$

We now show that setting $r(0)$ to 0 increases the probability with which the monitor sends message $m=0$. Since it also reduces the cost of retaliation, it must increase the agent's payoff.

A monitor sends a message $m=0$ if and only if

$$-(1-\sigma_0)r(0)+\sigma_0[v_M(c,m=0)-r(1)]\geq-(1-\sigma_1)r(0)+\sigma_1[v_M(c,m=1)-r(1)].\tag{11}$$

Since $\sigma_1\geq\sigma_0$, it follows that whenever (11) holds for a retaliation profile such that $r(0)>0$, it continues to hold when $r(0)$ is set to 0, everything else being kept equal. It follows from (10) that the agent benefits from setting $r(0)=0$. □

Side payments. The paper assumes that the agent uses only retaliation to provide incentives to the monitor. It is immediate that the analysis of Sections 4 and 5 can be extended to allow for side-payments (modeled as $r(i)<0$), provided that there are no rewards given conditional on no-intervention, i.e. provided that $r(i=0)=0$.

We now clarify circumstance in which this last requirement holds endogenously. The cost of retaliation $k(\cdot)\geq 0$ is extended over $\mathbb{R}$ and decreasing in r over $r\in(-\infty,0)$. For simplicity, we assume that k is everywhere differentiable, except at $r=0$, where there is a kink: $k'(0^-)<0\leq k'(0^+)$. We define $k'_-=\sup_{r<0}k'(r)$ and $k'_+=\inf_{r>0}k'(r)$

Lemma A.3. *Whenever*

$$\frac{\sigma_0}{1-\sigma_0} < \left|\frac{k'_-}{k'_+}\right| \tag{12}$$

the agent's optimal retaliation strategy is such that $r(0) = 0$.

Whenever the marginal cost of retaliation is low, the cost of transfers is high, and the probability of intervention conditional on message $m = 0$ is low, it is optimal for the agent never to give out rewards when there is no intervention. The intuition is clear when σ_0 approaches 0: rewards must be paid on the equilibrium path, whereas successful threats are costly only off of the equilibrium path.

Proof. By an argument identical to that of Lemma A.2, it follows that at any optimal retaliation profile, $r(0) \leq 0$. Assume that $r(0) < 0$. We show that for $\epsilon > 0$ small enough, it is welfare improving for the agent to reduce rewards by ϵ conditional on $i = 0$, and increase retaliation by ϵ conditional on $i = 1$, i.e. to use retaliation policy $r^\epsilon(i) \equiv r(i) + \epsilon$.

It is immediate that this change in retaliation policy induces the same messages from monitors: payoffs have been shifted by a constant. For all $m \in \{0, 1\}$, we have

$$-(1-\sigma_m)r^\epsilon(0) + \sigma_m\left[v_M(c,m) - r^\epsilon(1)\right] = -(1-\sigma_m)r(0) + \sigma_m\left[v_M(c,m) - r(1)\right] - \epsilon,$$

which implies that the monitor's IC constraints are unchanged, and retaliation profile r^ϵ induces the same message profile as r.

We now show that using r^ϵ rather than r reduces the agent's expected retaliation costs. If the agent uses retaliation, he must induce message $m = 0$. The change in the agent's retaliation costs is given by

$$\begin{aligned} &-k'(r(0))(1-\sigma_0)\epsilon - k'(r(1))\sigma_0\epsilon + o(\epsilon) \\ &\qquad \geq -k'_-(1-\sigma_0)\epsilon - k'_+\sigma_0\epsilon + o(\epsilon). \end{aligned}$$

Condition (12) implies that this last expression is positive for ϵ small enough. Increasing retaliation is optimal for the agent. This concludes the proof. □

A.2 Multiple Monitors

Our analysis can be extended to settings with multiple monitors. Imagine that there are now L monitors indexed by $i \in \{1, \cdots, L\}$, each of which observes the agent's crime decision $c \in \{0, 1\}$ and can send a binary message $m_i \in \{0, 1\}$ to the principal. We denote by $\overrightarrow{m} \in \{0, 1\}^L$ the vector of messages sent by the monitors. We abuse notation and denote by

0 the message profile in which all monitors report $m_i = 0$, and by 1 the message profile in which all monitors report $m_i = 1$. An intervention policy σ is now a map $\sigma : \{0,1\}^L \to [0,1]$. For example, likelihood of intervention may be an affine function of the number of complaints, $\sigma_{\overrightarrow{m}} = \sigma_0 + (\sigma_1 - \sigma_0)\frac{1}{L}\sum_{i=1}^{L} m_i$. Alternatively, it may follow a threshold rule, with threshold $\theta \in \mathbb{N}$, i.e. $\sigma_{\overrightarrow{m}} = \sigma_0 + (\sigma_1 - \sigma_0)\mathbf{1}_{\sum_{i=1}^{L} m_i > \theta}$. For simplicity, we consider intervention policies such that for all $\overrightarrow{m} \neq 0$, $\sigma_{\overrightarrow{m}} > \sigma_0$.

