Two Corruptible Monitors

[Latest version **here**]

Christopher Stapenhurst

Abstract

Do more monitors help to deter bribes? I compare optimal solutions for moral hazard problems in which a principal pays two corruptible monitors to provide hard evidence about an agent's action. If the agent can bribe one, but not both, of the monitors to conceal evidence, then the principal might prefer to have each monitor receive a different piece of evidence, because then the agent can never conceal all of it. But if the agent can bribe one or both of the monitors, then the principal always prefers to have one monitor receive all the evidence. Why? When each monitor receives a different piece of evidence, the principal must pay both monitors in order to deter bribes to conceal either piece. But when one monitor receives both pieces of evidence, the same payment that deters bribes to conceal one piece, also deters bribes to conceal the other piece, so the principal pays less overall. These results suggest that coalition formation frictions play an important role in incentive design problems with more than one monitor.

1 Introduction

The possibility of corruption is a serious problem in many incentive design problems. Whenever a principal relies on evidence provided by a third party monitor to reward or punish an agent, there is a risk that the agent might bribe the monitor to conceal the evidence. For example, factory owners bribe pollution inspectors to falsely report compliance (Duflo, Greenstone, Pande, and Ryan, 2013); construction firms bribe local politicians to grant planning permits[1]; firms bribe accountants to publish favourable audits[2]. Some institutions use multiple parties to monitor agents. For example, France now requires firms to be audited by two independent accountants (Vanstraelen, Richard, and R. Francis, 2009); the US False Claims Act rewards private citizens for providing evidence of Fraud to the Department of Justice (Beck, 2000); the US government allows private citizens to work independently of public agencies to provide evidence of environmental breeches (Langpap and Shimshack, 2010). One reason could be that more monitors can provide a large pool of information on which to condition incentive payments (e.g. Holmström, 1979). Another reason could be that competition between monitors may incentivise them to exert monitoring effort (e.g. Polinsky, 1980). A third possibility is that monitors may be able to deter each other from accepting bribes.

I focus on the third channel: do more independent monitors help to deter bribes? To answer this question, I study an incentive design problem in which a principal (she) designs incentives for an agent (he) to take the "right" action, by conditioning rewards for the agent on evidence provided by two monitors (they). Each monitor receives a binary signal. I interpret each signal as a piece of evidence that either materialises (outcome 1), or fails to materialise (outcome 0). For example, a piece of evidence could be a water sample taken to assess whether a nearby factory is emitting a legal amount of pollution. Evidence is informative and incriminating, because it is more likely to materialise when the agent takes the wrong action; but it is also noisy, because it might fail to materialise even though the agent takes the wrong

[1] the Guardian. 2022. Spain's biggest corruption trial ends with 53 people convicted. [online] Available at: ¡https://www.theguardian.com/world/2013/oct/04/spain-corruption-trial-marbella-mayor-convicted¿ [Accessed 1 April 2022].

[2] Encyclopedia Britannica. 2022. Enron scandal — Summary, Explained, History, & Facts. [online] Available at: ¡https://www.britannica.com/event/Enron-scandal¿ [Accessed 1 April 2022].

action, or it might materialise when the agent takes the right action. It is without loss of generality to assume that one monitor's signal ("inferior" evidence) is noisier than the other's ("superior" evidence). Thus there are four possible outcomes no matter what action the agent chooses: no evidence, inferior evidence materialises, superior evidence materialises, both inferior and superior evidence materialise. In general, the principal incentivises the agent to take the right action by rewarding exonerating outcomes, and/or punishing incriminating outcomes. A key result of this is paper will be that one of four agent-reward schemes is optimal. The "easy scheme" gives the agent a small reward whenever the monitors collectively report at most one piece of incriminating evidence; the "moderate scheme" gives the agent a moderate-sized reward so long as the monitors do not report superior evidence; and the "hard scheme" gives the agent a large reward only when neither monitor reports evidence. The fourth scheme, which I call the "cheeky" scheme, rewards the agent only when the monitors report the inferior evidence, but not the superior evidence. Most of the paper focuses on the easy, moderate and hard schemes both because they are more robust to bribes than the cheeky scheme, and because the cheeky scheme is only optimal for relatively extreme distributions of evidence.

