# Online Appendix

# Making Corruption Harder:

## Asymmetric Information, Collusion, and Crime

Juan Ortner — Boston University

Sylvain Chassang* — New York University

February 28, 2017

**Abstract**

This Online Appendix to "Making Corruption Harder: Asymmetric Information, Collusion, and Crime" provides several extensions. We analyze variants of our baseline model allowing for: ex ante and ex post bargaining, extortion from non-criminal agents, more sophisticated contracting, arbitrary bargaining mechanisms. .

KEYWORDS: monitoring, collusion, corruption, asymmetric information, random incentives, prior-free policy evaluation.

*Ortner: jortner@bu.edu, Chassang: chassang@nyu.edu.

# OA Extensions

## OA.1 Alternative timing of decisions

The model in the main text assumes that the monitor and the agent collude after the agent takes action $c \in \{0, 1\}$. This appendix studies the role of random incentives in settings in which the monitor and the agent can collude before the agent chooses her action.

We start by considering a model in which the agent chooses action $c \in \{0, 1\}$ after side-contracting with the monitor, but which is otherwise the same as the model in Section 3. At the side-contracting stage the agent makes a take-it-or-leave-it offer $\tau \geq 0$ to the monitor. If the monitor accepts the agent's offer, she commits to send report $m = 0$ to the principal regardless of the agent's action. Otherwise, if the monitor rejects the agent's offer, she sends the report $m \in \{0, 1\}$ that maximizes her expected payoff. The principal detects false messages with probability $q$. The monitor is compensated with an efficiency wage $w \geq 0$, and losses this wage if the principal detects that the message was false. We assume for now that all monitors have a type $\eta = 0$ and that all agents have type $\pi_A < k$. We relax these assumptions later.

**Lemma OA.1.** *The agent takes action $c = 1$ if and only if the monitor accepts her bribe.*

**Proof.** If the monitor accepts the agent's bribe $\tau$, the agent's payoffs from action $c = 1$ is $\pi_A - \tau$, while her payoff from action $c = 0$ is $-\tau$. If the monitor rejects the agent's bribe, the agent's payoff from $c = 1$ is $\pi_A - k < 0$ (since in this case the monitor will find it optimal to send message $m = 1$), while her payoff from action $c = 0$ is 0. Therefore, the agent takes action $c = 1$ if and only if the monitor accepts her bribe. ■

Lemma OA.1 implies that the monitor's payoff from accepting bribe $\tau$ is $\tau + (1 - q)w$, while her payoff from rejecting the bribe and sending a truthful message is $w$. Therefore, a monitor with wage $w$ accepts bribe $\tau$ if and only if $\tau > qw$.

We now consider the case in which the principal compensates the agent with a determin-

istic wage $w$. The following result generalizes Lemma 1 to the current setting; its proof is identical to the proof of Lemma 1 and hence omitted.

**Lemma OA.2.** *Suppose the principal uses a deterministic wage $w$. Under collusion, the minimum cost of wages needed to induce the agent to take action $c = 0$ is equal to $\frac{\pi_A}{q}$.*

Consider next the case in which the principal randomizes over the monitor's wage. Suppose the principal pays the monitor an efficiency wage drawn from the c.d.f. $F$. Note that the agent's payoff from making an offer $\tau \geq 0$ is $F(\tau/q) \times (\pi_A - \tau) + (1 - F(\tau/q)) \times 0$. Let $\tau_F^*$ be the smallest solution to $\max_\tau F(\tau/q)(\pi_A - \tau)$. For any distribution $F$, the principal's expected payoff is

$$F\left(\frac{\tau_F^*}{q}\right)\pi_P - \gamma_w \mathbb{E}_F[w] - \gamma_q q.$$

Under wage distribution $F$, the monitor accepts the agent's bribe when her wage is lower than $\tau_F^*/q$. In this case, the agent takes action $c = 1$ and the principal incurs cost $\pi_P < 0$.

**Proposition OA.1.** *Assume that the agent and monitor collude before the agent chooses $c \in \{0, 1\}$. Then, the optimal wage distribution $\widetilde{F}^*$ is described by,*

$$\forall w \in \left[0, \frac{\pi_A}{q}\left(1 - e^{\frac{q}{\gamma_w}\frac{\pi_P}{\pi_A}}\right)\right], \quad \widetilde{F}_w^*(w) = \frac{e^{\frac{q}{\gamma_w}\frac{\pi_P}{\pi_A}}\pi_A}{\pi_A - qw}. \tag{O1}$$

*When the principal pays the monitor a wage drawn from $\widetilde{F}_w^*$, the agent takes action $c = 1$ with probability $\widetilde{F}_w^*(0) \in (0, 1)$.*

**Proof.** Consider first distributions $F$ such that $F\left(\frac{\tau_F^*}{q}\right) = 0$. Note that $F\left(\frac{\tau_F^*}{q}\right) = 0$ implies that $0 \geq \max_\tau F(\tau/q)(\pi_A - \tau)$, and so $F(\tau/q) = 0$ for all $\tau < \pi_A$. Therefore, for distributions $F$ such that $F\left(\frac{\tau_F^*}{q}\right) = 0$, the minimum cost of wages is achieved with a distribution that puts all its mass at $w = \pi_A/q$. The principal's payoff under this distribution is $-\gamma_w \frac{\pi_A}{q} - \gamma_q q$. Our arguments below show that such a distribution is never optimal.

Consider next distributions $F$ such that $F\left(\frac{\tau_F^*}{q}\right) > 0$. Since $\tau_F^* \geq 0$ is the optimal offer,

for all $\tau \geq 0$,

$$F\left(\frac{\tau_F^*}{q}\right)(\pi_A - \tau_F^*) \geq F\left(\frac{\tau}{q}\right)(\pi_A - \tau) \Longleftrightarrow F\left(\frac{\tau}{q}\right) \leq F\left(\frac{\tau_F^*}{q}\right)\frac{\pi_A - \tau_F^*}{\pi_A - \tau}. \qquad \text{(O2)}$$

By first order stochastic dominance, an optimal wage distribution $F$ with $F\left(\frac{\tau_F^*}{q}\right) > 0$ must be such that (O2) holds with equality for all $\tau$ such that $F(\tau/q) < 1$.

Next, we show that the optimal distribution $F$ with $F\left(\frac{\tau_F^*}{q}\right) > 0$ must be such that $\tau_F^* = 0$. Let $F$ be such that $\tau_F^* > 0$, and let $\hat{F}$ be an alternative distribution described by: $\hat{F}(0) = F(\tau_F^*/q)$ and $\hat{F}(\tau/q) = \frac{\hat{F}(0)\pi_A}{\pi_A - \tau}$ for all $\tau \in [0, \pi_A(1 - \hat{F}(0))]$. By construction, bribe $\tau = 0$ maximizes $\hat{F}(\tau/q)(\pi_A - \tau)$. Since $\hat{F}(0) = F(\tau_F^*/q)$, the probability that the agent takes action $c = 1$ is the same under $\hat{F}$ than under $F$. Moreover, for all $\tau$ such that $\hat{F}(\tau/q) < 1$, $\hat{F}(\tau/q) = \hat{F}(0)\frac{\pi_A}{\pi_A - \tau} > F(\tau_F^*/q)\frac{\pi_A - \tau_F^*}{\pi_A - \tau} \geq F(\tau/q)$ (where the last inequality follows since offer $\tau_F^*$ is optimal under policy $F$). This implies that $\mathbb{E}_F[w] > \mathbb{E}_{\hat{F}}[w]$, so the principal's payoff is larger under $\hat{F}$ than under $F$.

