Monitoring, Disclosure, and Retaliation

Henrique Castro-Pires

July 25, 2023

Abstract

We analyze the effects of retaliation on optimal contracts in a hierarchy consisting of a principal, a monitor, and an agent. With probability m, the monitor observes a signal about the agent's effort and decides whether to reveal it to the principal. With probability $(1-m)$, the monitor is uninformed. The agent retaliates against the monitor and the principal whenever the disclosed signal reduces his compensation from the no-disclosure benchmark. We show that the principal's optimal contracting problem can be divided into two steps: first, an information acquisition stage. The principal chooses how much retaliation to tolerate, and more retaliation generates more informative signals (in the Blackwell sense) about the agent's effort. Second, given the information acquired, the principal designs the optimal payment schemes, which pool moderately (potentially all) bad agent's performances with the uninformative signal realization. The empirical literature documents that supervisors are reluctant to provide poor ratings and that performance reports are often inflated and compressed. We show that such a pattern can stem from retaliation concerns.

1 Introduction

When designing incentive schemes, firms often rely on supervisors to monitor and evaluate their agents. However, the supervisor might not reveal all the information she possesses about her subordinate's actions. There is substantial empirical evidence that supervisors are reluctant to disclose their subordinates' poor performance[1]. This paper studies optimal contracts when disclosing poor performance is costly to the supervisor and/or the organization.

The relationship between supervisor-subordinate is not without conflict. Often, agents who receive bad evaluations retaliate against supervisors and their organizations. The retaliation can take several forms, such as: talking back to the supervisor, spreading rumors, intentionally working slower, or sabotaging production. For instance, Stud Terkel - in his classic book *Working* (1972) - interviews Mike, a steelworker from Cicero, Illinois (Terkel (1972), p. xxxi-xxxv). Mike describes his conflicts with his foreman and how he retaliates by "not even try[ing] to think ", by refusing to say "Yes, sir" to the

[1] See Bol (2011) for supporting empirical evidence.

boss, and by occasionally "putting a dent in [the steel]" [2]. Sprouse (1992) collects several examples of employee retaliatory behavior in many different contexts, that range from jamming machinery in an industrial plant to a waitress serving spoiled food to ruin her employer's business. Moreover, extensive literature in accounting, psychology, and management examines retaliation in the workplace. Greenberg (1990) documents employee theft doubling after a pay decrease of 15%. In a sample of 240 manufacturing employees, Skarlicki and Folger (1997) documents a variety of retaliatory behaviors such as "gossiped about his or her boss", "left an unnecessary mess", and "talked back to his or her boss", among others[3].

In this article, we study the effect of an agent's retaliatory behavior on optimal incentive contracts. We characterize optimal contracts in a hierarchy consisting of a principal, a monitor, and an agent. The monitor privately observes a verifiable signal about the agent's performance with probability m and decides whether or not to disclose the information to the principal. The monitor is uninformed with probability $(1-m)$ and has no evidence to disclose. After receiving his compensation, the agent decides whether or not to retaliate against the monitor and the principal. We assume the agent retaliates whenever his payment is less than he would have been paid if no evidence was revealed.

We show that the optimal agent's wage schedule can be described by a reference payment in case of no disclosure and punishments/rewards depending on the disclosed performance. Depending on the magnitude of the retaliation losses, the contract offered to the agent takes one of two possible forms: the *carrot-only contract*, in which the agent is never paid below the reference value, or the *stick-and-carrot contract*, in which the agent might be paid above or below the reference value. Moreover, this reference value is endogenously determined and reflects the informational content of no-disclosure. That is, the less frequently the principal uses punishments, the more no disclosure is associated with low performance. The stronger the association between no-disclosure and low performance (hence, low effort), the smaller the optimal payment under no-disclosure.

Classic principal-agent models usually assume the signal structure observed by the principal — conditional on the agent's effort — to be exogenous. However, when there is the possibility of retaliation, the information the principal observes depends on what signal realizations the monitor is willing to reveal. For instance, the monitor may refrain from disclosing information that reduces the agent's compensation to avoid retaliation losses. In this case, the evidence the principal observes about the agent's action depends directly on the compensation scheme in place. When designing incentive schemes, the principal must not only choose the compensation plan for a given information structure but also consider what information is revealed for each chosen compensation scheme.

We find the optimal contracts in such a setting using three main steps: first, we show that retaliation must take a cutoff form. That is, if the evidence revealed is low enough, the agent retaliates. Second, we fix an amount of retaliation the principal is willing to tolerate and solve for the optimal compensation scheme. Third, we solve for the optimal amount of retaliation. On the one hand, the more retaliation the principal tolerates, the more information about the agent's performance she can use, and the cheapest it

[2] Both examples are also described by Akerlof and Kranton (2005).

[3] evidence of employee retaliation can be found at Greenberg (1990), Aquino and Douglas (2003), Krueger and Mas (2004), Mas (2008), Charness and Levine (2010), and Coviello et al. (2022).

is to compensate the agent for effort.

On the other hand, dealing with retaliation is costly. Either by directly suffering retaliatory damages and/or by compensating the monitor for her retaliation losses. We then show that the principal's choice of the retaliation cutoff is formally equivalent to an information acquisition problem. If the principal tolerates more retaliation — that is, if she implements a higher retaliation cutoff — she observes more informative signals about the agent's effort. However, tolerating more retaliation brings additional retaliatory costs.

The main insight is that the principal faces a trade-off between using more information and punishing poor performances but enduring retaliation losses or providing only positive incentives but wasting information. The stick-and-carrot contract allows the principal to use more information about the agent's effort, while the carrot-only contract avoids retaliation. In order to build intuition, it is helpful to consider two extreme situations. The first is the case where the losses caused by the agent's retaliation are negligible. In such a case, the principal does not care about retaliation and uses all information available. In particular, she rewards good performances and punishes bad ones. On the other extreme, when the retaliation costs are sufficiently high, the principal prefers to eliminate retaliatory behavior completely. Hence, she provides incentives only through rewards and disregards any information about bad performance.

Beyond the use of carrots and sticks, we show that the optimal mechanism can also be implemented by not providing incentives for the monitors to reveal their information fully. Moreover, we show that the optimal mechanism is consistent with allowing two performance review patterns extensively documented by the empirical literature and often perceived as sub-optimal: leniency and centrality. Leniency refers to the fact that managers often do not report their employees' low performances, while centrality refers to the observation that payments are less dispersed than realized performances. By letting managers hide moderately (potentially all) bad performances, the principal mitigates the costs associated with retaliation while generating lenient and centrally concentrated performance reviews.

Most of the analysis is conducted while keeping the monitoring effort as an exogenous variable. In Section 5, we extend the analysis to the case in which monitoring is costly to the monitor. In this case, the principal must also incentivize the monitor to acquire evidence about the agent's performance. In the optimal mechanism, the principal always requires the monitor to reveal all the information obtained and provides the monitor a bonus when evidence is disclosed. However, not all evidence is revealed to the agent. As with free monitoring, the principal hides moderately (potentially all) bad performances from the agent. Such a mechanism is consistent with the simultaneous use of two types of performance reports: an internal private one, observed only by the monitor and the principal, and a public one, also disclosed to the agent. Using two separate reports allows the principal to incentivize monitoring without necessarily generating more retaliation.

1.1 Related literature

This paper contributes to the literature on moral hazard with endogenous monitoring, such as Kvaløy and Olsen (2009), Georgiadis and Szentes (2020), and Li and Yang (2020). In this literature, before choosing the contracts, the principal acquires a monitoring technology at a given ex-ante cost and chooses optimal contracts given the information structure selected. In my paper, the monitoring costs are not determined ex-ante and depend on the signal realization. Retaliation occurs only when the evidence disclosed by the monitor reduces the agent's payment. Hence, what retaliation costs and information structure arise in equilibrium directly depends on the contracts.

As low signal realizations are helpful to the principal only if they generate low payments, and low payments imply retaliation costs, this paper is also related to the literature on costly verification, e.g., Townsend (1979), Gale and Hellwig (1985), and Hart and Moore (1998). In these models, an investor decides whether to verify the firm's performance at a cost. However, the cost is paid upfront, independently of the evidence realization. In our model, there is a cost only if the realized state generates low payments. Hence, high state realizations are always revealed at the optimal contracts, while low realizations are not.

A closely related article is that of Lang (2019), who studies the optimal use of subjective performance evaluations under partial and costly verification. They assume that evaluations are subjective but that the signal can be verified at a given cost. The agent demands a justification whenever he does not get the highest payment, which prevents the principal from underreporting signals to save on payments. His main finding is that the optimal contract pools high signal realizations to avoid such justification costs. My model has two main differences from Lang (2019): first, the information is verifiable if revealed; second, the one deciding whether or not to disclose is the monitor, who has no incentive to renege on payments (or conceal high signals). As a result, the optimal contracts in my model feature the pooling of low and intermediary signal realizations instead of high ones.

Another closely related strand of literature is the one on whistleblowing and retaliation, such as Chassang and Padró i Miquel (2019). In their model, a monitor perfectly observes whether an agent has committed a crime and decides whether to report it to the principal. The agent commits to a retaliation strategy to incentivize the monitor to use her preferred report. Like my results, the principal limits how much her response to the monitor's reports reveals about the reports themselves to allow for information transmission. However, there are two critical differences: first, the monitor is fully informed about a binary action in their model. Hence there is no loss in restricting attention to binary reports. In my model, the evidence observed by the monitor is not binary. Second, we allow the principal to control incentives fully; hence she can compensate the monitor for the retaliation. Their analysis focuses on cases where the principal cannot directly control the payoffs. Even with this additional tool, we show that the principal still prefers to commit not to use some of the information to avoid retaliation costs.