As in Section 5, the agent and monitors have arbitrary types, except for the fact that Assumption 1 is common knowledge among players. We assume that each monitor i's value conditional on intervention $v_{i,M}$ depends only on c and her own message m_i. The agent now commits to a profile of vector-valued retaliation intensities $\overrightarrow{r} : \{0,1\} \to [0,+\infty)^L$ associated with a cost function $k(\overrightarrow{r})$ that is increasing in all components of $\overrightarrow{r}$.

The vector of monitors' types is denoted by $\overrightarrow{\tau_M} = (\tau_{i,M})_{i\in\{1,\cdots,L\}}$. Note that now, each monitor's type must include a belief over other monitors' types. Furthermore, the agent's belief over $\overrightarrow{\tau_M}$ is now a joint distribution over T_M^L. We denote by $\overrightarrow{\mathbf{m}} \in \left(\{0,1\}^{T_M}\right)^L$ message functions mapping profiles of types to a profile of messages. Note that for all $i \in \{1,\cdots,L\}$ monitor i's message profile $m_i(\tau_{i,M})$ is only a function of monitor i's type.

The main properties identified in Section 4 and 5 continue to hold: for messages to be informative, it must be that all likelihood ratios of intervention rates be bounded away from 0; when policies σ are ordered along a ray, message profiles change only when crime decisions change, and crime must decrease along a ray going away from the origin.

One difficulty is that there may now be multiple messaging equilibria among agents conditional on a given retaliation policy. We work under the assumption that given a retaliation policy, the agent is able to select the equilibrium that most benefits him, and that this equilibrium is unique. We think of the agent as selecting a message profile $\overrightarrow{\mathbf{m}}$ under constraints corresponding to the monitors' incentive compatibility conditions.

Lemma A.4. *If $\sigma_0 = 0$ then all agents that benefit from crime will be criminal, and induce message profile $\overrightarrow{m} = 0$.*

Proof. The proof is identical to that of Proposition 1. By setting $r(i = 0) = 0$ and $r(i = 1) = r$ arbitrarily high, the agent is able to induce message $\overrightarrow{m} = 0$ at no cost in equilibrium, which insures that there is no intervention. □

Given an interior intervention profile σ, define $\overrightarrow{\lambda} = \left(\frac{\sigma_{\overrightarrow{m}}}{\sigma_0}\right)_{\overrightarrow{m}\in\{0,1\}^L}$ the vector of likelihood ratios of intervention.

Lemma A.5. *Fix a vector of intervention ratios $\overrightarrow{\lambda}$ and consider the ray of intervention policies $\{\sigma_0 \overrightarrow{\lambda}$ for $\sigma_0 \in [0,1]\}$. Along this ray the following properties hold:*

(i) conditional on a crime decision c, the message function $\overrightarrow{\mathbf{m}}$ that a given agent chooses to induce is constant along the ray;

(ii) the agent's decision to be criminal is decreasing in σ_0 along the ray.

Proof. Let us begin with point (i). Conditional on a crime decision $c \in \{0,1\}$, for any message profile $\overrightarrow{\mathbf{m}}$, we define the normalized cost $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\sigma)$ of inducing message function $\overrightarrow{\mathbf{m}}$ as

$$\begin{aligned} K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\sigma) =& \frac{1}{\sigma_0} \inf_{r\in[0,+\infty)} \int_{T_M^L} \sigma_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} k(r) \mathrm{d}\Phi(\overrightarrow{\tau_M}|\tau_A) \\ & \text{s.t. } \forall \overrightarrow{\tau_M} = (\tau_{i,M})_{i\in\{1,\cdots,L\}}, \ \ (m_i)_{i\in\{1,\cdots,L\}} = \overrightarrow{\mathbf{m}}(\tau_{i,M}) \text{ satisfies} \\ & \forall i \in \{1,\cdots,L\}, \\ & \mathbb{E}\left[\sigma_{(m_i,\overrightarrow{\mathbf{m}}_{-i})} v_{i,M}(m_i,c) - r_i \middle| m_i, \overrightarrow{\mathbf{m}}_{-i}, c\right] \geq \mathbb{E}\left[\sigma_{(\neg m_i,\overrightarrow{\mathbf{m}}_{-i})} v_{i,M}(\neg m_i,c) - r_i \middle| \neg m_i, \overrightarrow{\mathbf{m}}_{-i}, c\right]. \end{aligned}$$

It follows from inspection that $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}$ is a function of $\overrightarrow{\lambda}$ only. By convention $K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}$ is set to $+\infty$ whenever message function $\overrightarrow{\mathbf{m}}$ is not implementable. Given a crime decision c, the agent chooses to induce the message function $\overrightarrow{\mathbf{m}}$ solving

$$\sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(c) \right] - K^{\tau_A}_{c,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\}.$$

It follows that the optimal message function induced by the agent is a function of $\overrightarrow{\lambda}$ only, and, conditional on a crime decision, remains constant along rays.