Using the change in variable $w = \tau/q$, the two paragraphs above imply that the optimal wage distribution $F$ with $F\left(\frac{\tau_F^*}{q}\right) > 0$ is such that $\tau_F^* = 0$ and is described by

$$\forall w \in \left[0, \frac{\pi_A}{q}(1 - F(0))\right], \quad F(w) = \frac{F(0)\pi_A}{\pi_A - qw}.$$

The principal's expected payoff from using this wage distribution is

$$F(0)\pi_P - \gamma_w \mathbb{E}_F[w] - \gamma_q q = F(0)\pi_P - \gamma_w \frac{\pi_A}{q}(1 - F(0) + F(0)\ln F(0)) - \gamma_q q.$$

This expression is strictly concave in $F(0)$, and converges to $-\gamma_w \frac{\pi_A}{q} - \gamma_q q$ as $F(0) \to 0$. Maximizing this expression with respect to $F(0)$ yields $F(0) = e^{\frac{q}{\gamma_w}\frac{\pi_P}{\pi_A}} \in (0, 1)$. Therefore, the optimal wage distribution is given by (O1). ■

Proposition OA.1 shows that random incentives are optimal in this setting. We note

that the principal can improve upon deterministic wages using simpler schemes. Suppose the principal uses a two-wage distribution, paying the monitor wage $\underline{w} = 0$ with probability $x \in [0,1]$ and wage $\overline{w} = \frac{\pi_A}{q}(1-x)$ with probability $1-x$. Under this wage distribution, it is optimal for the agent to make a bribe offer of $\tau = 0$. The principal's payoff under this distribution is $x\pi_p - \gamma_w(1-x)^2\frac{\pi_A}{q} - \gamma_q q$, which is maximized by setting $x = \max\{0, 1+\frac{q}{\gamma_w}\frac{\pi_p}{2\pi_A}\}$.

**Ambiguous optimal policy.** Next, we extend Proposition 2 to this environment. As in Section 4, we assume that monitors and agents are privately informed about their types, with $\eta$ distributed according to c.d.f. $F_\eta$ with density $f_\eta$ and $\pi_A$ distributed according to c.d.f $F_{\pi_A}$ with density $f_{\pi_A}$.

Given wage distribution $F_w$, an agent with type $\pi_A$ offers bribe $\tau$ solving

$$\begin{aligned} U(\pi_A) &= \max_{\tau \in [0,\pi_A]} \text{prob}_{F_w}(qw + \eta < \tau)(\pi_A - \tau) \\ &= \max_{\tau \in [0,\pi_A]} \mathbb{E}_{F_w}[F_\eta(\tau - qw)](\pi_A - \tau). \end{aligned} \tag{O3}$$

Equation (O3) can be used to extend Proposition 2 to this environment. Indeed, whenever $F_\eta$ is strictly concave (strictly convex) over the range $[0, \pi_A]$, the wage profile that minimizes the agent's payoff under any budget $w_0 > 0$ is random (deterministic). Note, however, that these statements relate to the agent's payoff, and not to the probability that the agent is criminal. It is also possible to find conditions on $F_\eta$ under which the crime-minimizing policy is deterministic. For instance, if $F_\eta$ and $f_\eta$ are both strictly convex, and $f_\eta(\tau - qw_0)(\pi_A - \tau)$ is strictly decreasing in $\tau$, then the crime-minimizing policy is deterministic.[1]

[1]Proof: Fix a budget $w_0$ and let $F_w$ be any random policy with $\mathbb{E}_{F_w}[w] = w_0$. Let $\tau_0$ be the highest solution to $\max_\tau(\pi_A - \tau)F_\eta(\tau - qw_0)$ and $\tau_{F_w}$ be the highest solution to $\max_\tau \mathbb{E}_{F_w}[F_\eta(\tau - qw)](\pi_A - \tau)$. Suppose by contradiction that the probability with which the agent is criminal is higher under the deterministic policy than under policy $F_w$, so $F_\eta(\tau_0 - qw_0) \geq \mathbb{E}_{F_w}[F_\eta(\tau_{F_w} - qw)]$. Note that $F_\eta$ strictly convex implies $\tau_0 > \tau_{F_w}$. Then,

$$(\pi_A - \tau_0)f_\eta(\tau_0 - qw_0) = F_\eta(\tau_0 - qw_0) \geq \mathbb{E}_{F_w}[F_\eta(\tau_{F_w} - qw)] = (\pi_A - \tau_{F_w})\mathbb{E}_{F_w}[f_\eta(\tau_{F_w} - q_w)],$$

where the first and last equalities follow since $\tau_0$ and $\tau_{F_w}$ are optimal and satisfy the first-order conditions. Finally, since $\tau_0 > \tau_{F_w}$ and $f_\eta(\tau - qw_0)(\pi_A - \tau)$ is strictly decreasing in $\tau$, the inequality above implies that $(\pi_A - \tau_{F_w})f_\eta(\tau_{F_w} - qw_0) > (\pi_A - \tau_{F_w})\mathbb{E}_{F_w}[f_\eta(\tau_{F_w} - qw)]$, which cannot be since $f_\eta$ is strictly convex. Hence,

**Policy evaluation.** We now show how the policy evaluation results in Section 5 extend to an environment in which the interaction between monitors and agents may be ex-ante or ex-post. In particular, we consider a model in which a fraction $\mu \in (0,1)$ of agents interact with their monitors after taking action $c \in \{0,1\}$, as in the main text, and a fraction $1-\mu$ of agents interact with their monitors before taking action $c \in \{0,1\}$. Fraction $\mu$ is unknown to the principal. We assume that the agent has all the bargaining power at the side-contracting stage, and makes offers with probability 1.[2]

We allow monitors to report failed bribing attempts, in addition to reports of crime: monitors now send crime reports $m_c \in \{0,1\}$ and bribing attempt reports $m_b \in \{0,1\}$. As in the baseline model, an agent who was reported $m_c = 1$ incurs a cost of $k$ if criminal, and a cost $k_0 \leq k$ if not criminal. In addition, an agent who was reported $m_b = 1$ incurs a small fine $\phi > 0$ if she was not reported for crime; if she was reported $m_c = 1$, she incurs cost $k$ if criminal and cost $k_0$ if not criminal. We note that allowing monitors to report bribing attempts is needed to generate variation in the monitors' reports that can be used to evaluate how different policies affect crime among those agents that interact with monitors ex-ante. Indeed, by Lemma OA.1, agents who side-contract with monitors ex-ante take action $c = 1$ if and only if their monitor accepts the bribe. As a result, monitors who interact ex-ante with agents always report $m_c = 0$ regardless of the policy in place.