Finally, the paper also relates to the literature on contracts as reference points pioneered by Hart and Moore (2008). In my model, one can interpret the payment to the agent in the case of no disclosure as a reference point and the retaliation cost incurred by the monitor as a psychological cost of disclosing bad information about the agent's performance. Signal realizations that generate a payment higher than

the endogenously set reference point (the payment in the case of no disclosure) do not create retaliation losses. In contrast, signal realizations associated with lower payments do. We provide a tractable way to find the optimal contract with an endogenous reference point and relate it to an information acquisition problem.

The rest of the article is organized as follows: Section 2 describes the model and analyzes a benchmark case without retaliation. Section 3 introduces retaliation and characterizes optimal contracts. Section 4 reinterprets the optimal level of retaliation tolerated as an information acquisition problem. In Section 5, we introduce costly monitoring. Finally, Section 6 concludes.

2 Model

Consider a risk-neutral principal who hires a monitor (she) and an agent (he), both risk-averse. The principal proposes contracts that specify payments to both employees conditional on a verifiable signal, as described later. After signing contracts, the agent exerts effort $a \in [0,1]$, which is unobservable by the principal and the monitor. With probability $m \in (0,1)$, the monitor observes the realization of a verifiable signal denoted by $\mathbf{x}$ with support $X = [\underline{x}, \overline{x}]$. The monitor then decides whether or not to disclose the realized signal[4]. With probability $(1-m)$, the monitor is uninformed and has nothing to disclose. The verifiable signal $\mathbf{x}$ is drawn from a cumulative distribution function $P(\cdot|a)$, that admits a density given by

$$p(x|a) = ap_1(x) + (1-a)p_0(x),$$

where p_0, p_1 are densities strictly bounded away from zero and with support X. We define the score of a signal $s : X \times [0,1] \to \mathbb{R}$ as

$$s(x|a) := \frac{p_1(x) - p_0(x)}{p(x|a)}.$$

We assume that $s(\cdot|a)$ is strictly increasing, bounded, and continuously differentiable. The score is a strictly increasing transformation of the likelihood ratio of a signal and is a sufficient statistic in canonical moral hazard problems (see Holmström (1979)).

The timing of the game is the following:

1. The principal offers contracts specifying two measurable functions: the monitor and the agent's payments $w_M, w_A : X \cup \{\emptyset\} \to \mathbb{R}$. The symbol $\emptyset$ denotes that no evidence was disclosed.
2. The agent and the monitor decide whether or not to accept the contracts. If either of them rejects, the game ends, and they both get their respective outside option $\bar{u}_M$ and $\bar{u}_A$.
3. If both contracts are accepted, the agent chooses effort $a \in [0,1]$.
4. $\mathbf{x}$ is realized and the monitor observes the realization with probability m.
5. If informed, the monitor decides whether or not to disclose x.

[4]We denote random variables in bold font and typical realizations of the random variable in regular font.

6. Payments are realized.

7. The agent observes his payment, and retaliation takes place.

The agent's effort $a \in [0,1]$ generates a disutility given by $c_A : [0,1] \to \mathbb{R}_+$. The function c_A is twice continuously differentiable, strictly increasing, strictly convex, and $c_A(0) = 0$. The agent is risk averse with strictly increasing and strictly concave utility over income, denoted by the function $u_A : \mathbb{R} \to \mathbb{R}$ with a finite lower bound $\underline{u} < \bar{u}_A$[5]. If the agent exerts effort a and gets a wage w his utility is $u_A(w) - c_A(a)$.

The monitor privately observes the realization of the verifiable signal $\mathbf{x}$ with probability m. With probability $(1-m)$, she does not observe anything. The monitor then decides whether or not to disclose the signal to the principal. If the monitor is uninformed, she has nothing to disclose. However, if informed, she can hide the information and pretend to be uninformed. The monitor has preferences over income and suffers a utility loss $L_r \geq 0$ if retaliated against. If the monitor gets a wage w and is retaliated against with probability r, then her payoff is $u_M(w) - rL_R$, where $u_r : \mathbb{R} \to \mathbb{R}$ is strictly increasing and weakly concave.

The principal has a gross benefit from a given by $B(a)$, where B is strictly increasing and weakly concave. She is risk-neutral with respect to payments and incurs a cost $c_r \geq 0$ whenever the agent retaliates. For a given effort level a, payments w_A and w_M, and retaliation probability r, the principal's payoff is

$$\Pi(a, w_A, w_M, r) = B(a) - w_A - w_M - rc_r.$$

The principal commits to payments conditional on what she observes (disclosed signal or nothing). The monitor — if informed — decides whether or not to reveal the signal. The agent chooses his effort and, after observing his payment, whether or not to retaliate. We further assume that retaliation strictly harms at least one of the two (principal or monitor). That is, $max\{c_r, L_r\} > 0$.

2.1 Retaliation and Disclosure

We assume the agent retaliates whenever his payment under disclosure is smaller than without disclosure. That is, whenever the monitor's report reduces the agent's payments, he retaliates. Given a payment function $w_A : X \cup \{\emptyset\} \to \mathbb{R}$, an agent who received payment w, retaliates if and only if $w < w_A(\emptyset)$. With a slight abuse of notation, we write the retaliation strategy $r(x)$ as a function of x. For a given contract w_A, $r(x) = 1$ if $w_A(x) < w_A(\emptyset)$ and zero otherwise. This retaliation form is assumed to keep the model as simple as possible while still capturing the reciprocal nature of retaliation[6].

An alternative interpretation of the model is that the monitor directly dislikes reducing the agent's payments. That is, instead of thinking about the retaliation as an action by the agent, one can assume the monitor incurs a psychological cost whenever her report decreases the agent's payment. In such an interpretation, the parameters would satisfy $L_r > 0$ and $c_r = 0$.

[5] The finite lower bound assures the existence of optimal contracts. See Moroni and Swinkels (2014) for a detailed argument.

[6] An alternative approach that generates the same results is to assume the agent can commit to a retaliation strategy.

We denote the monitor's strategy about whether or not to disclose the signal as $d : X \to \{0,1\}$. Given contracts and a given realization x, the monitor's best response is to disclose the signal if her payoff is higher under disclosure. That is,

$$d(x) = 1 \text{ if and only if } u_M(w_M(x)) - r(x)L_r \geq u_M(w_M(\emptyset)).$$

For given contracts, a disclosure strategy, and an effort level a, the agent's expected utility is given by

$$U_A(a,r,d,m) = \int_X \Big\{ [1 - md(x)] u_A(w_A(\emptyset)) + md(x) u_A(w_A(x)) \Big\} p(x|a) dx - c(a), \tag{1}$$

while the monitor's expected utility is

$$U_M(a,r,d,m) = \int_X \Big\{ [1 - md(x)] u_M(w_M(\emptyset)) + md(x) \Big[u_M(w_M(x)) - L_r(r(x)) \Big] \Big\} p(x|a) dx. \tag{2}$$

2.2 Principal's Problem

Following the Grossman and Hart (1983) approach, we study the principal's problem of minimizing the cost of implementing a given effort level $a \in (0,1)$. By the revelation principle (Myerson (1982)), we can, without loss, focus on minimizing expected payments and retaliation costs by choosing contracts, recommending an effort level, a retaliation strategy, and a disclosure strategy such that monitor and agent are willing to participate and follow the recommendations. That is, the principal's problem can be written as

$$\min_{w_A, w_M, r, d} \mathbb{E}_{\mathbf{x}} \Big[(1 - md(\mathbf{x}))(w_A(\emptyset) + w_M(\emptyset)) + md(\mathbf{x})(w_a(\mathbf{x}) + w_M(\mathbf{x}) + c_r r(\mathbf{x})) \Big| a \Big] \tag{3}$$

subject to

$$U_A(a,r,d,m) \geq \bar{u}_A, \tag{IR_A}$$

$$U_M(a,r,d,m) \geq \bar{u}_M, \tag{IR_M}$$

$$a \in \underset{\hat{a} \in [0,1]}{argmax} \Big\{ U_A(\hat{a},r,d,m) \Big\}, \tag{IC_A}$$

$$[w_A(x) - w_A(\emptyset)][1 - r(x)] \geq 0 \quad \forall x \in X, \tag{IC_r}$$

$$[u_M(w_M(x)) - u_M(w_M(\emptyset)) - r(x)L_r] d(x) \geq 0 \quad \forall x \in X, \tag{IC_d}$$

where (IR_A) and (IR_M) denote the usual participation constraints, (IC_A) denotes the effort incentive compatibility constraint, (IC_d) assures the monitor is willing to follow the recommended disclosure strategy, and (IC_r) guarantees the recommended retaliation strategy is implemented.

The agent's payment when no information is disclosed $w_A(\emptyset)$ is an endogenous reference value determining when retaliation occurs. We refer to payments strictly below this reference as punishments (or sticks) and to payments strictly above as rewards (or carrots). Using sticks generates retaliation, while paying the reference value or using carrots does not.

2.3 Preliminary analysis: a convenient change of variables

As is standard, it is convenient to work in the space of utilities instead of payments. Therefore, we make the following change of variables

$$v_A(\emptyset) := u_A(w_A(\emptyset)) \text{ and } v_A(x) := [u_A(w_A(x)) - u_A(w_A(\emptyset))]$$

$$v_M(\emptyset) := u_M(w_M(\emptyset)) \text{ and } v_M(x) := [u_M(w_M(x)) - u_M(w_M(\emptyset))],$$

where $v_i(\emptyset)$ represents the monetary utility from payments in case no evidence was revealed, and $v_i(x)$ is the incremental utility (which might be negative) associated with signal x for each $i \in \{A, M\}$. The monetary utility associated with no disclosure $v_A(\emptyset)$ denotes the reference point determining whether there is retaliation. Signal realizations with $v_A(x) > 0$ are the ones associated with payments above the reference value $v_A(\emptyset)$ (the carrots) which do not generate retaliation. While signal realizations with $v_A(x) < 0$ generate lower payments (the stick) and are retaliated against.