We now turn to point (ii). An agent chooses to be non-criminal if and only if

$$\pi_A + \sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(1) \right] - K^{\tau_A}_{1,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\} \leq \sigma_0 \max_{\overrightarrow{\mathbf{m}}} \left\{ \mathbb{E}\left[\overrightarrow{\lambda}_{\overrightarrow{\mathbf{m}}(\overrightarrow{\tau_M})} v_A(0) \right] - K^{\tau_A}_{0,\overrightarrow{\mathbf{m}}}(\overrightarrow{\lambda}) \right\}. \quad (13)$$

Since $\pi_A \geq 0$ it follows that whenever (13) holds for σ_0, it must also hold for all $\sigma_0' \geq \sigma_0$. This proves point (ii). □

An implication is that changes in reporting patterns along a ray can be assigned to changes in crime. Consider two policies σ^α, σ^β such that $\overrightarrow{\lambda}^\alpha = \overrightarrow{\lambda}^\beta = \overrightarrow{\lambda}$ and $\sigma_0^\alpha < \sigma_0^\beta$. For any function $X : \overrightarrow{m} \in \{0,1\}^L \mapsto x \in \mathbb{R}^n$ computing a summary statistic of messages, denote by $\widehat{\mu}^\sigma_X$ the distribution over $x \in X(\{0,1\}^L)$ defined by $\widehat{\mu}^\sigma_X(x) = \int_T \mathbf{1}_{X(\overrightarrow{m}^*(\sigma,\tau))=x} \mathrm{d}\mu_T(\tau)$, where

$\overrightarrow{m}^*(\sigma, \tau)$ is the equilibrium vector of messages for a realized profile of types $\tau = (\tau_A, \overrightarrow{\tau}_M)$ given intervention policy σ. Given policies σ^α, σ^β, let D denote the distance between message distributions induced by σ^α and σ^β defined by

$$D \equiv \frac{1}{2} \sum_{x \in X(\{0,1\}^L)} |\mu_X^{\sigma^\beta}(x) - \mu_X^{\sigma^\alpha}(x)|.$$

Note that D can be computed from message data alone. Observing that D is one of the expression for the total variation distance between $\mu_X^{\sigma^\alpha}$ and $\mu_X^{\sigma^\beta}$, we have that

$$D = \max_{X' \subset X} |\mu_X^{\sigma^\alpha}(X') - \mu_X^{\sigma^\beta}(X')|$$

Proposition 6 extends as follows.

Proposition A.1 (inference)**.** *For all possible true distributions μ_T, we have that*

$$\int_{T_A} \left[c^*(\sigma^\alpha, \tau_A) - c^*(\sigma^\beta, \tau_A)\right] d\mu_T(\tau_A) \geq D$$

which implies that D is a lower bound for the mass $\int_{T_A}[1 - c^*(\sigma^\beta, \tau_A)]d\mu_T(\tau_A)$ of honest agents at policy σ^β as well as a lower bound for the mass $\int_{T_A} c^*(\sigma^\alpha, \tau_A)d\mu_T(\tau_A)$ of criminal agents at policy σ^α.

Proof. The proof is essentially identical to that of Proposition 6. From Proposition A.5, it follows that

$$\begin{aligned}
\int_{T_A} [c^*(\sigma^\alpha, \tau_A) - c^*(\sigma^\beta, \tau_A)] \mathrm{d}\mu_T(\tau_A) &\geq \int_{T_A} \mathbf{1}_{\overrightarrow{\mathbf{m}}_{\tau_A}^*(\sigma^\alpha) \neq \overrightarrow{\mathbf{m}}_{\tau_A}^*(\sigma^\beta)} \mathrm{d}\mu_T(\tau_A) \\
&\geq \int_{T_A} \mathbf{1}_{\mu_{X|\tau_A}^{\sigma^\alpha} \neq \mu_{X|\tau_A}^{\sigma^\beta}} \mathrm{d}\mu_T(\tau_A) \geq \int_{T_A} \max_{X' \subset X} |\mu_{X|\tau_A}^{\sigma^\alpha}(X') - \mu_{X|\tau_A}^{\sigma^\beta}(X')| \mathrm{d}\mu_T(\tau_A) \\
&\geq \max_{X' \subset X} |\mu_X^{\sigma^\alpha}(X') - \mu_X^{\sigma^\beta}(X')| = D
\end{aligned}$$

which concludes the proof. □

B Proofs

B.1 Proofs for Section 4

Proof of Proposition 1. We begin with point (i). Note that 0 is the highest payoff the principal can attain. Under intervention policy $\sigma_0 = 0$, $\sigma_1 = 1$, Assumption 2 implies that

it is optimal for the agent to choose $c = 0$. As a result, there will be no intervention on the equilibrium path. Hence the principal attains her highest possible payoff, and $\sigma_0 = 0$, $\sigma_1 = 1$ is indeed an optimal intervention policy.