We start by considering agents who interact with monitors ex-ante. Given wage distribution $F_w$, the expected payoff of an agent with type $\pi_A$ who interacts with her monitor ex-ante and who engages in bribing behavior is

$$\begin{aligned} U_{F_w}^{\mathsf{ante}}(\pi_A) &= \max_{\tau \in [0,\pi_A]} \text{prob}_{F_w}(qw + \eta < \tau)(\pi_A - \tau) + \text{prob}_{F_w}(qw + \eta > \tau)(-\phi) \\ &= \max_{\tau \in [0,\pi_A]} \text{prob}_{F_w}(qw + \eta < \tau)(\pi_A + \phi - \tau) - \phi. \end{aligned}$$

Such an agent will engage in bribing behavior if and only if $U_{F_w}^{\mathsf{ante}}(\pi_A) > 0$; if she engages

the crime-minimizing policy is deterministic.

[2]Our results extend to a setting with probabilistic take-it-or-leave-it offers provided that the monitor observes the agent's type.

in bribing behavior, she takes action $c = 1$ if and only if her bribe is accepted. We note that monitors with type $\eta > 0$ who interact with an agent ex-ante have a strict incentive to report failed bribing attempts. As a result, with probability 1, a monitor who interacts ex-ante with an agent engaging in bribing behavior will report $m_b = 0$ if she accepts the agent's bribe, and $m_b = 1$ if she rejects it. By our arguments above, monitors who interact ex-ante always report $m_c = 0$.

Consider next agents who interact with monitors ex-post. Given policy $F_w$, the expected payoff of a criminal agent of type $\pi_A$ who interacts with her monitor ex-post is

$$U^{\mathsf{post}}_{F_w}(\pi_A) = \pi_A - k + \max_{\tau \in [0,k]} (k - \tau) \text{prob}_{F_w}(qw + \eta < \tau).$$

An agent of type $\pi_A$ who interacts with her monitor ex-post chooses $c = 1$ if and only if $U^{\mathsf{ante}}_{F_w}(\pi_A) > 0$. A monitor who interacts with a criminal agent ex-post reports $m_c = m_b = 0$ if she accepts the bribe, and reports $m_c = m_b = 1$ if she rejects it (by assumption, in the latter case the agent incurs a punishment cost of $k$). On the other hand, a monitor who interacts with a non-criminal agent reports $m_b = m_c = 0$.

We now show how a principal can use reports from failed bribing attempts to perform local policy evaluations on agents who interact with monitors ex-ante. Take as given a wage distribution with c.d.f. $F^0_w$ and density $f^0_w$, and let $f^1_w$ be a policy with $\mathsf{supp}\, f^1_w \subset \mathsf{supp}\, f^0_w$ and $\mathbb{E}_{f^0_w}[w] = \mathbb{E}_{f^1_w}[w]$. For any such policy $f^1_w$ and any $\epsilon \in [0, 1]$, construct the mixture $f^\epsilon_w = (1 - \epsilon) f^0_w + \epsilon f^1_w$.

Given policy $f^\epsilon_w$, we denote by $\overline{R}^b_\epsilon(\pi_A)$ the proportion of monitors who report $m_b = 1$ and $m_c = 0$ among monitors matched with an agent of type $\pi_A$. We denote by $U^{\mathsf{ante}}_{f^\epsilon_w}(\pi_A)$ the payoff of an agent of type $\pi_A$ from engaging in bribing behavior:

$$U^{\mathsf{ante}}_{f^\epsilon_w}(\pi_A) = \max_\tau \text{prob}_{f^\epsilon_w}(qw + \eta < \tau)(\pi_A + \phi - \tau) - \phi$$

Fix a type $\pi_A$ such that $U^{\mathsf{ante}}_{f^0_w}(\pi_A) > 0$. For any $f^1_w$, denote by $\nabla_{f^1_w} U(\pi_A)$ the gradient of the

agent's payoff from bribing in policy direction $f_w^1$:

$$\nabla_{f_w^1} U^{\mathsf{ante}}(\pi_A) = \frac{\partial U^{\mathsf{ante}}_{f_w^\epsilon}(\pi_A)}{\partial \epsilon}\Big|_{\epsilon=0}.$$

For any wage $w \in \mathsf{supp}\, f_w^0$, let $R_0^b(w;\pi_A)$ be the fraction of reports $m_b = 1$ and $m_c = 0$ from monitors with wage $w$ who were matched with agents of type $\pi_A$ under policy $f_w^0$. For any $f_w^1$ with $\mathsf{supp}\, f_w^1 \subset \mathsf{supp}\, f_w^0$, construct counterfactual reports

$$R_0^b(f_w^1;\pi_A) \equiv \mathbb{E}_{f_w^0}\left[R_0^b(w;\pi_A) \times \frac{f_w^1(w)}{f_w^0(w)}\right]. \tag{O4}$$

The following result holds.

**Proposition OA.2.** *For every $\pi_A$ with $U^{\mathsf{ante}}_{f_w^0}(\pi_A) > 0$, there exists a fixed coefficient $\rho(\pi_A) > 0$ such that for all alternative policies $f_w^1$,*

$$\nabla_{f_w^1} U(\pi_A) = \rho(\pi_A)\left[\overline{R}_0^b(\pi_A) - R_0^b(f_w^1;\pi_A)\right].$$

**Proof.** Take as given a policy $f_w^1$. Under wage schedule $f_w^\epsilon$, the payoff of an agent with type $\pi_A$ who engages in bribing behavior is

$$U^{\mathsf{ante}}_{f_w^\epsilon}(\pi_A) = \max_\tau (\pi_A + \phi - \tau)[(1-\epsilon)\mathrm{prob}_{f_w^0}(qw+\eta<\tau) + \epsilon \mathrm{prob}_{f_w^1}(qw+\eta<\tau)] - \phi.$$

Let $\tau_0$ be the highest solution to this maximization problem for $\epsilon = 0$. By the Envelope Theorem,

$$\begin{aligned}
\nabla_{f_w^1} U(\pi_A) &= (\pi_A + \phi - \tau_0)\left[\mathrm{prob}_{f_w^1}(qw+\eta<\tau_0) - \mathrm{prob}_{f_w^0}(qw+\eta<\tau_0)\right] \\
&= (\pi_A + \phi - \tau_0)\frac{1}{1-\mu}\left[\overline{R}_0^b(\pi_A) - R_0^b(f_w^1;\pi_A)\right],
\end{aligned} \tag{O5}$$

where $1-\mu \in (0,1)$ is the fraction of agents that interact with monitors ex-ante. The second equality above follows from two observations. First, mean reports of failed bribing

attempts $\overline{R}_0^b(\pi_A)$ are equal to the product of the fraction of agents of type $\pi_A$ who interact with monitors ex-ante times the probability that their equilibrium bribes are refused:

$$\overline{R}_0^b(\pi_A) = (1-\mu) \times [1 - \text{prob}_{f_w^0}(qw + \eta < \tau_0)].$$

Second, for any $\tilde{w} \in \mathsf{supp}\, f_w^0$, mean reports $R_0^b(w; \pi_A)$ are equal to the product of the fraction of agents of type $\pi_A$ who interact with monitors ex-ante times the probability that a monitor with wage $\tilde{w}$ refuses their bribe:

$$\begin{aligned} \forall \tilde{w} \in \mathsf{supp}\, f_w^0, \quad & R_0^b(w; \pi_A) = (1-\mu) \times [1 - \text{prob}(q\tilde{w} + \eta < \tau_0)] \\ \Rightarrow \quad & R_0^b(f_w^1, \pi_A) = (1-\mu) \times [1 - \text{prob}_{f_w^1}(qw + \eta < \tau_0)]. \end{aligned}$$

This establishes the result. ■

Proposition OA.2 shows that, under this alternative timing, a principal who can condition on the type of the agent can evaluate how small changes in policy affect the agent's payoff from engaging in bribing behavior. We note that, even when the agent's type is unobservable, the identification result in Proposition OA.2 can still be useful if the principal can condition on a sufficiently rich set of covariates.