We can then re-write the agent's and monitor's participation constraints as:

$$U_A(a,r,d,m) = \int_X \Big\{ v_A(\emptyset) + m v_A(x) d(x) \Big\} p(x|a) dx - c(a) \geq \bar{u}_A, \tag{IR_A}$$

$$U_M(a,r,d,m) = \int_X \Big\{ v_M(\emptyset) + m v_M(x) d(x) - m r(x) d(x) L_r \Big\} p(x|a) dx \geq \bar{u}_M. \tag{IR_M}$$

Note that for any mechanism, the agent's payoff is strictly concave in the agent's effort. Therefore, effort incentive compatibility can be replaced by the first-order condition. That is, (IC'_A) is equivalent to

$$m \int_X v_A(x) d(x) s(x|a) p(x|a) dx = c'_A(a). \tag{IC'_A}$$

Finally, let $\varphi_i := u_i^{-1}$ denote the inverse of u_i. One can re-write the principal's problem as

$$\begin{aligned} \min_{v_A, v_M, r, d} \int_X \Big\{ & m d(x) \Big[\varphi_M\big(v_M(x) + v_M(\emptyset)\big) + \varphi_A\big(v_A(x) + v_A(\emptyset)\big) + c_r r(x) \Big] \\ & + (1 - m d(x)) \Big[\varphi_M\big(v_M(\emptyset)\big) + \varphi_A\big(v_A(\emptyset)\big) \Big] \Big\} p(x|a) dx \end{aligned} \tag{4}$$

subject to (IR_A), (IR_M), (IC'_A), (IC_M), (IC_d), (IC_r).

2.4 Harmless retaliation benchmark

The novel friction introduced in this model is the possibility of the agent inflicting damage on the principal and the monitor by retaliating. It is helpful to revisit the case in which there is no retaliation or equivalently when retaliation is harmless ($L_r = c_r = 0$). First, note that paying a flat wage to the monitor given by $v_M(\emptyset) = \bar{u}_M$ and $v_M(x) = 0$ for all $x \in [\underline{x}, \bar{x}]$ is enough to assure she participates and is willing

to disclose all the information she gets. The following standard moral hazard problem characterizes the optimal agent's compensation scheme:

$$\min_{v_A}\left\{(1-m)\varphi_A\big(v_A(\emptyset)\big)+m\int_X \varphi_A\big(v_A(x)+v_A(\emptyset)\big)d(x)p(x|a)dx\right\} \tag{5}$$

subject to (IR_A) and (IC'_A).

The solution to this problem has the standard Holmström-Mirrlees form (see Holmström (1979) and Mirrlees (1999)), for which the score $s(\cdot|a)$ is a sufficient statistic for the optimal contract. The optimal agent's compensation scheme denoted by $v_A^0 : X \cup \{\emptyset\} \to \mathbb{R}$ is given by

$$\varphi'_A(v_A^0(\emptyset)) = \lambda_A^0,$$

$$\varphi'_A(v_A^0(\emptyset)+v_A^0(x)) = \lambda_A^0+\mu_A^0 s(x|a),$$

where λ_A^0 and μ_A^0 are the dual multipliers associated with (IR_A) and (IC_A) respectively. Moreover, given the increasing score assumption, higher signal realizations are associated with higher payments. Below we present a graphical illustration of the optimal agent's compensation scheme. The dashed line represents the payments in case of no-disclosure, and the solid line is the payment for each realization x.

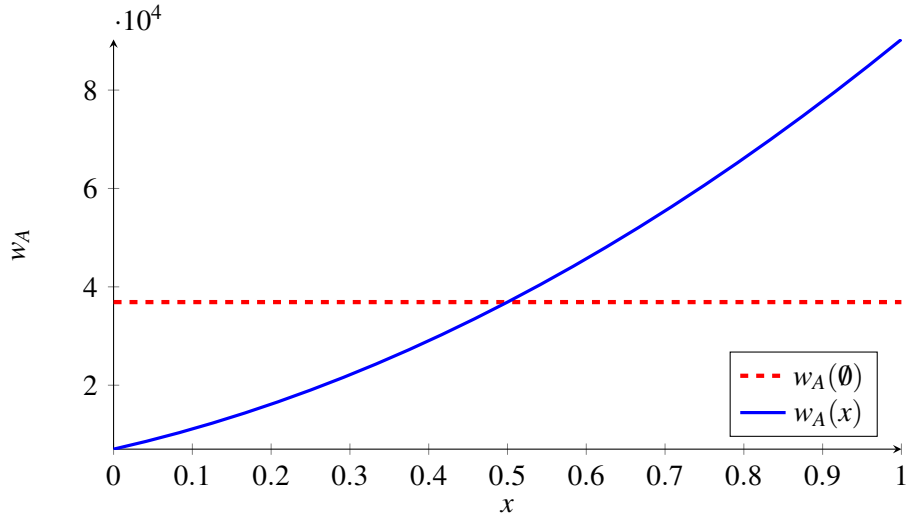

Figure 1: Harmless Retaliation Benchmark

There are a few important features of this benchmark to highlight. First, note that both rewards (payments above the no disclosure reference) and punishments (payments below the no disclosure reference) are used. Second, note that the payment to the agent under no disclosure is the same as the payment under the signal realization x_0 that has $s(x_0|a) = 0$. Moreover, all realizations below x_0 are associated with punishments and all above with rewards. The score of a signal realization is determined by how often one observes such a realization under low versus high effort. Realizations with negative scores are more likely to be realized under low effort levels, while positive ones occur more often for high effort levels. One can interpret negative score signal realizations as bad and positive score realizations as good ones. In the harmless retaliation benchmark, the principal punishes all bad realizations and rewards the good ones.

3 Harmful retaliation

We now reintroduce retaliation to the problem. That is, we proceed to analyze problem (4) assuming that retaliation is harmful ($max\{c_r, L_r\} > 0$). The approach to finding the optimal mechanism is divided into a few steps: first, we characterize the monitor's optimal compensation scheme as a function of the retaliation strategy implemented. Second, we show that retaliation must take a cutoff form. The agent retaliates only after a sufficiently low signal realization is revealed. Third, we find the cost-minimizing agent's contract for a given retaliation cutoff. Fourth, we find the optimal cutoff. We then contrast the optimal mechanism with harmful retaliation with the harmless retaliation benchmark.

Define a mechanism as feasible if it satisfies all the imposed constraints.

Definition 1. *We say a mechanism (v_M, v_A, r, d) is **feasible** if it satisfies (IR_M), (IR_A), (IC'_A), (IC_r) and (IC_d).*

We then address the monitor's disclosure decision and compensation scheme. The first observation is that there is no loss of optimality in restricting attention to mechanisms that recommend the monitor to disclose all observed signals.

Definition 2. *A mechanism satisfies **full disclosure** if $d(x) = 1$ for all $x \in X$.*

Lemma 1. *There is no loss of optimality in restricting attention to full disclosure mechanisms.*

Proof. In Appendix. □

The principal can always replicate the monitor's disclosure strategy directly on payments and implement full disclosure. Suppose that in a given mechanism (v_M, v_A, r, d) there is a set $\tilde{x} \subset X$ such that $d(x) = 0$ for all $x \in \tilde{X}$. It would be equivalent to recommend $\tilde{d}(x) = 1$ for all x, and change payments to $(\tilde{v}_M, \tilde{v}_A)$, with $(\tilde{v}_M(x), \tilde{v}_A(x)) = (v_M(\emptyset), v_A(\emptyset))$ for all $x \in \tilde{X}$, and $(\tilde{v}_M(x), \tilde{v}_A(x)) = (v_M(x), v_A(x))$ otherwise.

Knowing that the principal can, without loss, get full disclosure, we look for the cheapest monitor's payment function that implements full disclosure. It consists of paying the monitor her outside option whenever the disclosed signal is such that she is not retaliated against and compensating her for the retaliation loss whenever retaliation occurs.

Lemma 2. *For any given triple (v_M, v_A, r), the cheapest monitor's compensation that implements full disclosure is given by $v_M(x) = r(x)L_r$ and $v_M(\emptyset) = \bar{u}_M$.*

Proof. In Appendix. □

Lemmas 1 and 2 jointly characterize the optimal monitor's compensation scheme for a given retaliation recommendation. The following section shows that such a recommendation must take a cutoff form.

3.1 Retaliation as a cutoff

Retaliation occurs when the agent gets paid below what he would have been paid if the signal were not disclosed. We now show that in the optimal mechanism, it must be the case that the agent retaliates if the disclosed signal is low enough and does not retaliate otherwise[7].

Definition 3. *A function $g : X \to [0,1]$ is a **cutoff function** if there exists $x^* \in X$ such that $g(x) = 1$ for almost all $x < x^*$ and $g(x) = 0$ for almost all $x > x^*$.*

Proposition 1. *Take any arbitrary feasible full disclosure mechanism (v_M, v_A, r) such that r is not a cutoff function. Then, there exists an alternative feasible full disclosure mechanism $(\tilde{v}_M, \tilde{v}_A, \tilde{r})$ with a strictly lower implementation cost.*

Proof. In Appendix. □

Proposition 1 implies that the optimal mechanism must have a cutoff retaliation strategy recommendation. If the agent retaliates after a given signal realization is disclosed, he must retaliate if any lower signal is revealed. At first sight, the result might seem obvious; however, one must recall that retaliation is directly determined by the payments after each signal realization, which is an endogenous object. The proof is constructive: take an arbitrary feasible full disclosure mechanism such that r is not a cutoff function. Then, construct a strict improvement by switching payments from higher signal realizations with retaliation to lower realizations that did not generate retaliation while keeping the agent's expected utility the same. One can check that such a switch causes slackness in the effort incentive compatibility constraint, which allows the principal to offer less steep incentives and decrease expected payments to the agent.