Each monitor can conceal evidence by pretending that they didn't obtain it, when in fact they did. This possibility poses little threat to the principal in the absence of bribes, because the monitors are indifferent about reporting evidence (they do not face any costs of acquiring or reporting the evidence). But when the agent can bribe the monitors to conceal evidence, the principal must respond by designing rewards that induce the monitors to truthfully report their evidence. I assess the role of a second independent monitor by comparing three different scenarios that hold the total amount of available information fixed. In scenario 1, the benchmark case, one monitor accesses both pieces of evidence, and the agent can bribe them to conceal one or both of them. In scenario 2, each monitor accesses a different piece of evidence and the agent can bribe either one, but not both, monitors; In scenario 3, each monitor accesses only one type of evidence and the agent can bribe either one or both monitors.

In scenario 1, the agent will be willing to bribe the monitor in any outcome where the monitor can increase the agent's reward by concealing evidence. The principal can only deter bribes by paying the agent's reward to the monitor whenever the monitor reports evidence that forfeits the agent receiving it. Doing so ensures that the agent and the monitor receive the same joint surplus in all outcomes, so they can never increase their joint surplus by concealing evidence. Thus the principal can implement the easy scheme by rewarding the monitor whenever they report both pieces of evidence; she can implement the moderate scheme by rewarding the monitor whenever they report the superior evidence; and she can implement the hard scheme by rewarding the monitor whenever they report any type of evidence.

Which of the three schemes is the cheapest depends on the distribution of evidence. Two incriminating signals are incriminating, and two exonerating signals are exonerating; so all three schemes agree that the agent should be rewarded when both signals are exonerating, but not when both signals are incriminating. But what if one signal is exonerating and the other is incriminating? If the distribution of evidence is such that a single piece of exonerating evidence is enough to counteract a single piece of incriminating evidence then the principal should reward the agent whenever at least one signal is exonerating, so the easy scheme is the cheapest. This tends to be true when monitoring errors are more positively correlated when the agent takes the right action, than when he takes the wrong action. But if a single piece of incriminating evidence is enough to counteract a single piece of exonerating evidence then the principal should only reward the agent when both signals are exonerating, so the hard scheme is the cheapest. This tends to be true when monitoring errors are more positively correlated when the agent takes the wrong action. The remaining possibility is that the superior incriminating evidence is strong enough to counteract inferior exonerating evidence, but not vice versa. In this case, the moderate scheme is the cheapest. If monitoring errors are always perfectly correlated (i.e. if both monitors observe the same signal), then all three schemes cost the same.

In scenario 2, each monitor accesses a different type of evidence, and the agent can bribe one or other, but not both, of them to conceal it. The principal can implement the moderate reward scheme in the same way as for one monitor, because it only relies on the monitor with access to the superior type of evidence. The same is not true for the easy scheme. When both monitors have evidence, the agent can secure the reward by bribing either one of them. The principal can only deter such bribes by paying a reward to both of the monitors if and only if they both report evidence. This is more costly than the one monitor case, where the principal only had to pay one reward. On the other hand, the hard scheme is cheaper to implement in this scenario than in scenario 1. In the one monitor scenario, the principal needs to reward the monitor when they receive both pieces of evidence, because the monitor has the ability to conceal them both, thereby allowing the agent to claim the reward. But if the agent cannot coordinate with both monitors, then there is no need to pay either of them a reward when they both report evidence, because the agent cannot gain by bribing only one of them.

Whether the principal is better off in scenario 1 or scenario 2 depends on the distribution of evidence. If the distribution is one favouring the easy scheme (neither type of evidence is incriminating by itself), then the principal is better off with one monitor, because she can implement the easy scheme more cheaply than with two monitors. However, if the distribution is one favouring the hard scheme (both types of evidence are incriminating by themselves), then the principal is better off with two monitors, so long as the agent cannot bribe them both. In other cases, including all the cases where both monitors receive the same evidence, the principal is indifferent.