Proposition OA.2 can be used to identify directions of policy change that lead to less bribing behavior. We now show how this result can be leveraged to evaluate the effect of local policy changes on crime rates among agents who interact ex-ante.

Let $f_w^0$ be the original policy in place. For any policy $f_w$, we let $\overline{\pi}_A^{\mathsf{ante}}(f_w)$ denote the cutoff such that all agents with $\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w)$ who interact ex-ante engage in bribing behavior under policy $f_w$. Let $f_w^1$ be a policy direction that reduces the set of agents who engage in bribing behavior; i.e., a policy direction with $\nabla_{f_w^1} U(\overline{\pi}_A^{\mathsf{ante}}(f_w^0)) < 0$. Fix $\epsilon > 0$ small, and let $f_w^\epsilon = (1-\epsilon) f_w^0 + \epsilon f_w^1$. Suppose that the principal changes her policy from $f_w^0$ to $f_w^\epsilon$.

Let $\overline{C}_0^{\mathsf{ante}}$ and $\overline{C}_\epsilon^{\mathsf{ante}}$ denote, respectively, the fraction of agents who interact ex-ante that

take action $c = 1$ under policies $f_w^0$ and $f_w^\epsilon$. Let $\overline{R}_0^b$ and $\overline{R}_\epsilon^b$ denote, respectively, the fraction of monitors who report bribing attempts and don't report crime under policies $f_w^0$ and $f_w^\epsilon$; i.e., the fraction of monitors who report $m_b = 1$ and $m_c = 0$. The following result holds.

**Proposition OA.3.** *Fix a policy $f_w^0$ and a policy direction $f_w^1$ such that $\nabla_{f_w^1} U(\overline{\pi}_A^{ante}(f_w^0)) < 0$. Then, there exists a constant $\kappa > 0$ such that*

$$\overline{C}_\epsilon^{ante} - \overline{C}_0^{ante} \leq \kappa \times [\overline{R}_0^b - \overline{R}_\epsilon^b].$$

**Proof.** For every type $\pi_A$ and any policy $f_w$, let $\tau(\pi_A; f_w)$ denote the bribe that agents of type $\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w)$ who interact with monitors ex-ante offer under policy $f_w$; i.e., $\tau(\pi_A; f_w)$ maximizes $\text{prob}_{f_w}(qw + \eta < \tau)(\pi_A + \phi - \tau)$. Note that an agent of type $\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w)$ who interacts ex-ante takes action $c = 1$ only when her bribe is accepted. Then, for $x \in \{0, \epsilon\}$, the fraction of agents who interact ex-ante that take action $c = 1$ under policy $f_w^x$ is

$$\overline{C}_x^{\mathsf{ante}} = (1 - F_{\pi_A}(\overline{\pi}_A^{\mathsf{ante}}(f_w^x))) \times \mathbb{E}_{F_{\pi_A}}[\text{prob}_{f_w^x}(qw + \eta < \tau(\pi_A; f_w^x))|\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w^x)]. \quad \text{(O6)}$$

Note next that, under policy $f_w$, an agent of type $\pi_A$ who interacts ex-ante gets reported for a failed bribing attempt with probability $\mathbf{1}_{\{\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w)\}} \times (1 - \text{prob}_{f_w}(qw + \eta < \tau(\pi_A; f_w)))$. Moreover, only monitors who interact with agents ex-ante send reports $m_b = 1$ and $m_c = 0$.[3] Thus, for $x \in \{0, \epsilon\}$, the share of monitors reporting $m_b = 1$ and $m_c = 0$ under policy $f_w^x$ is

$$\begin{aligned}\overline{R}_x^b &= (1 - \mu) \times (1 - F_{\pi_A}(\overline{\pi}_A^{\mathsf{ante}}(f_w^x))) \times \mathbb{E}_{F_{\pi_A}}[1 - \text{prob}_{f_w^x}(qw + \eta < \tau(\pi_A; f_w^x))|\pi_A > \overline{\pi}_A^{\mathsf{ante}}(f_w^x)] \\ &= (1 - \mu) \times [(1 - F_{\pi_A}(\overline{\pi}_A^{\mathsf{ante}}(f_w^x)) - \overline{C}_x^{\mathsf{ante}}], \quad \text{(O7)}\end{aligned}$$

where we used equation (O6). Since policy direction $f_w^1$ is such that $\nabla_{f_w^1} U(\overline{\pi}_A^{\mathsf{ante}}(f_w^0)) < 0$, it

[3] Monitors who interact with agents ex-post send either $m_b = m_c = 0$ or $m_b = m_c = 1$.

follows that $\overline{\pi}_A^{\mathsf{ante}}(f_w^0) < \overline{\pi}_A^{\mathsf{ante}}(f_w^\epsilon)$. Using this together with equation (O7) yields

$$\overline{C}_\epsilon^{\mathsf{ante}} - \overline{C}_0^{\mathsf{ante}} \leq \frac{1}{1-\mu}[\overline{R}_0^b - \overline{R}_\epsilon^b],$$

which establishes the result. ■

We end this section by noting that the principal can perform local policy evaluations on agents who interact ex-post using reports of crime. Indeed, since reports $m_c = 1$ come exclusively from monitors who interact with agents ex-post, Proposition 4 continues to hold in this setting. This, combined with Propositions OA.2 and OA.3, allows the principal to find policy directions that reduce overall crime rates.

## OA.2 Extortion

This section shows how our results extend to settings in which the monitor can extort transfers from non-criminal agents by committing to send a false report. The framework we consider is essentially the same as in Section 4. The only difference is that a monitor who makes an offer at the side-contracting stage can commit to sending a false report if the agent rejects her proposal. A report $m = 1$ triggers an exogenous judiciary process that imposes an expected cost $k > \pi_A$ on criminal agents and an expected cost $k_0 \in (0, k]$ on non-criminal agents.