From now on, we denote the recommended retaliation by a cutoff $x_r \in X$. That is, $r(x) = 1$ if $x < x_r$ and $r(x) = 0$ otherwise. We then denote mechanisms by a triple (v_M, v_A, x_r). Our next step is to show that there is no loss in restricting attention to cutoffs $x_r \leq x_0$[8]. That is, the principal would never use a mechanism in which retaliation happens after good signal realizations.

Lemma 3. *For any feasible full disclosure mechanism (v_M, v_A, x_r) such that $x_r > x_0$, there exists an alternative feasible full disclosure mechanism $(\tilde{v}_M, \tilde{v}_A, x_0)$ with lower implementation costs.*

Proof. In Appendix. □

In the harmless retaliation benchmark, signal realizations above x_0 did not generate retaliation because they had higher scores (and consequently payments) than the empty signal. There is no reason for the principal to pay less for such realizations than for the empty signal, which implies that the retaliation cutoff must be below x_0.

[7] Note that we have not restricted the contracting space to increasing payments. Hence, the retaliation strategy assumed does not directly imply that retaliation must take a cutoff form.

[8] Recall that x_0 is such that $s(x_0|a) = 0$.

3.2 Optimal agent's compensation for a given retaliation cutoff

Knowing that retaliation must take a cutoff form, we break the characterization of the optimal mechanism into two steps. First, we fix a retaliation cutoff and find the agent's contract that minimizes the principal's expected cost. Then, we look for the best retaliation cutoff. In Section 4, we provide an information acquisition interpretation for this two-step procedure.

For a fix x_r, Lemma 2 characterizes the monitor's compensation scheme v_M. Hence, we can write the problem of minimizing the principal's cost by choosing the agent's contract v_A as

$$\begin{aligned} C(x_r,a,m) := \min_{v_A}\Bigg\{ & \varphi_M(\bar{u}_M)\Big(1-m+m\int_{x_r}^{\bar{x}} p(t|a)dt\Big) + m\big[\varphi_M(\bar{u}_M+L_r)+c_r\big]\int_{\underline{x}}^{x_r} p(t|a)dt \\ & + \int_X \Big[m\varphi_A\big(v_A(x)+v_A(\emptyset)\big)+(1-m)\varphi_A\big(v_A(\emptyset)\big)\Big]p(x|a)dx\Bigg\} \end{aligned} \tag{6}$$

subject to

$$m\int_X v_A(x)s(x|a)p(x|a)dx = c_A'(a). \tag{IC_A'}$$

$$\int_X \Big\{v_A(\emptyset)+mv_A(x)\Big\}p(x|a)dx - c_A(a) \geq \bar{u}_A \tag{IR_A}$$

$$v_A(x)[1-\chi^x_{[\underline{x},x_r)}]mp(x|a) \geq 0 \quad \forall x \in X, \tag{IC_r}$$

where $\chi^x_{[\underline{x},x_r)}$ denotes an indicator function of whether $x \in [\underline{x},x_r)$ or not.

The first line of equation (6) denotes the expected payments to the monitor plus the expected direct retaliation losses suffered by the principal. The second line denotes the expected payments to the agent. We look for the v_A that minimizes the principal's expected costs while satisfying (IR_A), (IC_A'), and implementing the desired retaliation cutoff.

Proposition 2. *Given $x_r \leq x_0$, the optimal agent's compensation scheme is characterized by*

$$\varphi_A'(v_A(\emptyset)) = \lambda_A + \mu_A s(\tilde{x}|a),$$

$$\varphi_A'(v_A(\emptyset)+v_A(x)) = \lambda_A + \mu_A s(\tilde{x}|a) \text{ for } x \in (x_r,\tilde{x}],$$

$$\varphi_A'(v_A(\emptyset)+v_A(x)) = \lambda_A + \mu_A s(x|a) \text{ for } x \in X \setminus (x_r,\tilde{x}],$$

where $\tilde{x}$ is uniquely characterized by

$$s(\tilde{x}|a) = \frac{m}{1-m\Big[1-\int_{x_r}^{\tilde{x}} p(x|a)dx\Big]}\int_{x_r}^{\tilde{x}} s(x|a)p(x|a)dx, \tag{7}$$

and λ_A and μ_A are the dual multipliers associated with (IR_A) and (IC_A) respectively.

Proof. In Appendix. □

There are a few characteristics of the optimal compensation scheme worth highlighting. First, note that it takes a modified Holmström-Mirrlees form, in which the score remains a sufficient statistic. Second, it implies that whenever the principal wants to reduce retaliation from the harmless retaliation

benchmark ($x_r < x_0$), there is a concentration in payments at the reference payment. That is, the agent receives the reference payment not only when the monitor is uninformed but also for any disclosed signal in $(x_r, \tilde{x}]$. Third, note that the reference payment is no longer associated with a score equal to zero but with $s(\tilde{x}|a) < 0$, which means that preventing retaliation decreases the performance standards for receiving the reference value. Figure 3.2 plots an example of an optimal compensation scheme.

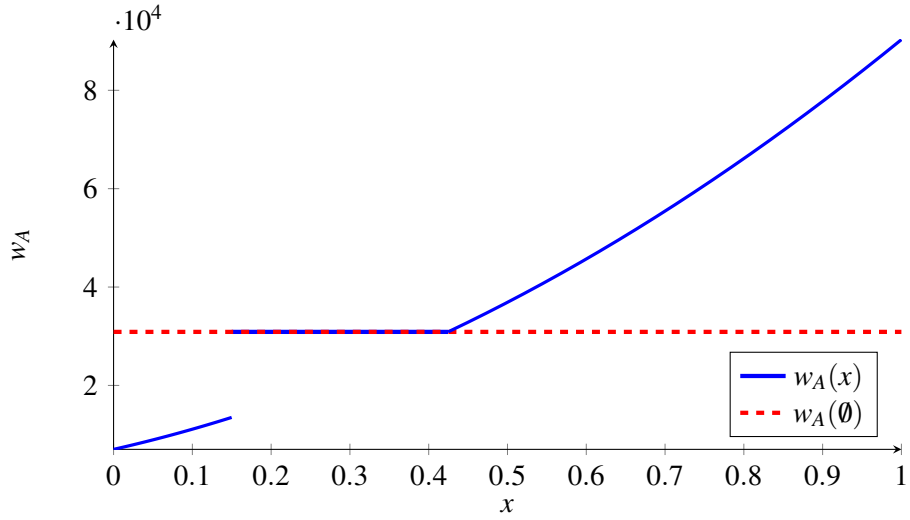

Figure 2: Optimal Agent's Compensation for a Given x_r

Moreover, note that as payments when the disclosed signal belongs to $(x_r, \tilde{x}]$ are the same as when the monitor is uninformed, the optimal compensation scheme is as if the monitor revealed all signals but the ones in $(x_r, \tilde{x}]$. That is, the principal could equivalently implement the same effort at the same cost by recommending the monitor not to reveal signal realizations in $(x_r, \tilde{x}]$.

Under full disclosure, the observation of no signal has a score equal to 0. As seen in the harmless retaliation benchmark, if the principal disregards retaliation, all realizations below x_0 would generate retaliation. When fixing x_r, we establish the most retaliation the principal is willing to tolerate. To avoid retaliation for signals above x_r and below x_0, the principal must increase such payments up to how much she pays in case of no information, that is, $v_A(\emptyset)$. However, paying the same for a given x as the no information case is equivalent to asking the monitor not to reveal such x. When the monitor does not disclose a subset of possible signal realizations, those realizations are pooled with the empty signal, which decreases the score associated with observing the empty signal. As the score is a sufficient statistic for payments, pooling low signals with $\emptyset$ decreases how much the principal pays the agent if there is no disclosure. Equivalently, the principal can recommend the monitor not to disclose all signals starting from x_r up to $\tilde{x}$. Where $\tilde{x}$ is such that it has the same score as the empty signal when there is no disclosure in $(x_r, \tilde{x}]$. Any signals above $\tilde{x}$ have a higher score. Hence, they generate higher payments than the empty signal and do not generate retaliation.

Note also that $\tilde{x}$ is fully characterized by x_r and does not depend on v_A. Hence, fixing x_r is equivalent to fixing the information that is going to be revealed to the principal. For a given x_r, solving the problem (6) is equivalent to solving a canonical moral hazard problem in which the information structure is exogenous and the principal does not observe realizations in $(x_r, \tilde{x}]$.

Remark 1. *The implementation described above, which requests the monitor not to reveal signal realizations in $(x_r, \tilde{x}]$ does not contradict the Lemma 1. Showing that full disclosure is without loss facilitates the characterization of the optimal contracts. However, it does not imply that full disclosure is necessary for optimality. Using the full disclosure implementation of the mechanism, the principal commits "not to use" signal realizations in $(x_r, \tilde{x}]$. She commits to pay the agent the same under such realizations or without disclosure. Hence, asking the monitor not to report such signals is equivalent.*

It is interesting to notice that the reference payment is an endogenous object. When the principal changes the amount of retaliation tolerated, she affects the set of signals disclosed by the monitor. Hence, she affects the informational content of observing no-disclosure. Next, we describe how different x_r's affect the score associated with no disclosure.

Corollary 1. *The score associated with the reference payment $s(\tilde{x}|a)$ is increasing in x_r. That is, decreasing the retaliation cutoff implies reducing the standards for the reference payment.*

Proof. In Appendix. □

By Proposition 2, we know that the informational content of a given signal is captured by its score. When the principal tolerates less retaliation, it requires lower performances not to be revealed, which reduces the standards for the reference no disclosure payment. That is, a firm that tolerates less retaliation faces a lower benchmark for performance.

4 Retaliation and information acquisition

We have established that the choice of the retaliation cutoff implies the information observed by the principal. We now argue that tolerating more retaliation (higher x_r) is formally equivalent to costly acquiring more information (in a Blackwell sense) about the agent's effort.