Let us turn to point (ii). Consider policies σ such that $\lambda = \frac{\sigma_1}{\sigma_0} > 2$ and the retaliation profile under which the agent retaliates by an amount $r \equiv 2v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$. Retaliation level r is chosen so that whenever $c = 1$, the monitor prefers to send message $m = 0$. Indeed, the monitor prefers to send message $m = 0$ if and only if

$$\begin{aligned} &\sigma_1[v_M(c = 1, m = 1) - r] \leq \sigma_0[v_M(c = 1, m = 0) - r] \\ \iff &r \geq \frac{\lambda v_M(c = 1, m = 1) - v_M(c = 1, m = 0)}{\lambda - 1}. \end{aligned} \tag{14}$$

Noting that the right-hand side of (14) is decreasing in λ and that $\lambda > 2$, we obtain that the monitor indeed sends message m whenever $r \geq 2v_M(c = 1, m = 1) - v_M(c = 1, m = 0)$.

It follows that the expected payoff of a criminal agent is

$$\pi_A + \sigma_0[v_A(c = 1) - k(r)] \geq \pi_A + \frac{1}{\lambda}[v_A(c = 1) - k(r)].$$

Since $\pi_A > 0$, it follows that this strategy guarantees the agent a strictly positive payoff for λ sufficiently large. Given that the highest possible payoff for an agent choosing $c = 0$ is equal to 0, it follows that for λ large enough the agent will engage in crime.

Given decision $c = 1$, we now show that the agent will also use retaliation. Under no retaliation the agent obtains an expected payoff equal to $\pi_A + \sigma_1 v_A(c = 1)$. Under the retaliation strategy described above, the agent obtains a payoff equal to $\pi_A + \frac{\sigma_1}{\lambda}[v_A(c = 1) - k(r)]$. Since $v_A(c = 1) < 0$ it follows that for λ large enough, it is optimal for the agent to commit to retaliation. ■

Proof of Lemma 1. Lemma 1 is an immediate implication from expression (3) and Assumption 1. ■

Proof of Lemma 2. It follows from expression (4) that a criminal agent will induce message $m = 0$ if and only if

$$k(r_\lambda^1) \leq -(\lambda - 1)v_A(1).$$

Since $v_A(1) \leq 0$ and r_λ^1 is decreasing in λ it follows that the optimal message manipulation strategy of the agent takes a threshold form: there exists λ_1 such that the agent induces

message $m = 0$ for all $\lambda \geq \lambda_1$.

If $c = 0$ and the monitor is not malicious, the agent can induce message $m = 0$ without committing to retaliation. Consider now the case where the monitor is malicious: $v_M(c = 0, m = 1) > 0$. As λ approaches 1, it follows from (3) that the retaliation r needed to induce $m = 0$ goes to 0. In turn, as λ approaches $+\infty$, the equilibrium cost of retaliation $\sigma_0 k(r)$ also approaches 0. Hence as λ approaches either 0, or $+\infty$, an honest agent induces message $m = 0$. We now show that there exist λ and k such that the agent allows message $m = 1$. Pick any value λ large enough such that $r^0_\lambda > 0$. Then for a cost of retaliation k sufficiently high, the agent will abstain from retaliation threats and induce message $m = 1$. ■

Proof of Proposition 2. Assume that there are no malicious monitors and $v_A(c = 0) = 0$. Condition (6) implies that $\lambda_\alpha \geq \lambda_\beta$. Hence $r^1_{\lambda_\alpha} \leq r^1_{\lambda_\beta}$. This implies that

$$\pi_A + \max\{\sigma^\alpha_1 v_A(c = 1),\, \sigma^\alpha_0[v_A(c = 1) - k(r^1_{\lambda_\alpha})]\} \\ \geq \pi_A + \max\{\sigma^\beta_1 v_A(c = 1),\ \sigma^\beta_0[v_A(c = 1) - k(r^1_{\lambda_\beta})]\}.$$

Hence $c^\beta \leq c^\alpha$.

Consider the case where $m^\alpha = 1$. Since there are no malicious monitors, it must be that $c^\alpha = 1$. Furthermore, λ_α must be low enough that $\lambda_\alpha v_A(1) \geq v_A(1) - k(r^1_{\lambda_\alpha})$. Hence, it must also be that $\lambda_\beta v_A(1) \geq v_A(1) - k(r^1_{\lambda_\beta})$, which implies that if $c^\beta = 1$, then $m^\beta = 1$. Since $m^\beta = 0$, it must be that $c^\beta = 0$.