**Lemma OA.3.** *If the monitor acts as proposer when the agent is non-criminal, she demands a bribe $\tau = k_0$ if her type is $\eta < k_0$, and she demands no bribe (i.e. $\tau = 0$) if her type is $\eta \geq k_0$. A non-criminal agent accepts any offer $\tau \leq k_0$.*

**Proof.** Suppose the monitor makes an offer $\tau$ to a non-criminal agent and commits to sending a false message if her proposal is rejected. In this case, it is optimal for a non-criminal agent to accept the offer if and only if $\tau \leq k_0$: her payoff from accepting such an offer is $-\tau$, while her payoff from rejecting the offer is $-k_0$. The monitor's payoff from making an offer

$\tau \in (0, k_0]$ is $\tau - \eta$, while her payoff from not demanding a bribe is 0. A type $\eta$ monitor finds it optimal to make an offer $\tau = k_0$ if only if $\eta < k_0$. ■

Lemma OA.3 implies that the payoff of a non-criminal agent is $-(1-\lambda)k_0 F_\eta(k_0)$. On the other hand, by the same arguments as in Section 4, the payoff of a criminal agent of type $\pi_A$ is $\pi_A - k + \lambda \max_\tau (k-\tau)\text{prob}(qw + \eta < \tau)$. Therefore, when the monitor can commit to sending a false report, an agent of type $\pi_A$ will take action $c = 0$ if only if

$$\pi_A - (k - (1-\lambda)k_0 F_\eta(k_0)) + \lambda \max_{\tau \in [0,k]} (k-\tau)\text{prob}(qw + \eta < \tau) \leq 0.$$

From the principal's perspective, the possibility of extortion by the monitor reduces the effective punishment cost that a criminal agent incurs when the monitor sends report $m = 1$ to $k - (1-\lambda)k_0 F_\eta(k_0)$. Note that this term does not depend on the distribution of wages. Hence, all the results in Sections 4 and 5 continue to hold when the monitor can commit to sending a false message.

## OA.3 Efficient contracting between the principal and monitor

Throughout the paper we assume that the principal compensates the monitor with an efficiency wage contract. This appendix shows that random incentives continue to be valuable when we allow for arbitrary contracts. We consider the same environment as in Section 3, with one minor modification: we impose a participation constraint that the agent's payoff cannot be negative. We stress, however, that the results in the main text would remain unchanged if we added this constraint.[4] We also assume that the cost $k_0$ that a non-criminal agent expects from the judiciary is strictly positive.[5]

[4]Indeed, when the monitor is compensated with an efficiency wage $w \geq 0$ the agent can guarantee herself a payoff of 0 by taking action $c = 0$. When we allow for arbitrary contracts, the agent's participation constraint rules out wage structures under which the agent needs to bribe the monitor to get a favorable report after taking action $c = 0$.

[5]As in the main text, we assume that the probability $q$ of detecting a false report of $m = 0$ when the agent took action $c = 1$ is the same as the probability of detecting a false report $m = 1$ when the agent took

Let $s \in \{\emptyset, f\}$ denote the signal that the principal observes by scrutinizing the monitor's report: the principal observes signal $s = f$ when she detects that the monitor's report is false, and observes signal $s = \emptyset$ otherwise.[6] The principal offers a wage contract $w(m, s)$ to the monitor, which determines the monitor's compensation as a function of the report she sends and the principal's signal. By limited liability, $w(m, s) \geq 0$ for all $(m, s) \in \{0, 1\} \times \{\emptyset, f\}$.

We begin by noting that a monitor who is compensated with contract $w(m, s)$ accepts a bribe $\tau$ from a criminal agent if and only if $\tau > w(1, \emptyset) - (1 - q)w(0, \emptyset) - qw(0, f)$.

**Lemma OA.4.** *Let $w(m, s)$ be a contract that induces the monitor to send message $m = 0$ when the agent takes action $c = 0$ and offers bribe $\tau = 0$. Then, it must be that $w(0, \emptyset) \geq (1 - q)w(1, \emptyset) + qw(1, f)$.*

**Proof.** When the agent takes action $c = 0$ and offers bribe $\tau = 0$, the monitor's payoff from sending message $m = 0$ is $w(0, \emptyset)$, while her payoff from sending message $m = 1$ is $(1 - q)w(1, \emptyset) + qw(1, f)$. The monitor sends message $m = c = 0$ if and only if $w(0, \emptyset) \geq (1 - q)w(1, \emptyset) + qw(1, f)$. ■

**Lemma OA.5.** *Under an optimal incentive scheme (either deterministic or random), a principal who wants to induce the agent to take action $c = 0$ offers the monitor contracts $w(m, s)$ with $w(0, \emptyset) = (1 - q)w(1, \emptyset)$ and $w(m, f) = 0$ for $m = 0, 1$.*

**Proof.** Suppose the incentive scheme induces the agent to take action $c = 0$ and satisfies the agent's participation constraint. By Lemma OA.4, any contract $w(m, s)$ that the principal offers to the monitor with positive probability must satisfy $w(0, \emptyset) \geq (1-q)w(1, \emptyset)+qw(1, f)$; otherwise the agent's expected payoff from action $c = 0$ would be strictly negative, either because with positive probability the monitor sends a false report $m = 1$, or because the

action $c = 0$. Our results remain qualitatively unchanged if we allow these two probabilities to be different.

[6] When the monitor sends report $m \neq c$, the principal observes signal $s = f$ with probability $q$ and signal $s = \emptyset$ with probability $1 - q$. When the monitor sends report $m = c$, the principal observes signal $s = \emptyset$ with probability 1.

agent needs to bribe the monitor for a report $m = 0$. In either case, this would violate the agent's participation constraint.

This implies that under an optimal incentive scheme that induces the agent to take action $c = 0$, on the equilibrium path the monitor sends report $m = 0$ and receives a wage $w(0, \emptyset)$. If $w(0, \emptyset) > (1 - q)w(1, \emptyset) + qw(1, f)$ for some contract $w(m, s)$ that is offered with positive probability, the principal would be strictly better-off by reducing $w(0, \emptyset)$ as this would reduce wage payments and would also increase the cost of bribing the monitor.

By limited liability it must be that $w(m, f) \geq 0$ for $m = 0, 1$. Setting $w(0, f) = 0$ is optimal as it increases the cost of bribing the monitor. Finally, since $w(0, \emptyset) = (1 - q)w(1, \emptyset) + qw(1, f)$, setting $w(1, f) = 0$ reduces the wage $w(0, \emptyset)$ that the principal pays on the equilibrium path and also increases the cost of bribing the monitor. ■

We now consider the case in which the principal compensates the agent with a deterministic contract $w(m, s)$. The following result generalizes Lemma 1 to the current setting.

**Lemma OA.6.** *Suppose the principal uses a deterministic contract $w(m, s)$. Under collusion, the minimum cost of wages needed to induce the agent to be non-criminal is equal to $\frac{1-q}{2-q}\frac{\pi_A}{q}$.*

**Proof.** A monitor with contract $w(m, s)$ accepts a bribe $\tau$ from a criminal agent if and only if $\tau > w(1, \emptyset) - (1-q)w(0, \emptyset) - qw(0, f) = w(1, \emptyset) - (1-q)w(0, \emptyset)$, where the equality follows from OA.5. The agent's payoff from taking action $c = 1$ is then $\pi_A - \min\{k, w(1, \emptyset) - (1 - q)w(0, \emptyset)\}$, while her payoff from taking action $c = 0$ is 0. To induce the agent to take action $c = 0$, it must be that $w(1, \emptyset) - (1-q)w(0, \emptyset) \geq \pi_A$. By Lemma OA.5, $w(0, \emptyset) = (1-q)w(1, \emptyset)$, so the previous inequality yields $w(0, \emptyset) \geq \frac{1-q}{2-q}\frac{\pi_A}{q}$. ■

Consider next the case in which the principal randomizes over the monitor's contract $w(m, s)$. By Lemma OA.5, it is optimal for the principal to offer contracts $w(m, s)$ such that $w(0, \emptyset) = (1 - q)w(1, \emptyset)$ and $w(m, f) = 0$ for $m = 0, 1$. Therefore, it is without loss

of optimality to focus on distributions over wages $w(0,\emptyset)$, with the understanding that a contract with $w(0,\emptyset) = w \geq 0$ has $w(1,\emptyset) = \frac{w}{1-q}$ and $w(m,f) = 0$ for $m = 0,1$.