As is typical in moral hazard problems, the score $s(\cdot|a)$ is a sufficient statistic for the optimal compensation problem. In particular, by construction of $\tilde{x}$, when $x \in (x_r, \tilde{x}]$ is not disclosed, the score of no-disclosure (the $\emptyset$ signal) is given by $s(\tilde{x}|a)$. Denote by $\rho(\cdot, a)$ the inverse of the score function for a given effort level a, that is, $\rho\big(s(x|a), a\big) = x$[9]. Define $F^{x_r}(\cdot|a)$ as the cumulative probability function of the scores observed by the principal when the monitor discloses all signals but $x \in (x_r, \tilde{x})$. That is,

$$F^{x_r}(y|a) := \begin{cases} m \int_{\underline{x}}^{\rho(y,a)} p(t|a)dt & \text{if } y \leq s(x_r|a), \\ m \int_{\underline{x}}^{x_r} p(t|a)dt & \text{if } y \in [s(x_r|a), s(\tilde{x}|a)), \\ (1-m) + m \int_{\underline{x}}^{\rho(y,a)} p(t|a)dt & \text{otherwise.} \end{cases}$$

Scores below $s(x_r|a)$ are associated with realizations of $\mathbf{x}$ below x_r, which are disclosed by the monitor whenever observed. Hence, the first part of the definition of F^{x_r}. The second part stems from the fact

[9] As $s(\cdot|a)$ is strictly increasing, there is a well-defined inverse.

that $\mathbf{x}$'s realizations in $(x_r, \tilde{x}]$ are not disclosed. Thus, the cdf remains constant at $F^{x_r}(s(x_r|a)|a)$ until $y = s(\tilde{x}|a)$. The last part reflects the fact that at $y = s(\tilde{x}|a)$, there is a mass point regarding all the non-disclosed signals together with the probability of an uninformed monitor.

Theorem 1. *If the principal chooses a higher x_r, she gets more information (in the Blackwell sense) about the agent's effort. That is, if $x_r' \geq x_r''$, then $F^{x_r'} \gtrsim F^{x_r''}$, where $\gtrsim$ denotes the second-order stochastic dominance relation.*

Proof. For any x_r, the cumulative distribution $F^{x_r}(y|a)$ co-moves with the distribution of $\mathbf{x}$ up to $y = s(x_r|a)$. Then, all realizations in $(x_r, \tilde{x}]$ are pooled with the empty signal. Hence, the cumulative distribution $F^{x_r}(\cdot|a)$ is flat on $(s(x_r|a), s(\tilde{x}|a))$. At $s(\tilde{x}|a)$, there is a mass point corresponding to all signals that were not disclosed and pooled with the non-informative signal realization. For realizations above $\tilde{x}$, the distribution again co-moves with $\mathbf{x}$'s distribution. Below, we plot $F^{x_r}(\cdot|a)$ for the harmless retaliation benchmark ($x_r = x_0$), as well as for examples with $x_0 > x_r' > x_r''$. Figure 4 illustrates that $F^{x_r''}$ single crosses $F^{x_r'}$ from below when $x_r' \geq x_r''$. Therefore, $F^{x_r'}$ Blackwell-dominates $F^{x_r''}$.

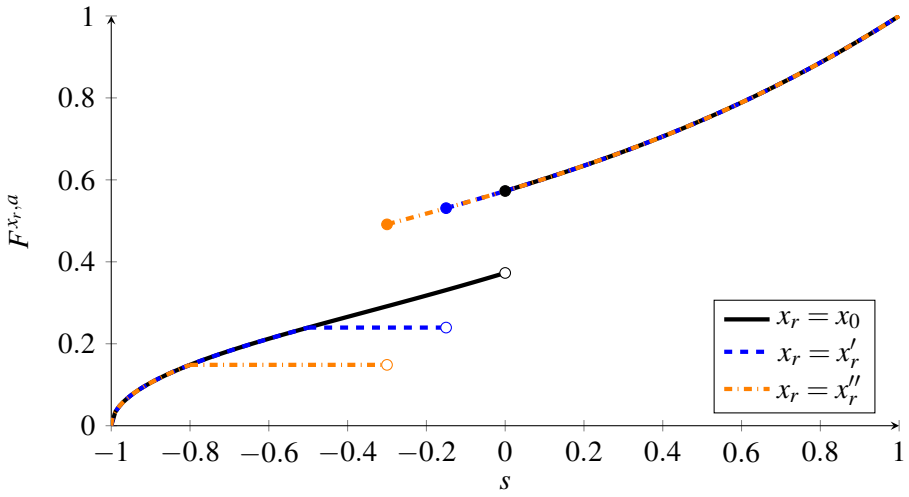

Figure 3: Distribution of Scores as a Function of x_r

□

Theorem 1 implies that a higher choice of x_r corresponds to a more informative signal about the agent's effort. However, a higher x_r implies dealing with retaliation costs more often. For each $x < x_r$, the principal incurs the cost $\kappa_r := \left[\varphi_M\left(\bar{u}_M + L_r\right) - \varphi_M\left(\bar{u}_M\right) + c_r\right]$ which implies that tolerating more retaliation (choosing a higher x_r) corresponds to acquiring more information, while paying the additional κ_r retaliation costs. Firms then face a trade-off between acquiring better information about their workers' performance or disregarding information about poor performances and avoiding retaliation costs.

4.1 The optimal retaliation cutoff

We have established the principal's cost and associated contracts for each amount of retaliation the principal is willing to tolerate. The final step in characterizing the optimal mechanism is finding the optimal retaliation cutoff. We are then ready to state our main result, which describes many features of the optimal mechanism when retaliation is harmful.

Theorem 2. *Suppose retaliation is harmful. Then, for a given pair* $(a,m) \in (0,1)^2$ *the cost-minimizing mechanism is consistent with the following features:*

i- The monitor is lenient: she refrains from revealing moderately (potentially all) signals perceived as bad;

ii- Payments are compressed: multiple different realized performances generate the same agent's payment;

iii- No news is bad news: no-disclosure is associated with a strictly negative score.

iv- For low retaliation costs, the optimal agent's contract uses punishments and rewards to motivate the agent (stick-and-carrot contract);

v- For high retaliation costs, the optimal agent's contract uses only rewards (carrot-only contract).

Proof. In Appendix. □

Parts (i) and (ii) stem from the fact that there is always a region in which payments are flat and equal to the payment in case of no-disclosure. As dealing with retaliation is costly, the principal is better off pooling signals with a score below but close to the one of no-disclosure. That is, such signal realizations have an information value close to the no-disclosure signal but entail a strictly positive retaliation cost. Hence, the principal is better off pooling the payments of such signals with the benchmark and avoiding retaliation costs. This pattern arises optimally, but it is consistent with two empirical features of performance evaluations often perceived as harmful biases: leniency and centrality.

Leniency refers to monitors refraining from revealing bad performance by their subordinates. We have shown that the optimal mechanism can be implemented by asking the monitor not to disclose signals in $(x_r,\tilde{x}]$, which have a score strictly lower than no-disclosure score. Centrality refers to agents with different performances receiving the same compensation. As all agents with performance in $(x_r,\tilde{x}]$ get the same reward, the optimal agent's compensation has centrality as a feature.

Part (iii) is a direct consequence of this manifestation of leniency and centrality. As the principal optimally lets the manager pool signals with a negative score with the no-disclosure benchmark, it is as if no-disclosure had a negative score. Refraining from revealing bad performances deteriorates the benchmark since no-disclosure now arises when the performance is low.

Parts (iv) and (v) describe cases in which the principal uses a stick-and–carrot or carrot-only contract. On the one hand, when κ_r is sufficiently small, the informational gains from punishing extremely bad

performances overweight retaliation costs, and the principal uses both punishments and rewards to motivate the agent. In this case, the agent's compensation scheme takes the form illustrated in Figure 3.2. On the other hand, when κ_r is large, the principal is better off by completely eliminating retaliation. In such a case, the agent's compensation scheme has a flat region for low signals and increases after $\tilde{x}$, as plotted in the example below.

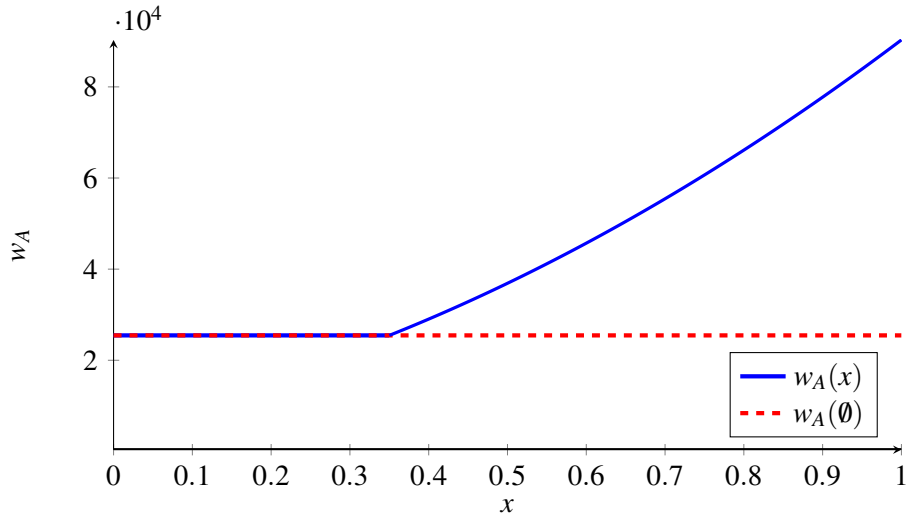

Figure 4: Optimal Compensation with High Retaliation Cost

Note also that when the retaliation costs are high, the agent's compensation has a flat region for low values of x. The agent is only paid above the base wage when the performance is sufficiently high. To the best of our knowledge, this is the first paper that generates a flat payment region for low-performance levels without relying on a binding minimum payment constraint. When a binding minimum payment constraint exists, the optimal contract naturally presents a flat initial region (see Jewitt et al. (2008)). However, in practice, we often observe flat payments for low-performance levels even when there is no binding minimum wage constraint. In our model, the level at which the flat payment occurs is determined by the outside option, not an exogenous payment constraint. Hence, it may occur even for job positions where the minimum wage is never binding.