In scenario 3, each monitor accesses a different type of evidence, and the agent can bribe either one or both of them. The same easy and moderate schemes that deter bribes in scenario 2 continue to deter bribes here, because the agent only needs to conceal one piece of evidence to obtain a reward in those schemes. But hard reward scheme that deters bribes in scenario 2, no longer deters bribes when both monitors obtain evidence — the agent would like to conceal both pieces of evidence, and can easily bribe the monitors to do so since neither of them is being rewarded to report it. In order to make the hard scheme robust to coalitions containing both monitors, the principal must reward one or other (or a mix of both) monitors when they both report evidence. But then the total rewards paid to both monitors are exactly equal to the size of the reward paid in the one monitor scenario. This implies that the principal is always weakly better off with a single monitor who accesses all the evidence, and strictly so when the distribution of evidence favours the easy reward scheme (i.e. when neither type of evidence is incriminating by itself). When both monitors access the same evidence, the easy and moderate schemes are weakly optimal, so once again, the principal is indifferent between the one and two monitors, even though the agent can bribe both of them.

I conclude that whether or not more monitors help to deter bribes depends on the ease with which the agent can bribe larger coalitions of monitors. If the agent cannot bribe both monitors simultaneously, and both types of evidence are sufficiently incriminating, then the principal is indeed better off with two, strategically independent monitors. If the agent can bribe both monitors then they are strategically equivalent to a single monitor. In this case, the principal is weakly better off with a single, omniscient monitor. In practice, the environment is likely to be somewhere between these two extremes because of frictions created by exogenous sources of private information, such as the precise characteristics of the evidence, the legal environment, or the players' 'moral' cost of bribery. It seems likely that larger coalitions suffer more from these frictions, in which case it will be harder for the agent to bribe both monitors than one. Moreover, the principal may be able to endogenously create private information to undermine collusive coalitions (see chapter 2). In these cases, it will generally be preferable to have different monitors receive different pieces of evidence.

The problem of relying on a corruptible monitor to provide incentives for an agent taking a hidden action dates back to the seminal contribution of Tirole (1986), and has been followed by an extensive literature (e.g. Felli and Hortala-Vallve, 2016; Strausz, 1997; Vafaï, 2005). An extensive literature studies whether competition between monitors can incentive costly monitoring effort (e.g. Liu, Wang, and Yin, 2022; McAfee, Mialon, and Mialon, 2008; Polinsky, 1980). Rahman (2012) tackles the same problem, but by having monitors monitor one another's effort, rather than by relying directly on market mechanisms. However, there are few results about the role that a multiplicity of monitors can play in deterring corruption. Kofman and Lawarrée (1993) introduce a second costly-but-incorruptible 'external' monitor. The optimal contract in their setting entails employing the internal monitor to monitor the agent and employing the external monitor to monitor the internal monitor. In all of these models, the principal has some means of obtaining her own unbiased information about the agent, either by observing output or by recourse to costly monitoring or by employing an external monitor. A key difference of my model is that the principal must rely exclusively on corruptible monitors to obtain information.

My problem is also closely related to the broader literature on mechanism design. Ben-Porath and Lipman (2012); Green and Laffont (1986); Koessler and Perez-Richet (2019), and others, study the role of hard evidence. Che and Kim (2006); Crémer (1996); Green and Laffont (1979); Laffont and Martimort (1997, 2000), and others, study the role of collusion on the set of social choice functions can be implemented. My principal also faces an implementation problem with hard evidence and collusion: she needs to implement a particular reward function for the agent (to induce him to take the right action), and she is able to choose transfers for the monitors in order to induce them to report their information. Hard evidence is strictly necessary for the principal because the monitors have no intrinsic preference over the agent's rewards, but they can always be bribed to report in the agent's favour.

Another related literature studies mutual monitoring in repeated games. Ben-Porath and Kahneman (1996) find that all individually rational payoff can be attained with equilibrium strategies if and only if each player's action is observed by at least two other players. Aoyagi (2005) allow players to collude by means of a communication device, and find that individually rational payoffs can be attained with equilibrium strategies if and only if each player's action is almost perfectly observed by all other players collectively. I get a similar result in a static setting: when players can collude, the *number* of monitors observing the agent's action does not matter, only the *quality* of the information they collectively provide.