The following result generalizes Proposition 1 to the current setting.

**Proposition OA.4.** *Under collusion, it is optimal for the principal to use random contracts. The cost-minimizing distribution $\hat{F}_w^*$ over wages $w(0,\emptyset)$ that induces the agent to be non-criminal is described by*

$$\forall w \in \left[0, \frac{\pi_A}{q}\frac{1-q}{2-q}\right], \quad \hat{F}_w^*(w) = \frac{k-\pi_A}{k-qw\frac{2-q}{1-q}}. \tag{O8}$$

*The corresponding cost of wages $\hat{W}^*(\pi_A) \equiv \mathbb{E}_{\hat{F}^*}[w]$ is*

$$\hat{W}^*(\pi_A) = \frac{1-q}{2-q}\frac{\pi_A}{q}\left[1 - \frac{k-\pi_A}{\pi_A}\log\left(1+\frac{\pi_A}{k-\pi_A}\right)\right] < \frac{1-q}{2-q}\frac{\pi_A}{q}\frac{\pi_A}{k}. \tag{O9}$$

**Proof.** By our arguments above, a monitor with contract $w(m,s)$ accepts a bribe $\tau$ from a criminal agent if and only if $\tau > w(1,\emptyset) - (1-q)w(0,\emptyset) - qw(0,f) = \frac{2-q}{1-q}qw(0,\emptyset)$, where the last equality follows since $w(1,\emptyset) = \frac{w(0,\emptyset)}{1-q}$ and $w(m,f) = 0$ for $m = 0,1$ (Lemma OA.5). A distribution $F$ over wages $w(0,\emptyset)$ induces the agent to take action $c = 0$ if and only if, for every bribe offer $\tau \geq 0$, $\pi_A - k + (k-\tau)\text{prob}(\tau > \frac{2-q}{1-q}qw) \leq 0$, or equivalently, if and only if, for every $\tau \geq 0$, $F\left(\frac{\tau}{q}\frac{1-q}{2-q}\right) \leq \frac{k-\pi_A}{k-\tau}$. Using the change in variable $w = \frac{\tau}{q}\frac{1-q}{2-q}$, we obtain that wage distribution $F$ induces the agent to take action $c = 0$ if and only if,

$$\forall w \in \left[0, \frac{\pi_A}{q}\frac{1-q}{2-q}\right], \quad F(w) \leq \frac{k-\pi_A}{k-qw\frac{2-q}{1-q}}. \tag{O10}$$

By first-order stochastic dominance, it follows that in order to minimize expected wages, the optimal distribution must satisfy (O10) with equality. This implies that the optimal wage distribution is described by (O8). Expected cost expression (O9) follows from integration and straightforward computations. ■

## OA.4 Arbitrary bargaining

The model of Sections 3 and 4 simplifies the side-contracting stage by assuming take-it-or-leave-it offers. This appendix shows that random wages remain valuable under arbitrary bargaining mechanisms. We study a model in which the monitor and the agent can use any individually rational and incentive compatible mechanism at the side-contracting stage, but that is otherwise identical to the basic model in Section 3.

By the revelation principle, we can restrict attention to mechanisms under which the monitor announces her private information (i.e. her wage) and this announcement determines the bargaining outcome. Such a bargaining mechanism is characterized by two functions: (i) $P(w)$, the probability with which monitor and agent reach an agreement when the monitor's wage is $w$; and (ii) $\tau(w)$, the expected transfer from the agent to the monitor when the monitor's wage is $w$. The monitor commits to send message $m = 0$ if there is an agreement. If there is no agreement, the monitor sends the message that maximizes her final payoff (i.e., she sends a truthful message).

Given a wage schedule $F$ and a mechanism $(P, \tau)$, the agent's expected payoff from crime is $U_A = \pi_A - k + \int (P(w)k - \tau(w))\, dF(w)$. The individual rationality constraint of a criminal agent is $U_A \geq \pi_A - k$, since a criminal agent can guarantee $\pi_A - k$ by not participating in the mechanism.

The payoff that a monitor with wage $w$ who announces wage $w'$ gets under mechanism $(P, \tau)$ when the agent is criminal is $\tilde{U}_M(w, w') = \tau(w') + (1 - P(w')q)w$. By incentive compatibility, $U_M(w) \equiv \tilde{U}_M(w, w) \geq \tilde{U}_M(w, w')$ for all $w' \neq w$. By individual rationality, $U_M(w) \geq w$ for all $w$, since a monitor with wage $w$ obtains a payoff of $w$ by not participating in the mechanism and sending a truthful report.

Given a mechanism $(P, \tau)$ and a wage distribution $F$, the weighted sum of the agent's and monitor's payoff when the agent is criminal is

$$(1 - \lambda) \int U_M(w) dF(w) + \lambda U_A, \tag{O11}$$

where the weight $\lambda \in [0,1]$ represents the monitor's bargaining power. For every wage schedule $F$ and every $\lambda \in [0,1]$, let $\Gamma(F,\lambda)$ be the set of incentive compatible and individually rational bargaining mechanisms that maximize (O11). We assume that, at the side-contracting stage, the monitor and the agent use a bargaining mechanism in $\Gamma(F,\lambda)$. Let $\tilde{U}_A(F,\lambda)$ be the lowest utility that a criminal agent gets under a bargaining mechanism in $\Gamma(F,\lambda)$. The agent has an incentive to be non-criminal if $\tilde{U}_A(F,\lambda) \leq 0$.

The following result generalizes Proposition 1 to this setting.

**Proposition OA.5.** *Suppose that, at the collusion stage, the monitor and the agent use an incentive compatible and individually rational mechanism that maximizes* (O11).