5 Incentives for monitoring

We have assumed that the monitor observes the signal $\mathbf{x}$ for free. One can enrich the model by assuming that $m \in [\underline{m}, \overline{m}] \subset (0,1)$ is a monitor's choice, for which the monitor incurs a private cost $c_M(m)$, which is strictly increasing, strictly convex and such that $c_M(0) = 0$. The principal now must motivate the monitor to exert the desired monitoring effort.

The agent's compensation for a given x_r remains unaltered. The only change is that the principal must incentivize monitoring. That is, the monitor's participation constraint must account for monitoring costs, and there is a monitor's incentive compatibility constraint:

$$U_M(m) \geq c_M(m) + \bar{u}_M, \qquad (IR_M)$$

$$\int_X v_M(x)p(x|a)dx - L_r \int_{\underline{x}}^{x_r} p(x|a) = c'_M(m). \qquad (IC_M)$$

The agent's compensation problem remains unaltered.

Proposition 3. *Fix $(a,m) \in (0,1)^2$ and a $x_r \in [\underline{x}, x_0)$. The optimal manager's compensation scheme is given by $v_M(\emptyset) = \bar{u}_M$ and*

$$v_M(x) = \begin{cases} c'_M(m) + L_r & \text{if } x \in [\underline{x}, x_r], \\ c'_M(m) & \text{otherwise.} \end{cases}$$

While v_A is characterized in Proposition 2.

Disclosing the signal becomes evidence that the monitor has exerted monitoring effort. Then, the principal offers a bonus in case of disclosure to incentivize monitoring. However, the agent's flat compensation region remains present to save on retaliation costs. One interpretation of this result is that there are two reports on the performance of the agent: one shared only with the principal, fully revealing the evidence, and another also shared with the agent, which conceals information in case the realized signal is in $[x_r, \tilde{x})$.

6 Discussion and Conclusion

Relationships inside firms are not without conflict. Often unsatisfied employees retaliate against their supervisors and organizations. This paper analyzes how the possibility of retaliation shapes optimal incentive contracts. We assume that an agent retaliates whenever her realized payment is below an endogenous benchmark level — the payment under no information. The optimal agent's contract features centrality and leniency, often perceived as the monitor's biases, but here they arise from the principal's optimal disclosure recommendation to the monitor.

We show that tolerating more retaliation allows the principal to be better informed about the agent's effort. One can characterize optimal contracts using a two-step procedure: first, fix the amount of retaliation tolerated and minimize expected payments. Second, choose the optimal level of retaliation. The first step is a classic moral hazard problem with an exogenous information structure. The second can be interpreted as an information acquisition problem where a higher retaliation cutoff generates higher retaliation costs but also more information (in a Blackwell sense) about the agent's effort.

Our model is a first step toward studying the interaction between retaliation and incentive contracts. There is extensive literature and anecdotal evidence documenting retaliation inside organizations. A better understanding of how retaliation patterns interact with monitoring and compensation schemes can improve firm performance and reduce losses due to internal conflicts.

A Appendix A - Proofs

Proof of Lemma 1. Take any mechanism $\left(v_A, v_M, r, d\right)$ that satisfies (IR_A), (IC'_A), (IR_M), (IC_r) and (IC_d). Let $\tilde{x} := \{x \in X : d(x) \neq 1\}$. We construct an alternative mechanism that satisfies all constraints, generates

the same payments to agent and monitor, and has $\hat{d}(x) = 1$ for all $x \in X$.

Let $\left(\hat{v}_A, \hat{v}_M, r, \hat{d}\right)$ be such that

$$\hat{v}_A(s) = \begin{cases} v_A(\emptyset) & \text{if } s = \emptyset, \\ \left[v_A(\emptyset)(1-d(s)) + d(s)v_A(s)\right] & \text{if } s \in \tilde{X}, \\ v_A(s) & \text{otherwise.} \end{cases}$$

$$\hat{v}_M(s) = \begin{cases} v_M(\emptyset) \text{ if } s \in \tilde{X} \cup \{\emptyset\} \\ v_M(s) \text{ otherwise.} \end{cases}$$

The monitor is now willing to disclose any signal realization. The change did not alter the monitor's payments on the equilibrium path. The agent's expected utility conditional on each possible realization of $\mathbf{x}$ also did not change. Instead of paying $v_A(x)$ with probability $d(x)$ and $v_A(\emptyset)$ with the complementary probability, the principal pays the certain equivalent, which is cheaper. If $d(x) = 0$ for all $x \in \tilde{X}$, the payments do not change. □

Proof of Lemma 2. We have proved that $d(x) = 1$ is without loss of generality. The payment suggested implements that disclosure rule and keeps the monitor at his outside option for every possible x. Hence, the principal cannot do better. □

Proof of Proposition 1. As r is not a cutoff function, there must exist two positive measure (with respect to p) sets $X^*, \check{X}$ such that $x^* > \check{x}$ and $v_A(x^*) < 0 \leq v_A(\check{x})$ for all $x^* \in X^*$ and $\check{x} \in \check{X}$. Then, there exists $\varepsilon > 0$ such that there exists a positive measure set $X^*_\varepsilon \subseteq X^*$ such that $v_A(x) < -\varepsilon$ for all $x \in X^*_\varepsilon$.

Let $\hat{S} \subset X^*_\varepsilon$ and $\check{S} \subset \check{X}$ be positive measure subsets with the same measure $k > 0$. That is, let $k = \int_{\hat{S}} p(x|a)dx = \int_{\check{S}} p(x|a)dx > 0$.

Let $n \in \mathbb{N}$ and $\{\hat{S}_1, ..., \hat{S}_n\}$, $\{\check{S}_1, ..., \check{S}_n\}$ be successively finer partitions of $\hat{S}$ and $\check{S}$ such that $\int_{\hat{S}_i} p(x|a)dx = \int_{\check{S}_i} p(x|a)dx = \frac{k}{n}$. Also, let the sets in each partition be ordered, in the sense that all elements of subset i are smaller than all elements of subset $j > i$ for any $i, j \in \{1, ..., n\}$.

First, we approximate v_A by the following sequence of functions

$$v_{An}(s) = \begin{cases} v_A(\emptyset) \text{ if } s = \emptyset, \\ \int_{\hat{S}_i} v_A(x) \frac{p(s|a)}{k/n} dx \text{ if } x \in \hat{S}_i \text{ for } i \in \{1, ...n\}, \\ \int_{\check{S}_i} v_A(x) \frac{p(s|a)}{k/n} dx \text{ if } x \in \check{S}_i \text{ for } i \in \{1, ...n\}, \\ v_A(s) \text{ otherwise.} \end{cases}$$

Note that v_{An} converges almost everywhere to v_A. Also, as v_{An}'s are mean preserving contractions of v_A, they each generate a strictly lower expected payment. We now construct a sequence of $\tilde{\tilde{v}}_{An}$ switching

payments in $\hat{S}$ and $\check{S}$ and satisfying (IR_A) and (IC_A). Define

$$\tilde{\tilde{v}}_{An}(s) = \begin{cases} v_A(\emptyset) \text{ if } s = \emptyset, \\ \int_{\check{S}_i} v_A(x) \frac{p(x|a)}{k/n} dx \text{ if } x \in \hat{S}_i \text{ for } i \in \{1,...n\}, \\ \int_{\hat{S}_i} v_A(x) \frac{p(x|a)}{k/n} dx \text{ if } x \in \check{S}_i \text{ for } i \in \{1,...n\}, \\ v_A(s) \text{ otherwise.} \end{cases}$$

Note that

$$\begin{aligned} &\int_{\hat{S}\cup\check{S}} \Big[\tilde{\tilde{v}}_{An}(x) - v_{An}(x)\Big] s(x|a)p(x|a)dx \\ &= \sum_{i=1}^{n} \Big[\int_{\hat{S}_i} s(x|a)p(x|a)dx - \int_{\check{S}_i} s(x|a)p(x|a)dx\Big] \Big[\int_{\check{S}_i} v_A(x)\frac{p(x|a)}{k/n}dx - \int_{\hat{S}_i} v_A(x)\frac{p(x|a)}{k/n}dx\Big] \\ &\geq \varepsilon \Big[\int_{\hat{S}} s(x|a)p(x|a)dx - \int_{\check{S}} s(x|a)p(x|a)dx\Big] > 0. \end{aligned}$$

The first inequality is direct from the construction of v_{An} and $\tilde{\tilde{v}}_{An}$. The square brackets in the third line is strictly positive because $\hat{S} > \check{S}$. Note that for n sufficiently large

$$\begin{aligned} &\int_{\hat{S}\cup\check{S}} \Big[\tilde{\tilde{v}}_{An}(x) - v_A(x)\Big] s(x|a)p(x|a)dx \\ &= \int_{\hat{S}\cup\check{S}} \Big[\tilde{\tilde{v}}_{An}(x) - v_{An}(x)\Big] s(x|a)p(x|a)dx + \int_{\hat{S}\cup\check{S}} \Big[v_{An}(x) - v_A(x)\Big] s(x|a)p(x|a)dx \\ &\geq \varepsilon \Big[\int_{\hat{S}} s(x|a)p(x|a)dx - \int_{\check{S}} s(x|a)p(x|a)dx\Big] + \int_{\hat{S}\cup\check{S}} \Big[v_{An}(x) - v_A(x)\Big] s(x|a)p(x|a)dx. \end{aligned}$$