Consider now the case where monitors may be malicious and $v_A(c = 0)$ may be strictly negative. The agent's decision to engage in crime is determined by condition (5). We first provide an example where inequality $c^\beta \leq c^\alpha$ fails to hold. Assume that $v_A(c = 0) < 0$, but maintain the assumption that the monitor is non-malicious. This implies that $r^0_\lambda = 0$ for all λs: an honest agent induces message $m = 0$ without threats. In turn we can pick v_M and k such that $k(r^1_\lambda)$ is large, so that a criminal agent would prefer to induce message $m = 1$. Pick π_A such that under σ^α, the agent is almost indifferent between $c = 0$ and $c = 1$, but prefers $c = 0$ by an arbitrarily small amount. Then pick σ^β such that $\sigma^\beta_0 > \sigma^\alpha_0$ and $\sigma^\beta_1 = \sigma^\beta_0$. This policy satisfies (6), keeps the payoff of a criminal agent the same, and strictly decreases the payoff of a non-criminal agent. As a result $c^\beta = 1 > c^\alpha = 0$.

In turn, consider an environment in which the monitor is malicious ($v_M(c = 0, m) > 0$), and $v_A(c = 0) < 0$. This implies that to induce message $m = 0$, an honest agent will threaten to retaliate a positive amount. Furthermore, we know from (3) that we can find payoffs v_M

such that $r^0_{\lambda_\beta} < r^0_{\lambda_\alpha}$. This implies that on can find a convex and increasing cost function k such that

$$\sigma_1^\alpha v_A(0) > \sigma_0^\alpha(v_A(0) - k(r^0_{\lambda_\alpha}))$$
$$\sigma_1^\beta v_A(0) < \sigma_0^\beta(v_A(0) - k(r^0_{\lambda_\beta})).$$

In addition, one can pick π_A and $v_A(c=1)$ to ensure that under both σ^α and σ^β, the agent chooses $c=0$. This yields that $m^\alpha = 1$, $m^\beta = 0$ while $c^\alpha = c^\beta = 0$. ■

Proof of Proposition 3. First, note that the optimal policy must induce $m = 0$. If not the principal may as well set $\sigma_0^* = \sigma_1^*$. This implies that the optimal policy solves

$$\min_{\sigma_0,\sigma_1} \sigma_0$$
$$s.t. \quad \sigma_0[v_A(c=0) - k(r^0_\lambda)] \geq \pi_A + \max\{\sigma_1 v_A(c=1),\ \sigma_0[v_A(c=1) - k(r^1_\lambda)]\} \tag{15}$$
$$\sigma_0[v_A(c=0) - k(r^0_\lambda)] \geq \sigma_1 v_A(c=0). \tag{16}$$

Constraint (15) must be binding, otherwise $\sigma_0 = \sigma_1 = 0$ would be a solution. If the monitor is not malicious, $r^0_\lambda = 0$. It follows that (16) is not binding. If the monitor is malicious, (16) may or may not be binding.

Assume (16) is not binding. It must be that $\sigma_1 v_A(c=1) = \sigma_0[v_A(c=1) - k(r^1_\lambda)]$ at the optimal policy. Assume instead that $\sigma_1 v_A(c=1) > \sigma_0[v_A(c=1) - k(r^1_\lambda)]$. Since $v_A(c=0) - k(r^0_\lambda) \leq 0$ this implies that one could reduce σ_0^*. Assume instead that $\sigma_1 v_A(c=1) < \sigma_0[v_A(c=1) - k(r^1_\lambda)]$. Reducing σ_1 strictly relaxes (15), allowing for a reduction in σ_0. Hence $\frac{\sigma_1^*}{\sigma_0^*} = \lambda_1$. If instead, (16) is binding, then, by definition $\frac{\sigma_1^*}{\sigma_0^*} \in \Lambda_0$.

This proves the first two points of the proposition. We now establish that the optimal intervention is generically interior.

We know that $\sigma_0 \in (0,1)$ from Proposition 1 and the assumption that $\pi_A + v_A(c=1) < v_A(c=0)$. Assume that $\sigma_1 = 1$ and (16) is not binding. It must be the case that $\pi_A + v_A(c=1) = \sigma_0[v_A(c=0) - k(r^0_\lambda)] \geq v_A(c=0)$. This contradicts Assumption 2. Hence, if (16) is not binding it must be the case that $\sigma_1 < 1$.