*(i) If $\lambda \in (1/2,1]$, the cost minimizing wage distribution $\tilde{F}_w^*$ that induces the agent to be non-criminal is described by*

$$\forall w \in [0,\pi_A/q], \quad \tilde{F}_w^*(w) = \left(\frac{k-\pi_A}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}. \tag{O12}$$

*(ii) If $\lambda \in [0,1/2]$, the cost minimizing wage distribution $\tilde{F}_w^*$ that induces the agent to be non-criminal has $\tilde{F}_w^*(0) = 1$.*

**Proof.** By standard arguments, any incentive compatible mechanism $(P,\tau)$ must satisfy: (i) $P(w)$ is decreasing, and (ii) $U_M'(w) = 1 - qP(w)$ a.e.. This last condition and the monitor's individual rationality constraint (i.e., $U_M(w) \geq w$ for all $w$) imply that $U_M(w) = \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} + w + c$ for some constant $c \geq 0$ (where $\overline{w}$ is the highest wage in the support of $F$). Since $U_M(w) = \tau(w) + (1 - qP(w))w$, $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} + c$. The weighted

sum of payoffs when the agent is criminal is

$$
\begin{aligned}
&(1-\lambda)\int_{\underline{w}}^{\overline{w}} U_M(w)dF(w) + \lambda U_A \\
&= \int_{\underline{w}}^{\overline{w}} \left[(1-\lambda)(\tau(w) + (1-qP(w))w) + \lambda(P(w)k - \tau(w))\right] dF(w) + \lambda(\pi_A - k) \\
&= \int_{\underline{w}}^{\overline{w}} \left[P(w)\lambda\left(k - qw\right) + (1-\lambda)w\right] dF(w) + \lambda(\pi_A - k) + (1-2\lambda)\left(\int_{\underline{w}}^{\overline{w}} qP(w)F(w)dw + c\right).
\end{aligned} \tag{O13}
$$

We use the following lemma.

**Lemma OA.7.** *For all $\lambda \in (1/2, 1]$, the mechanism $(P, \tau)$ that maximizes* (O13) *has: (i) $P(w) = 1$ if $w < w^*$ and $P(w) = 0$ if $w > w^*$ for some $w^* \in [\underline{w}, \overline{w}]$, and (ii) $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$.*

**Proof.** Note first that (O13) is maximized by setting $c = 0$ when $\lambda \in (1/2, 1]$. Moreover, when $\lambda \in (1/2, 1]$ any mechanism $(P, \tau)$ that maximizes (O13) must be such $P(w) = 0$ for all $w \geq k/q$.

We now show that the mechanism that maximize (O13) is such that $P(w)$ only takes values 0 or 1. From above, we know that $P(w) = 0$ for all $w \geq k/q$. Suppose by contradiction that there exists an interval $V \subset [0, k/q]$ such that $P(w) \in (0, 1)$ for all $w \in V$, and let $H \equiv \int_V \lambda(k - qw)dF(w) + (1-2\lambda)\int_V qF(w)dw$. If $H \geq 0$, increasing $P(w)$ over this interval (subject to the constraint that $P$ is decreasing) makes (O13) larger. If $H < 0$, decreasing $P(w)$ over this interval (subject to the constraint that $P$ is decreasing) also makes (O13) larger. Such improvements are exhausted when $P(w)$ only takes values 0 and 1.[7] Since $P(\cdot)$ is decreasing, when $P(\cdot)$ only takes values 0 or 1 there must exist a wage $w^*$ such that $P(w) = 1$ if $w < w^*$ and $P(w) = 0$ if $w > w^*$. Finally, since (O13) is maximized by setting

[7]Note that these changes in $P(w)$ do not conflict with the participation constraints of monitor and agent. Indeed, $U_M(w) = \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} + w \geq w$ for any incentive compatible mechanism $(P, \tau)$. Moreover, for all $w$, $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w} \leq P(w)k$, where the inequality follows since any mechanism that maximizes (O13) has $P(w) = 0$ for all $w \geq k/q$ and since $P(\cdot)$ is decreasing. Hence, $U_A = \pi_A - k + \int(P(w)k - \tau(w))dF(w) \geq \pi_A - k$.

$c = 0$ when $\lambda \in (1/2, 1]$, $\tau(w) = P(w)qw + \int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}$. Since $P(w) = 1$ if $w < w^*$ and $P(w) = 0$ if $w > w^*$, it follows that $\tau(w) = qw^*$ if $w < w^*$ and $\tau(w) = 0$ if $w > w^*$. ■

We now conclude the proof of Proposition OA.5, beginning with point $(i)$. Fix $\lambda \in (1/2, 1]$ and let $F$ be a cost-minimizing wage schedule that induces the agent to be non-criminal. Let $(P, \tau)$ be the mechanism that maximizes the weighted sum of payoffs (O13) under distribution $F$. By Lemma OA.7, $P(w) = \mathbf{1}_{w \leq w^*}$ and $\tau(w) = qw^*\mathbf{1}_{w \leq w^*}$ for some $w^*$. Under this mechanism (O13) becomes

$$\begin{aligned}
&\lambda \left[ F(w^*)k - \int_0^{w^*} qwdF(w) + \pi_A - k \right] + (1-\lambda) \int wdF(w) + (1-2\lambda) \int_0^{w^*} qF(w)dw \\
=&\lambda \left[ F(w^*)(k - qw^*) + \pi_A - k \right] + (1-\lambda) \int wdF(w) + (1-\lambda) \int_0^{w^*} qF(w)dw,
\end{aligned}$$

where we used $\int_0^{w^*} qwdF(w) = qw^*F(w^*) - \int_0^{w^*} qF(w)dw$. Since $(P, \tau)$ maximizes the weighted sum of payoffs, for all $\hat{w} \neq w^*$ it must be that

$$\lambda F(w^*)(k - qw^*) + (1-\lambda) \int_0^{w^*} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda) \int_0^{\hat{w}} qF(w)dw$$

Otherwise, if the inequality did not hold for some $\hat{w} \neq w^*$, the weighted sum of payoffs would be strictly larger under mechanism $(\hat{P}, \hat{\tau})$ with $\hat{P}(w) = 1$ if $w < \hat{w}$ and $\hat{P}(w) = 0$ if $w > \hat{w}$.

For any $\hat{w} \in \mathsf{supp}\, F$, let $(P_{\hat{w}}, \tau_{\hat{w}})$ be the mechanism with $P_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}}$ and $\tau_{\hat{w}}(w) = \mathbf{1}_{\{w \leq \hat{w}\}} q\hat{w}$. Recall that $\Gamma(F, \lambda)$ is the set of bargaining mechanisms that maximize (O13) and that $\tilde{U}_A(F, \lambda)$ is the lowest utility that a criminal agent gets under a mechanism in $\Gamma(F, \lambda)$. By our arguments above,

$$\Gamma(F, \lambda) = \left\{ (P_{\hat{w}}, \tau_{\hat{w}}) : \hat{w} \in \arg\max_{w'} \lambda F(w')(k - qw') + (1-\lambda) \int_0^{w'} qF(w)dw \right\}.$$

Suppose that there exists $w_1$ and $w_2 > w_1$ such that $(P_w, \tau_w) \in \Gamma(F, \lambda)$ for $w = w_1, w_2$. Note that the agent's payoff from being criminal under mechanism $(P_w, \tau_w)$ is $F(w)(k - qw) + \pi_A - k$.

Since $(P_w, \tau_w) \in \Gamma(F, \lambda)$ for $w = w_1, w_2$,

$$\lambda F(w_1)(k - qw_1) + (1-\lambda)\int_0^{w_1} qF(w)dw = \lambda F(w_2)(k - qw_2) + (1-\lambda)\int_0^{w_2} qF(w)dw$$

and so $F(w_2)(k-qw_2) < F(w_1)(k-qw_1)$. This implies that, $\tilde{U}_A(F,\lambda) = F(\tilde{w})(k-q\tilde{w})+\pi_A-k$, where $\tilde{w} \equiv \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. Since $F$ induces the agent to be non-criminal, $\tilde{U}_A(F,\lambda) = F(\tilde{w})(k-q\tilde{w}) + \pi_A - k \leq 0$.