The outline of the chapter is as follows: Section 2 defines the players, their strategies and payoffs, and each of the three scenarios. Section 3 shows that the principal can restrict attention to collusion proof schemes. Section 4 solves the principal's problem in two stages. The first stage takes the agent's reward scheme as given and solves for the cheapest monitor wage scheme that deters bribes. The second stage then uses the costs implied by the monitor's wages to solve for the cheapest agent-reward scheme. Section 5 obtains the main result by comparing the optimal schemes across the three scenarios. Section 6 discusses the robustness of the baseline model by considering different assumptions about limited liability, collusion, and the distribution of evidence. Section 7 discusses the results in the context of applications to journalism, whistle-blowing and financial auditing. Section 8 concludes.

2 Model

A risk neutral agent (he) chooses between a "right" action and a "wrong" action. Other things equal, the right action yields a payoff of 0, whereas the wrong action yields a payoff of 1. There are two binary signals s_1 and s_2 which are informative about the agent's action. Each signal takes the values 1 ("incriminating evidence") or 0 ("no evidence"), so the joint signal $s = (s_1, s_2)$ is an element of the set $S := \{0,1\}^2$ (where ":=" denotes the definition of a symbol). The distribution of s depends on the agent's choice of action. If the agent takes the wrong action, then s occurs with probability τ_s; if the agent takes the right action, then s occurs with probability π_s. If $\Delta_s := \pi_s - \tau_s > 0$, then s is more likely to occur when the agent takes the right action than when he takes the wrong action, so we can think of s as *exonerating evidence*. Similarly, if $\Delta_s < 0$, then s constitutes *incriminating evidence*, and if $\Delta_s = 0$, then it is neutral. Note that $\sum_{s\in S} \pi_s = \sum_{s\in S} \tau_s = 1$ implies $\sum_{s\in S} \Delta_s = 0$, so a realisation s can only be incriminating if some other realisation, s', is exonerating. The main results assume that $\Delta_{11} < 0$ and $\Delta_{00} > 0$, so two pieces of incrimination evidence are incriminating, and no evidence is exonerating. Moreover, I assume that $\Delta_{11} \leq \Delta_{10} \leq \Delta_{00}$ and $\Delta_{11} \leq \Delta_{01} \leq \Delta_{00}$, so a single piece of evidence is more incriminating than no evidence, but less incriminating than both pieces of evidence. These assumptions imply that $\Delta_{00} + \Delta_{01} > \Delta_{10} + \Delta_{11}$ and $\Delta_{00} + \Delta_{10} > \Delta_{01} + \Delta_{11}$ which means that $s_i = 1$ is incriminating when s_{-i} is not observed. It is without loss of generality to further assume that $\Delta_{01} \geq \Delta_{10}$. Consequently, $\Delta_{01} + \Delta_{11} \geq \Delta_{10} + \Delta_{11}$ and $\Delta_{00} + \Delta_{01} \geq \Delta_{00} + \Delta_{10}$, which means that $s_1 = 1$ is more incriminating than $s_2 = 1$ and $s_1 = 0$ is more exonerating than $s_2 = 0$. Thus the (superior) signal s_1 is more accurate than the (inferior) signal s_2. If $\Delta_{00} = 1$ (resp. $\Delta_{11} = -1$) then we arrive in the case of perfect monitoring because it must be that $\pi_{00} = 1$ and $\tau_{00} = 0$ (resp. $\pi_{00} = 0$ and $\tau_{00} = 1$). In this special case, the principal can perfectly infer that the agent must have taken the right action if and only if $s = (0,0)$ (resp. $s \neq (1,1)$). Other cases are considered in section 6.

The principal wants to incentivise the agent to take the right action. She can only do so by committing to reward or punish him on the basis of the signal s. Specifically, the principal promises to pay the agent a transfer $T(s)$ when signal s is reported. This transfer function $T : S \to \mathbb{R}$ needs to satisfy

$$\sum_{s \in S} \pi_s T(s) \geq 1 + \sum_{s \in S} \tau_s T(s)$$

or equivalently,

$$\sum_{s \in S} \Delta_s T(s) \geq 1, \tag{IC}$$

to ensure that the agent's payoff from taking the right action is greater than his payoff from taking the wrong action. This is the agent's *incentive compatibility* or (IC) constraint. I assume that indifferences are broken in favour of the principal, so the agent complies if and only if (IC) is satisfied. It is clear that this constraint must be satisfied by either rewarding the agent when exonerating evidence is realised, or by punishing the agent when incriminating evidence is realised, because these configurations will ensure that the product $\Delta_s T(s)$ is positive. For most of the paper, I rule out punishments by assuming that the agent has limited liability, so the principal must rely on rewarding exonerating evidence. If (IC) is satisfied, then the agent chooses the right action in equilibrium so the cost of transfer function T is $\sum_{s \in S} \pi_s T(s)$.