As v_{An} converges to v_A, for n sufficiently large

$$\int_{\hat{S}\cup\check{S}} \Big[\tilde{\tilde{v}}_{An}(x) - v_A(x)\Big] s(x|a)p(x|a)dx > 0.$$

Hence, there exists $\gamma_n \in (0,1)$ such that

$$\gamma_n \int_X \tilde{\tilde{v}}_{An}(x)s(x|a)p(x|a)dx = \int_X v_A(x)s(x|a)dx.$$

Let $\alpha_n \in \mathbb{R}_+$ be such that

$$\alpha_n = m[1-\gamma_n] \int_X \tilde{\tilde{v}}_{An}(x)p(x|a)dx.$$

Define

$$\tilde{v}_{An}(s) = \begin{cases} \big(v_A(\emptyset) + \alpha_n\big) \text{ if } s = \emptyset, \\ \gamma_n \tilde{\tilde{v}}_{An}(s) \text{ otherwise.} \end{cases}$$

By construction, for sufficiently large n, $\tilde{v}_{An}$ satisfies (IR_A), (IC_A), and is a mean preserving contraction of v_A. Hence, it strictly reduces expected payments. One can complete the new mechanism definition by letting

$$\tilde{r}(x) = 1 \text{ if and only if } \tilde{v}_A(x) < 0,$$

and letting $\tilde{v}_M$ be defined as in Lemma 2. □

Proof of Lemma 3. Take any (v_M, v_A, x_r) as stated. We construct an alternative full disclosure feasible mechanism $(\tilde{v}_M, \tilde{v}_A, x_0)$ with lower implementation costs. Let

$$\gamma := \frac{c_A'(a)}{\int_{X\setminus[x_0,x_r]} v_A(x)s(x|a)p(x|a)dx}.$$

Note that as the original mechanism is feasible, $\gamma \in (0,1)$. Let

$$\alpha := -m \int_{X_0}^{x_r} v_A(x) p(x|a) dx > 0.$$

$$\tilde{v}_A(x) := \begin{cases} 0 \text{ if } x \in [x_0, x_r] \\ \gamma v_A(x) \text{ otherwise,} \end{cases}$$

and

$$\tilde{v}_A(\emptyset) = v_A(\emptyset) + \alpha.$$

By construction, $\tilde{v}_A$ satisfies (IR_A) and (IC_A). Then, the mechanism $(\tilde{v}_M, \tilde{v}_A, x_0)$, where $\tilde{v}_M$ is defined as in Lemma 2, is feasible. Also, as $\tilde{v}_A$ is a mean preserving contraction of v_A, the implementation costs are smaller.

□

Proof of Proposition 2. The first order conditions of problem (6) are:

$$(1-m)\varphi_A'(v_A(\emptyset)) + m \int_X \varphi_A'(v_A(\emptyset) + v_A(x)) p(x|a) dx = \lambda_A; \tag{8}$$

$$\varphi_A'(v_A(\emptyset) + v_A(x)) = \lambda_A + \mu_A s(x|a) + \mu^r(x)[1 - \chi^x_{[\underline{x}, x_r)}]; \tag{9}$$

By equations (8) and (9)

$$\varphi_A'(v_A(\emptyset)) = \lambda_A - \frac{m}{1-m} \int_{x_r}^{\bar{x}} \mu^r(x) p(x|a) dx. \tag{10}$$

We now prove that the multipliers λ_A and μ_A must be strictly positive.

Lemma 4. *$\lambda_A > 0$ and $\mu_A > 0$.*

Proof. By equation (8) we have $\lambda_A > 0$.

We now need to show that $\mu_A > 0$. For the sake of obtaining a contradiction, suppose that $\mu_A = 0$. Hence, for all x

$$\varphi_A'(v_A(\emptyset) + v_A(x)) = \lambda_A + \mu^r(x)[1 - \chi^x_{[\underline{x}, x_r)}].$$

For all x such that $v_A(x) > 0$, then $\mu^r(x) = 0$ and $\varphi_A'(v_A(\emptyset) + v_A(x)) = \lambda_A$.

For all x such that $v_A(x) = 0$,

$$\varphi_A'(v_A(\emptyset) + v_A(x)) = \lambda_A + \mu^r(x)[1 - \chi^x_{[\underline{x}, x_r)}] = \lambda_A - \frac{m}{1-m} \int_{x_r}^{\bar{x}} \mu^r(x) p(x|a) dx = \varphi_A'(v_A(\emptyset))$$

which can only hold if $\mu^r(x) = 0$ almost everywhere. Thus, $\varphi_A'(v_A(\emptyset) + v_A(x)) = \varphi_A'(v_A(\emptyset))$ almost everywhere. Then, (IC_A) cannot be satisfied. A contradiction. □

Define $\tilde{x} = inf\{x \in X : v_A(x) > 0\}$. By definition $x_r \leq \tilde{x}$. For $x > \tilde{x}$, $\mu^r(x) = 0$. For $x \in (x_r, \tilde{x})$

$$\mu^r(x) = -\mu_A s(x|a) - \frac{m}{1-m} \int_{x_r}^{\tilde{x}} \mu^r(x) p(x|a) dx. \tag{11}$$

$\tilde{x}$ is given by $\mu^r(\tilde{x}) = 0$.

Lemma 5. *For a given x_r, $\tilde{x}$ is uniquely characterized by*

$$s(\tilde{x}|a) = \frac{m}{1-m\left[1-\int_{x_r}^{\tilde{x}} p(x|a)dx\right]} \int_{x_r}^{\tilde{x}} s(x|a)p(x|a)dx \tag{12}$$

Proof. Take equation (11) multiply both sides by $p(x|a)$ and integrate with respect to x in $(x_r, \tilde{x})$. We then have that,

$$\int_{x_r}^{\tilde{x}} \mu^r(x)p(x|a)dx = \frac{-(1-m)}{1-m\left[1-\int_{x_r}^{\tilde{x}} p(x|a)dx\right]} \mu_A \int_{x_r}^{\tilde{x}} s(x|a)p(x|a)dx. \tag{13}$$

Replace (13) into (11) and equate it to 0. We then get the equation in the lemma. It remains to show that for each $x_r \in [\underline{x}, x_0]$ there is a unique $\tilde{x}$ satisfying equation (7). Note that

$$s(x_r|a) - \frac{m}{1-m\left[1-\int_{x_r}^{x_r} p(x|a)dx\right]} \int_{x_r}^{x_r} s(x|a)p(x|a)dx = s(x_r|a) < 0.$$

Also,

$$s(x_0|a) - \frac{m\int_{x_r}^{x_0} s(x|a)p(x|a)dx}{1-m\left[1-\int_{x_r}^{x_0} p(x|a)dx\right]} = -\frac{m\int_{x_r}^{x_0} s(x|a)p(x|a)dx}{1-m\left[1-\int_{x_r}^{x_0} p(x|a)dx\right]} > 0.$$

By the intermediate value theorem there exists a $\tilde{x} \in (x_r, x_0)$ such that the equality is satisfied.

Rearranging (7), we can write it as

$$s(\tilde{x}|a)\left[(1-m)+m\int_{x_r}^{\tilde{x}} p(x|a)dx\right] - m\int_{x_r}^{\tilde{x}} s(x|a)p(x|a)dx = 0. \tag{14}$$

Taking the derivative of the lhs with respect to $\tilde{x}$ we get

$$\frac{\partial s(\tilde{x}|a)}{\partial x}\left[(1-m)+m\int_{x_r}^{\tilde{x}} p(x|a)dx\right] > 0.$$

Hence, there exists only one $\tilde{x}$. □

We then replace $\mu^r(x)$ and $\tilde{x}$ in the first-order conditions and get the expressions stated in the proposition.

Note that given x_r we can directly characterize $\tilde{x}$. Moreover, given $[x_r, \tilde{x})$, the problem is equivalent to a canonical moral hazard problem. Hence, existence is guaranteed by the same arguments as in Moroni and Swinkels (2014).

□

Proof of Corollary 1. When we fully differentiate (12) with respect to x_r and rearrange terms we get

$$\frac{d\tilde{x}}{dx_r}\frac{\partial s(\tilde{x}|a)}{\partial x}\left[1-m\left[1-\int_{x_r}^{\tilde{x}} p(x|a)dx\right]\right] = -(1-m)p(x_r|a)s(x_r|a).$$

As $\partial s(\tilde{x}|a)/\partial x > 0$ and $s(x_r|a)$, we have $\frac{d\tilde{x}}{dx_r} < 0$. As the score is increasing in x, we have that $s(\tilde{x}|a)$ is increasing in x_r. □

Proof of Theorem 1. Note that for any x_r the expected score is zero, that is, $\int x dF^{x_r}(x|a) = 0$. Hence, it remains to show that for any $x_r' > x_r''$ and any $t \in \mathbb{R}$

$$\int_{-\infty}^{t} F^{x_r''}(x|a)dx \leq \int_{-\infty}^{t} F^{x_r'}(x|a)dx. \tag{15}$$

Let $\tilde{x}', \tilde{x}''$ be the respective $\tilde{x}$ for x_r', x_r''. First notice that for any $x \leq s(x_r''|a)$

$$F^{x_r'}(x|a) = F^{x_r''}(x|a).$$

Second, note that for any $x \in (s(x_r''|a), s(\tilde{x}''|a))$, $F^{x_r''}(\cdot|a)$ is constant while $F^{x_r'}(\cdot|a)$ initially increases. That is,

$$F^{x_r''}(x|a) = F^{x_r''}(s(x_r''|a)|a) = F^{x_r'}(s(x_r''|a)|a) < F^{x_r'}(x|a).$$

Third, note that for any $x \geq s(\tilde{x}''|a)$

$$F^{x_r''}(x|a) = (1-m) + m\int_{\underline{x}}^{\rho(y,a)} p(x|a)dx \geq F^{x_r'}(x|a).$$

That is, $F^{x_r''}(\cdot|a)$ single crosses $F^{x_r'}(\cdot|a)$ from below. Hence, $F^{x_r'} \gtrsim F^{x_r''}$.