Assume (16) is binding. Then we have two candidate solutions. The first candidate is defined by taking (16) with equality and by $\pi_A + \sigma_1 v_A(c=1) = \sigma_0[v_A(c=0) - k(r^0_\lambda)]$. A similar argument establishes that these two equations cannot be solved by $\sigma_1 = 1$ and

respect Assumption 2: they imply $\pi_A + v_A(c=1) = \sigma_0[v_A(c=0) - k(r^0_\lambda)] = v_A(c=0)$. Hence this first candidate is necessarily interior. The second candidate is defined by taking (16) with equality and by $\pi_A + \sigma_0[v_A(c=1) - k(r^1_\lambda)] = \sigma_0[v_A(c=0) - k(r^0_\lambda)]$. These are two equations in two unknowns, and a solution with $\sigma_1 = 1$ is only possible in a knife-edge case. This would only be the overall optimal policy if this candidate σ_0 is smaller than the σ_0 in the first candidate solution. ■

Proof of Proposition 4. Point (i) follows from the fact that increasing $v_A(c=0)$ and decreasing $v_A(c=1)$ both relax condition (8).

Point (ii) follows from the fact that increasing $v_M(c=1,m=1) - v_M(c=1,m=0)$ weakly increases r^1_λ, which weakly reduces the right-hand side of (8).

Point (iii) follows from the fact that increasing $v_M(c=0,m=0) - v_M(c=0,m=1)$ weakly decreases r^0_λ, which weakly increases the left-hand side of (8). ■

B.2 Proofs for Section 5

Proof of Proposition 5. Fix σ and a distribution μ_T such that $\int_T m^*(\sigma,\tau)\mathrm{d}\mu_T(\tau) = M \in [0,1]$. Fix $C \in [0,1]$. We show that there exists $\widehat{\mu}_T$ such that $\int_T m^*(\sigma,\tau)\mathrm{d}\widehat{\mu}_T(\tau) = M$ and $\int_T c^*(\sigma,\tau_A)\mathrm{d}\mu_T(\tau) = C$.

It is sufficient to work with type spaces such that the agent knows the type of the monitor, provided we allow payoffs to be correlated. A possible environment is as follows. With probability C, the agent gets a strictly positive payoff $\pi_A > 0$ from crime. Conditional on $\pi_A > 0$, with probability γ, the monitor has positive value for intervention against criminal agents, i.e. $v_M(c=1,m) = v > 0 = v_M(c=0,m)$; with probability $1-\gamma$, the monitor has a low value for intervention on criminal agents: $v_M(c,m) = 0$ for all $(c,m) \in \{0,1\}^2$. The cost of retaliation for the agent is such that k is convex and strictly increasing. For $v_A(c=1) > 0$ appropriately low, it will be optimal for the agent to be criminal, and commit to an arbitrarily low retaliation profile so that the monitor with a low value for intervention sends message $m = 0$ and the monitor with a high value for intervention sends message $m = 1$.

With complementary probability $1 - C$ the agent gets a payoff $\pi_A = 0$ from crime and has an arbitrarily high cost of retaliation. The agent's values upon intervention are such that $v_A(c=1) < v_A(c=0)$. With probability ν, the monitor has negative value for intervention against a non-criminal agent $v_M(c=0,m) < 0$. With probability $1-\nu$ the monitor gets

a positive payoff $v > 0$ from intervention against the agent, regardless of his crime status. For v and a cost of retaliation k sufficiently high, the agent will choose not to be criminal, the non-malicious monitor will send message $m = 0$, and the malicious monitor will send message $m = 1$.

For any $C \in [0,1]$ and $M \in [0,1]$, one can find γ and ν such that $C\gamma + (1-C)\nu = M$. This concludes the proof. ■

Proof of Proposition 6. The proof is a special case of Proposition A.1. We provide a brief outline.

Given an induced message function $\mathbf{m} : \tau_M \mapsto \mathbf{m}(\tau_M) \in \{0,1\}$, we define the normalized cost $K^{\tau_A}_{c,\mathbf{m}}(\sigma)$ of inducing message function $\mathbf{m}$ as

$$\begin{aligned} K^{\tau_A}_{c,\mathbf{m}}(\sigma) = & \frac{1}{\sigma_0} \inf_{r \in [0,+\infty)} \int_{T^L_M} \sigma_{\mathbf{m}}(\tau_M) k(r) \mathrm{d}\Phi(\tau_M | \tau_A) \\ & \text{s.t.} \quad \forall \tau_M, \quad m = \mathbf{m}(\tau_{i,M}) \text{ satisfies} \\ & \forall i \in \{1, \cdots, L\}, \\ & \sigma_m(v_M(m,c) - r) \geq \sigma_{\neg m}(v_M(\neg m, c) - r). \end{aligned}$$

It follows from inspection that $K^{\tau_A}_{c,\mathbf{m}}$ is a function of λ only. By convention $K^{\tau_A}_{c,\mathbf{m}}$ is set to $+\infty$ whenever message function $\mathbf{m}$ is not implementable. For $m \in \{0,1\}$, let $\lambda_m \equiv \frac{\sigma_m}{\sigma_0}$. Given a crime decision c, the agent chooses to induce the message function $\mathbf{m}$ solving

$$\sigma_0 \max_{\mathbf{m}} \left\{ \mathbb{E}\left[\lambda_{\mathbf{m}(\tau_M)} v_A(c) \right] - K^{\tau_A}_{c,\mathbf{m}}(\lambda) \right\}.$$

It follows that conditional on a crime decision, the optimal message function induced by the agent is a function of λ only.