If the principal directly observed both signals s_1 and s_2, then her objective would simply be to choose a transfer function to minimise the cost of satisfying the agent's (IC) constraint. However, the key feature of the model is that the signal is not observed by the principal, but by either one or two monitors. Moreover, the monitors are corruptible: they can be bribed to conceal evidence. If monitor $i \in \{1,2\}$ receives a 1 signal, then they may send either a 1 or a 0 report to the principal. But if they receive a 0 signal, then they have no choice but to send a 0 report to the principal. The interpretation is that 1 is hard incriminating evidence, and 0 is a dearth of hard incriminating evidence. Hard evidence can be concealed, but not fabricated. Define a component-wise ordering, '$\leq$', by $m \leq s$ if and only if $m_i \leq s_i$ for all $i = 1,2$, which is true only if m can be obtained from s by fully or partially concealing evidence. The 'lower contour set' of an outcome $s \in S$ is the set of outcomes that can be reached from it by concealing evidence, $\{m \in S \mid m_i \leq s_i, i = 1,2\}$.

The principal anticipates the possibility that the agent could bribe the monitor to conceal evidence. For instance, if the principal chooses $T(0,0) > T(1,0)$, then the agent would be willing to bribe the monitor anything up to $T(0,0) - T(1,0)$ to report $(0,0)$ instead of $(1,0)$. The principal can mitigate against bribery by rewarding the monitors for reporting hard evidence. It does so by committing to a wage function $w_i : S \to \mathbb{R}_+$ which specifies how much monitor i gets paid when they collectively report evidence s. The main result assumes that both monitors have limited liability, so wages must be weakly positive (this assumption is relaxed in section 6). A *scheme* (T, w_1, w_2) specifies a transfer function for the agent together with a pair of wage functions for the monitors. The cost of a scheme (T, w_1, w_2) is $c(T, w_1, w_2) := \sum_{s \in S} \pi(s)(T(s) + w_1(s) + w_2(s))$. The principal's problem is to choose a scheme of transfers and wages to minimise the expected cost, subject to the agent's incentive compatibility constraint, the agent's liability constraint, the monitors' liability constraints.

The agent has an incentive to bribe the monitors to conceal evidence whenever doing so would lead to him being rewarded. I assume that bribery negotiations take place after the monitor obtains evidence (if the monitor does not obtain evidence then there is no scope for bribes because the monitor cannot change the monitor's reward), and that the evidence realisation s is common knowledge. The latter assumption ensures that any scheme that is robust to bribes in this environment, will continue to be robust to bribes in environments where the players negotiate bribes without perfect knowledge of the evidence realisation. It does not necessarily conflict with the principal's need to rely on monitors to report the evidence, because evidence needs to be verifiable in most practical contexts. Here, I restrict attention to deterministic bribes — all the results continue to hold when stochastic bribes are permitted.

The precise form of a bribe depends on which monitor(s) have evidence and which collusive coalitions can form. I compare three scenarios:

- **Scenario 1.** Monitor 1 receives both signals s_1 and s_2, and the agent can bribe the monitor to conceal one or both of them. In outcome s, a bribe $(b(s), m(s))$ specifies a report $m(s) = (m_1(s), m_2(s))$ and with $m \leq s$ for $i = 1, 2$, and a bribe b to be paid to monitor 1. The agent is willing to pay the (non-trivial) bribe if

$$T(m(s)) - b(s) > T(s), \tag{1}$$

  and monitor 1 is willing to accept the (non-trivial) bribe if

$$w_1(m(s)) + b(s) > w_1(s). \tag{2}$$

- **Scenario 2.** Monitor i receives signal s_i, and the agent can either bribe one, but not both of them. In outcome s, a bribe $(b_i(s), m_i(s))$ specifies a report $m_i(s) \leq s_i$, and a bribe b_i to be paid to monitor $i = 1, 2$. The agent is willing to pay the bribe if