Let $\overline{w}$ be the highest wage in the support of $F$. We now show that, if $F$ is an optimal distribution, it must be that $\overline{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw$. Suppose by contradiction that this is not true, so that $\overline{w} > \tilde{w} = \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. Pick $\epsilon \in (0, \overline{w} - \tilde{w})$ small and let $F^\epsilon$ be a c.d.f. with $F^\epsilon(w) = F(w)$ for all $w < \overline{w} - \epsilon$ and $F^\epsilon(\overline{w} - \epsilon) = 1$. By first-order stochastic dominance, $\mathbb{E}_{F^\epsilon}[w] < \mathbb{E}_F[w]$. By the definition of $\tilde{w}$,

$$\lambda F(\tilde{w})(k-q\tilde{w}) + (1-\lambda)\int_0^{\tilde{w}} qF(w)dw \geq \lambda F(\hat{w})(k-q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw,$$

for all $\hat{w}$, with strict inequality for all $\hat{w} \in (\tilde{w}, \overline{w}]$. Therefore, there exists $\epsilon > 0$ small enough such that, for all $\hat{w}$,

$$\lambda F^\epsilon(\tilde{w})(k-q\tilde{w}) + (1-\lambda)\int_0^{\tilde{w}} qF^\epsilon(w)dw \geq \lambda F^\epsilon(\hat{w})(k-q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF^\epsilon(w)dw$$

This implies that mechanism $(P_{\tilde{w}}, \tau_{\tilde{w}})$ is still optimal under distribution $F^\epsilon$, and so $\tilde{U}_A(F^\epsilon, \lambda) \leq F(\tilde{w})(k-q\tilde{w}) + \pi_A - k \leq 0$. But this cannot be, since $F$ is a cost-minimizing distribution that induces the agent to be non-criminal. Therefore, if $F$ is optimal it must be that $\overline{w} = \sup\{\hat{w} \in \mathsf{supp}\, F : \hat{w} \in \arg\max_{w'} \lambda F(w')(k-qw') + (1-\lambda)\int_0^{w'} qF(w)dw\}$. The agent's payoff from being criminal under mechanism $(P_{\overline{w}}, \tau_{\overline{w}})$ is $k - q\overline{w} + \pi_A - k \leq 0 \iff \overline{w} \geq \frac{\pi_A}{q}$.

By the arguments above, for all $\hat{w} \in [0, \overline{w}]$,

$$\lambda(k - q\overline{w}) + (1-\lambda)\int_0^{\overline{w}} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw$$

$$\iff \lambda(k - q\overline{w}) + (1-\lambda)\int_{\hat{w}}^{\overline{w}} qF(w)dw \geq \lambda F(\hat{w})(k - q\hat{w}) \tag{O14}$$

We now show that, if $F$ is an optimal distribution, (O14) must hold with equality for all $\hat{w} \in [0, \overline{w}]$. Suppose by contradiction that there is an interval $[w_1, w_2] \subset [0, \overline{w})$ such that (O14) is slack for all $\hat{w} \in [w_1, w_2]$. By first-order stochastic dominance, increasing $F(\cdot)$ over $[w_1, w_2]$ (subject to the constraint that $F$ is increasing) reduces expected wage payments. Moreover, increasing $F(\cdot)$ over $[w_1, w_2]$ relaxes (O14) for all $\hat{w} < w_1$ and does not affect (O14) for all $\hat{w} > w_2$. This implies that mechanism $(P_{\overline{w}}, \tau_{\overline{w}})$ still maximizes the weighted sum of payoffs (O13) after increasing $F(\cdot)$ slightly over $[w_1, w_2]$, and so the agent's payoff from being criminal is $k - q\overline{w} + \pi_A - k \leq 0$. But this cannot be, since $F$ is a cost-minimizing distribution that induces the agent to be non-criminal. Therefore, if $F$ is optimal, (O14) must hold with equality for all $\hat{w} \leq \overline{w}$.

Since (O14) holds with equality for all $\hat{w} \leq \overline{w}$, $\lambda F(\hat{w})(k - q\hat{w}) + (1-\lambda)\int_0^{\hat{w}} qF(w)dw$ is constant over $[0, \overline{w}]$. Differentiating this expression with respect to $\hat{w}$, it must be that

$$F'(\hat{w})\lambda[k - q\hat{w}] + qF(\hat{w})(1 - 2\lambda) = 0. \tag{O15}$$

The solution to the differential equation (O15) is $F(w) = C\left(\frac{1}{k-qw}\right)^{\frac{2\lambda-1}{\lambda}}$, where $C$ is a constant such that $F(\overline{w}) = 1$; i.e., $C = (k - q\overline{w})^{\frac{2\lambda-1}{\lambda}}$. Finally, by our arguments above, under distribution $F$ the agent will have an incentive to be non-criminal as long as $k - q\overline{w} + \pi_A - k \leq 0 \iff \overline{w} \geq \frac{\pi_A}{q}$. Since the constant $C$ is decreasing in $\overline{w}$, an optimal distribution must have $\overline{w} = \frac{\pi_A}{q}$. Hence, $C = (k - \pi_A)^{\frac{2\lambda-1}{\lambda}}$, so the optimal distribution is (O12).

We now turn to point $(ii)$. When $\lambda \leq 1/2$, the mechanism $(P, \tau)$ that maximizes (O13) must make the constant $c$ as large as possible, subject to the agent's IR constraint; that is, subject to $\pi_A - k + \int [P(w)k - \tau(w)]dF(w) \geq \pi_A - k$. Recall that $\tau(w) = P(w)qw +$

$\int_w^{\overline{w}} qP(\tilde{w})d\tilde{w}+c$. The maximum is achieved by choosing $c$ such that $\int[P(w)k-\tau(w)]dF(w) = 0$. Therefore, for $\lambda \leq 1/2$ the agent's payoff from engaging in crime under a mechanism that maximizes (O13) is $\pi_A - k < 0$, regardless of the wage schedule. This implies that the agent has an incentive to be non-criminal even when $F$ has all its mass at $w = 0$. ■

We end this appendix by noting that the results above generalize to settings in which the agent is privately informed about the benefit $\pi_A$ from crime. Given a wage profile $F_w$, the payoff an agent of type $\pi_A$ gets from taking action $c = 1$ is $U_A(\pi_A) = \pi_A - k + \int (P(w; F_w)k - \tau(w; F_w))\, dF(w)$, where $(P(w), \tau(w)$ is the mechanism that maximizes the weighted sum of payoffs (O13).[8] Since $U_A(\pi_A)$ is increasing in $\pi_A$, agents follow a threshold strategy: for any wage schedule $F_w$, there is a cutoff $\overline{\pi}_A(F_w)$ such that an agent of type $\pi_A$ is criminal if and only if $\pi_A > \overline{\pi}_A(F_w)$. For any cutoff $\pi_A$, Proposition OA.5 characterizes the cheapest wage distribution that attains this cutoff.

[8]Note that, given wage profile $F_w$, the mechanism $(P(w), \tau(w))$ that maximizes (O13) is independent of the agent's type $\pi_A$.