An agent chooses to be non-criminal if and only if

$$\pi_A + \sigma_0 \max_{\mathbf{m}} \left\{ \mathbb{E}\left[\lambda_{\mathbf{m}(\tau_M)} v_A(1) \right] - K^{\tau_A}_{1,\mathbf{m}}(\lambda) \right\} \leq \sigma_0 \max_{\mathbf{m}} \left\{ \mathbb{E}\left[\lambda_{\mathbf{m}(\tau_M)} v_A(0) \right] - K^{\tau_A}_{0,\mathbf{m}}(\lambda) \right\}. \tag{17}$$

Since $\pi_A \geq 0$ it follows that whenever (17) holds for σ_0, it must also hold for all $\sigma_0' \geq \sigma_0$. Hence $c^\alpha \geq c^\beta$ and $m^\alpha \neq m^\beta \Rightarrow c^\alpha \neq c^\beta$.

This implies that for all τ_A, $c^\alpha - c^\beta \geq |m^\alpha - m^\beta|$. Integrating over μ_T and Jensen's inequality yields Proposition 6. ■

Proof of Corollary 2. The first inequality follows from the fact that $\widehat{C}(\sigma)$ is an upper bound to the amount of crime at policy σ.

Data-driven heuristic policy $\widehat{\sigma}(\mu_T)$ is weakly undominated because it is exactly optimal whenever payoffs are common knowledge between the agent and the monitor, satisfy Assumption 1, Assumption 2, with the adjustment that $v_A(c = 0) = 0$, and monitors are non-malicious, i.e. $v_M(m, c = 0) \leq 0$. ■

References

Bac, M. (2009): "An Economic Rationale for Firing Whistleblowers," *European Journal of Law and Economics*, 27, 233–256.

Banerjee, A., S. Chassang, S. Montero, and E. Snowberg (2017): "A theory of experimenters," Tech. rep., National Bureau of Economic Research.

Banerjee, A. and E. Duflo (2006): "Addressing Absence," *Journal of Economic Perspectives*, 20, 117–132.

Basu, K., K. Basu, and T. Cordella (2014): "Asymmetric punishment as an instrument of corruption control," *World Bank Policy Research Working Paper*.

Becker, G. S. (1968): "Crime and Punishment: An Economic Approach," *Journal of Political*, 76, 169–217.

Becker, G. S. and G. J. Stigler (1974): "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, 3, 1–18.

Berman, E., J. N. Shapiro, and J. H. Felter (2011): "Can hearts and minds be bought? The economics of counterinsurgency in Iraq," *Journal of Political Economy*, 119, 766–819.

Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan (2007): "Obtaining a driver's license in India: an experimental approach to studying corruption," *The Quarterly Journal of Economics*, 122, 1639–1676.

Carroll, G. (2013): "Robustness and Linear Contracts," *Stanford University Working Paper*.

Chassang, S. (2013): "Calibrated incentive contracts," *Econometrica*, 81, 1935–1971.

Chassang, S. and G. Padro i Miquel (2014): "Corruption, Intimidation, and Whistleblowing: a Theory of Inference from Unverifiable Reports," Tech. rep., National Bureau of Economic Research.

Chassang, S., G. Padró i Miquel, and E. Snowberg (2012): "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 102, 1279–1309.

Dal Bó, E. (2007): "Bribing voters," *American Journal of Political Science*, 51, 789–803.

Ederer, F., R. Holden, and M. Meyer (2017): "Gaming and Strategic Opacity in Incentive Provisin," Tech. rep., Cowles Foundation Discussion Paper.

Eeckhout, J., N. Persico, and P. Todd (2010): "A theory of optimal random crackdowns," *The American Economic Review*, 100, 1104–1135.

Ensminger, J. (2013): "Inside Corruption Networks: Following the Money in Community Driven Development," *Unpublished manuscript, Caltech*.

Faure-Grimaud, A., J.-J. Laffont, and D. Martimort (2003): "Collusion, delegation and supervision with soft information," *The Review of Economic Studies*, 70, 253–279.

Frankel, A. (2014): "Aligned delegation," *The American Economic Review*, 104, 66–83.

Fudenberg, D. and D. K. Levine (1992): "Maintaining a reputation when strategies are imperfectly observed," *The Review of Economic Studies*, 59, 561–579.

Galula, D. (1964): "Counter-insurgency warfare: theory and practice," .

Ghosh, A. and A. Roth (2010): "Selling privacy at auction," *Arxiv preprint arXiv:1011.1375*.

Gradwohl, R. (2012): "Privacy in Implementation," .