□

Proof of Theorem 2. The first observation is that there exists and optimal retaliation cutoff which is always strictly below x_0. Signal realizations close to x_0 have scores very close to 0. Hence, payments for such realizations have a low impact on effort incentives. However, negative $v_A(x)$'s generate a discrete cost with retaliation. Therefore, the principal is better off by reducing the retaliation cutoff and avoiding such expenses.

Lemma 6. *There exists an optimal retaliation cutoff and it is strictly below x_0. That is, $x_r^* < x_0$.*

Proof of Lemma 6. Let $\mathscr{V}$ be the set of mappings $v_A : X \cup \{\emptyset\} \to \mathbb{R}$ that satisfy (IR_A), (IC_A) and (IC_r). Let $g : \mathscr{V} \times [\underline{x}, x_0] \to \mathbb{R}$ be

$$g(v_A, x_r) := \int_X \left[m\varphi_A\big(v_A(x) + v_A(\emptyset)\big) + (1-m)\varphi_A\big(v_A(\emptyset)\big) + m\kappa_r \int_{\underline{x}}^{x_r} p(t|a)dt + \varphi_M(\bar{u}_M) \right] p(x|a)dx.$$

Note that $g(v_A, .)$ is absolutely continuous for all $v_A \in \mathscr{V}$. Also, note that

$$|g_{X_r}(v_A, x_r)| = m\kappa_r p(x_r|a) \leq \sup_{x \in X}\{m\kappa_r p(x|a)\}.$$

Hence, $C(\cdot, a, m)$ is absolutely continuous. Therefore, there exists $x_r^* \in \underset{x_r \in [\underline{x}, x_0]}{argmin}\big\{C(x_r, a, m)\big\}$.

Take the derivative of $C(x_r, a, m)$ with respect to x_r and evaluate it at x_0. We get

$$\frac{d\mathscr{C}(x_0, a, m)}{dx_r} = \frac{d\mathscr{L}}{dx_r} = mp(x_0|a)\Big[\varphi_M(\bar{u}_M + L_r) - \varphi_M(\bar{u}_M) + c_r\Big] > 0.$$

Hence, reducing x_r strictly lowers implementation costs and $x_r^* < x_0$. □

Parts (i) and (ii): they are directly implied by Lemma 6. Note that as $x_r^* < x_0$ and $\tilde{x} \in (x_r^*, x_0)$, then all realizations $x \in (x_r^*, \tilde{x})$ are pooled with the no-disclosure payment. Hence, such signals receive the same payments (centrality), and such payments reflect a score $s(\tilde{s}|a)$ larger than the score of their actual performance (leniency).

Part (iii): note that the payment to no disclosure is associated with $s(\tilde{x}|a)$, which is strictly negative since $\tilde{x} < x_0$.

Part (iv): note that the necessary first order condition for minimizing $C(x_r, a, m)$ over x_r is given by

$$\kappa_r + \Big[\varphi_A\Big(v_A(x_r^*) + v_A(\emptyset)\Big) - \varphi_A\Big(v_A(\emptyset)\Big)\Big] - v_A(x_r^*)[\lambda_A + \mu_A s(x_r^*|a)] \geq 0 \quad \Big(= \text{ if } x_r^* > \underline{x}\Big).$$

Therefore, if

$$\kappa_r < v_A(\underline{x})[\lambda_A + \mu_A s(\underline{x}|a)] - \Big[\varphi_A\Big(v_A(\underline{x}) + v_A(\emptyset)\Big) - \varphi_A\Big(v_A(\emptyset)\Big)\Big],$$

the retaliation cutoff must be interior[10]. For that to be possible, we must show that the right-hand-side of the equation above is strictly positive. Note that as φ_A is strictly convex and $v_A(\underline{x}) < 0$, we have

$$\Big[\varphi_A\Big(v_A(\underline{x}) + v_A(\emptyset)\Big) - \varphi_A\Big(v_A(\emptyset)\Big)\Big] < v_A(\underline{x})\varphi_A'\Big(v_A(\underline{x}) + v_A(\emptyset)\Big) = v_A(\underline{x})\big[\lambda_A + \mu_A s(\underline{x}|a)\big].$$

Hence, there exists a strictly positive $\underline{\kappa}$ such that $x_r^* > \underline{x}$, and consequently a stick-and-carrot contract is used.

Part (v): note that λ_A and μ_A are characterized by (IR_A) and (IC_A) binding. As both constraints are continuous in x_r, the multipliers are also continuous functions of x_r. Let

$$K := \max_{x_r \in [\underline{x}, x_0]} \Big\{\Big[\varphi_A\Big(v_A(x_r) + v_A(\emptyset)\Big) - \varphi_A\Big(v_A(\emptyset)\Big)\Big] - v_A(x_r)[\lambda_A + \mu_A s(x_r|a)]\Big\}.$$

The maximum exists because the function inside the brackets is continuous in x_r. If $\kappa_r > K$, then $\frac{\partial \hat{C}(x_r,a)}{\partial x_r} < 0$ for all $x_r \in [\underline{x}, x_0]$. Hence, $x_r^* = \underline{x}$ and a carrot-only contract is optimal. □

Proof of Proposition 3. Note that for a given $m \in (0,1)$, agent's compensation problem remains unaltered. Hence, Proposition 2 characterizes the optimal scheme.

Now We characterize the optimal monitor's compensation. For a given x_r, the principal minimizes:

$$\min_{v_M}\Big\{\int_X \Big[m\varphi_M(v_M(\emptyset) + v_M(x)) + (1-m)\varphi_M(v_M(\emptyset))\Big] p(x|a)dx\Big\}$$

subject to

$$m\int_X \Big\{v_M(x) - \frac{L_r}{p(x|a)}\int_{\underline{x}}^{x_r} p(t|a)dt\Big\}p(x|a)dx = c_M'(m). \qquad (IC_M)$$

$$\int_X \Big\{v_M(\emptyset) + m v_M(x) - m\frac{L_r}{p(x|a)}\int_{\underline{x}}^{x_r} p(t|a)dt\Big\}p(x|a)dx - c_M(m) \geq \bar{u}_M \qquad (IR_M)$$

$$v_M(x)[1 - \chi^x_{[x_f, x_r]}]mp(x|a) \geq mp(x|a)L_r \quad \forall x \in X. \qquad (IC_d)$$

The problem above has a strictly convex objective function and linear constraints. The pointwise solution delivers the result. □

[10] An important remark is that λ_A and μ_A are functions of x_r. We omit this dependency not to overload the notation.

References

Akerlof, G. A. and R. E. Kranton (2005, March). Identity and the economics of organizations. *Journal of Economic Perspectives 19*(1), 9–32.

Aquino, K. and S. Douglas (2003). Identity threat and antisocial behavior in organizations: The moderating effects of individual differences, aggressive modeling, and hierarchical status. *Organizational Behavior and Human Decision Processes 90*(1), 195–208.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review 86*(5), 1549–1575.

Charness, G. and D. I. Levine (2010). When is employee retaliation acceptable at work? evidence from quasi-experiments. *Industrial Relations 49*(4), 499 – 523.

Chassang, S. and G. Padró i Miquel (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of economic studies 86*(6), 2530–2553.

Coviello, D., E. Deserranno, and N. Persico (2022). Counterproductive worker behavior after a pay cut. *Journal of the European Economic Association 20*(1), 222–263.

Gale, D. and M. Hellwig (1985, 10). Incentive-Compatible Debt Contracts: The One-Period Problem. *The Review of Economic Studies 52*(4), 647–663.

Georgiadis, G. and B. Szentes (2020). Optimal monitoring design. *Econometrica 88*(5), 2075–2107.

Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology*, 561–568.

Grossman, S. J. and O. D. Hart (1983). An analysis of the principal-agent problem. *Econometrica 51*(1), 7–45.

Hart, O. and J. Moore (1998). Default and renegotiation: A dynamic model of debt. *The Quarterly journal of economics 113*(1), 1–41.

Hart, O. and J. Moore (2008). Contracts as reference points. *The Quarterly journal of economics 123*(1), 1–48.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*(1), 74–91.

Jewitt, I., O. Kadan, and J. M. Swinkels (2008). Moral hazard with bounded payments. *Journal of Economic Theory 143*(1), 59–82.

Krueger, A. and A. Mas (2004). Strikes, scabs, and tread separations: Labor strife and the production of defective bridgestone/firestone tires. *Journal of Political Economy 112*(2), 253–289.

Kvaløy, O. and T. E. Olsen (2009, December). Endogenous verifiability and relational contracting. *American Economic Review 99*(5), 2193–2208.

Lang, M. (2019). Communicating subjective evaluations. *Journal of Economic Theory 179*, 163–199.

Li, A. and M. Yang (2020). Optimal incentive contract with endogenous monitoring technology. *Theoretical Economics 15*(3), 1135–1173.

Mas, A. (2008, 01). Labour Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market. *The Review of Economic Studies 75*(1), 229–258.

Mirrlees, J. (1999). The theory of moral hazard and unobservable behaviour: Part i. *Review of Economic Studies 66*(1), 3–21.

Moroni, S. and J. Swinkels (2014). Existence and non-existence in the moral hazard problem. *Journal of Economic Theory 150*, 668–682.

Myerson, R. B. (1982). Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics 10*(1), 67–81.

Skarlicki, D. P. and R. Folger (1997). Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology 82*(3), 434–443.

Sprouse, M. (1992). . *Sabotage in the American Workplace: Anecdotes of Dissatisfaction, Mischief and Revenge*. San Francisco: Pressure Drop Press.

Terkel, S. (1972). *Working: People Talk About What They Do All Day and How They Feel About What They Do*. New York: Pantheon.

Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of economic theory 21*(2), 265–293.