Hartline, J. D. and T. Roughgarden (2008): "Optimal Mechanism Design and Money Burning," in *Symposium on Theory Of Computing (STOC)*, 75–84.

Heyes, A. and S. Kapur (2009): "An Economic Model of Whistle-Blower Policy," *Journal of Law, Economics and Organizations*, 25, 157–182.

Hurwicz, L. and L. Shapiro (1978): "Incentive structures maximizing residual gain under incomplete information," *The Bell Journal of Economics*, 9, 180–191.

Izmalkov, S., M. Lepinski, and S. Micali (2011): "Perfect implementation," *Games and Economic Behavior*, 71, 121–140.

Kaplan, S. E., K. Pany, J. A. Samuels, and J. Zhang (2009): "An Examination of the Effects of Procedural Safeguards on Intentions to Anonymously Report Fraud," *Auditing: A Journal of Practice & Theory*, 28, 273–288.

Kaplan, S. E. and J. J. Schultz (2007): "Intentions to Report Questionable Acts: An Examination of the Influence of Anonymous Reporting Channel, Internal Audit Quality, and Setting," *Journal of Business Ethics*, 71, 109–124.

Karlan, D. and J. Zinman (2009): "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, 77, 1993–2008.

Kilcullen, D. (2009): *The Accidental Guerrilla: Fighting Small Wars in the Midst of a Big One*, Oxford University Press.

Laffont, J. and D. Martimort (1997): "Collusion under asymmetric information," *Econometrica: Journal of the Econometric Society*, 875–911.

Laffont, J.-J. and D. Martimort (2000): "Mechanism design with collusion and correlation," *Econometrica*, 68, 309–342.

Madarász, K. and A. Prat (2010): "Screening with an Approximate Type Space," *Working Paper, London School of Economics*.

Makowsky, M. and S. Wang (2018): "Embezzlement, Whistle-Blowing, and Organizational Architecture: An Experimental Investigation," *Journal of Economic Behavior & Organization*, 147, 58–75.

Miceli, M. P., M. Rehg, J. P. Near, and C. C. Ryan (1999): "Can Laws Protect Whistle-Blowers? Results of a Naturally Occuring Field Experiment," *Work and Occupations*, 26, 129–151.

Mookherjee, D. and I. Png (1995): "Corruptible Law Enforcers: How Should They Be Compensated?" *Economic Journal*, 105, 145–59.

Myerson, R. B. (1986): "Multistage games with communication," *Econometrica: Journal of the Econometric Society*, 323–358.

Nagl, J. A. (2002): *Counterinsurgency lessons from Malaya and Vietnam: Learning to eat soup with a knife*, Praeger Publishers.

Nagl, J. A., J. F. Amos, S. Sewall, and D. H. Petraeus (2008): *The US Army/Marine Corps Counterinsurgency Field Manual*, University of Chicago Press.

Near, J. and M. P. Miceli (1995): "Effective Whistleblowing," *Academy of Management Review*, 679–708.

Nissim, K., C. Orlandi, and R. Smorodinsky (2011): "Privacy-aware mechanism design," *Arxiv preprint arXiv:1111.3350*.

Olken, B. (2007): "Monitoring corruption: evidence from a field experiment in Indonesia," *Journal of Political Economy*, 115, 200–249.

Ortner, J. and S. Chassang (2018): "Making collusion hard: asymmetric information as a counter-corruption measure," *Journal of Political Economy*.

Polinsky, A. M. and S. Shavell (2000): "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature*, 36, 45–76.

——— (2001): "Corruption and Optimal Law Enforcement," *Journal of Public Economics*, 81, 1–24.

Prendergast, C. (2000): "Investigating Corruption," *working paper, World Bank development group*.

Punch, M. (2009): *Police Corruption: Deviance, Accountability and Reform in Policing*, Willan Publishing.

Rahman, D. (2012): "But who will monitor the monitor?" *The American Economic Review*, 102, 2767–2797.

Ricks, T. E. (2006): *Fiasco: the American military adventure in Iraq*, Penguin.

——— (2009): *The gamble: General David Petraeus and the American military adventure in Iraq, 2006-2008*, Penguin.

Segal, I. (2003): "Optimal pricing mechanisms with unknown demand," *The American economic review*, 93, 509–529.

Spagnolo, G. (2008): "Leniency and Whistleblowers in Antitrust," in *Handbook of Antitrust Economics*, ed. by P. Buccirossi, MIT Press.

Tirole, J. (1986): "Hierarchies and bureaucracies: On the role of collusion in organizations," *Journal of Law, Economics, and Organizations*, 2, 181.

US Army (2006): "Field Manual 2-22.3: Human Intelligence Collector Operations," *Department of the Army, Washington DC*.

US Army and Marine Corps (2006): "Field Manual 3-24: Counterinsurgency," *Department of the Army, Washington DC*.

Warner, S. L. (1965): "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, 60, 63–69.