$$T(m_i(s), s_{-i}) - b_i(s) > T(s), \tag{3}$$

  and monitor i is willing to accept the bribe if

$$w_i(m_i(s), s_{-i}) + b_i(s) > w_i(s). \tag{4}$$

- **Scenario 3.** Monitor i receives signal s_i, and the agent can bribe either one or both of them. In outcome s, a bribe $(b_1(s), b_2(s), m_1(s), m_2(s))$ specifies reports $m_i(s) \leq s_i$, and bribes b_i, to be paid to one or both monitors $i = 1, 2$. The agent is willing to bribe one monitor if (3) holds, or both monitors if

$$T(m_1(s), m_2(s)) - \sum_{i=1,2} b_i(s) > T(s) \tag{5}$$

  holds; and monitor $i = 1, 2$ is willing to accept the bribe if

$$w_i(m_1(s), m_2(s)) + b_i(s) > w_i(s). \tag{6}$$

The inequalities are strict because of my assumption that indifferences are broken in favour of the principal. This assumption simplifies the problem because it will ensure that the principal's feasible set is compact; but has no substantial bearing on the solution because the principal can give the players a strict preference to reject bribes by adding some arbitrarily small amount to the monitors' wages. A bribe is *feasible* in outcome s if the monitor is willing to accept it and the agent is willing to pay it. I discuss the implications of allowing other coalitions (excluding the agent and/or including the principal) in section 6. *How* the players choose between different feasible bribes does not matter, because lemma 2 will show that the principal can restrict attention to schemes for which there are no feasible, non-trivial bribes.

How do bribes affect the principal's problem? To incentivise the agent to take the right action, the principal must use a 'robustly incentive compatible' scheme. A scheme is *robustly incentive compatible* if it satisfies (IC), no matter what feasible bribes take place. For example, in scenario 3, a scheme (T, w_1, w_2) must satisfy

$$\sum_{s \in S} \Delta_s (T(m_1(s), m_2(s)) - b_1(s) - b_2(s)) \geq 1, \tag{RIC}$$

for every feasible bribe $(b(s), m(s))$, and every outcome $s \in S$. Bribes also affect the cost of the scheme. Erring on the side of caution, I assume that the cost of a scheme is equal to the maximum amount that the principal might have to pay out in rewards following any feasible bribes. Indeed, this is how much the principal pays if the agent and the monitor conceal evidence to maximise their joint surplus. For example, in scenario 1, the cost of a scheme (T, w_1, w_2) is

$$\max_{\substack{(b(s),m(s))\\ \text{is feasible}}} \sum_{s\in S} \pi_s(T(m(s)) + w_1(m(s)) + w_2(m(s)).$$

The cost is defined analogously in other scenarios, with the appropriate bribe as a choice variable.

I conclude this section by summarising the timing of the model:

i) the principal proposes a scheme;

ii) the agent chooses either the right or the wrong action;

iii) the monitor(s) obtain evidence s;

iv) the players negotiate a bribe;

v) the monitor(s) report evidence;

vi) all players' payoffs are realised.

3 Bribe Proof Schemes

Anticipating all the possible bribes that may take place is unnecessary, because the principal can simplify her problem by restricting attention to "bribe proof" schemes in which all bribes are deterred. A scheme (T, w_1, w_2) is bribe proof if and only if there are no feasible bribes in any outcome s. Precisely what this means will depend on which scenario is under consideration, because the scenario determines which bribes are feasible. Specifically, a scheme is *j-bribe proof* if it is bribe proof in scenario $j = 1, 2, 3$. Lemma 1 shows that a scheme is j-bribe proof if and only if no coalition can generate surplus by suppressing evidence.

Lemma 1. *A scheme (T, w_1, w_2) is*

- *1-bribe proof if and only if*

$$T(s) + w_1(s) \geq T(m) + w_1(m) \tag{1-BP}$$

for all $m \leq s$;

- *2-bribe proof if and only if*

$$T(s) + w_i(s) \geq T(m_i, s_{-i}) + w_i(m_i, s_{-i}) \tag{2-BP}$$

for all $m_i \leq s_i$, and for all $i = 1, 2$;

- *3-bribe proof if and only if it is 2-bribe and*

$$T(s) + w_1(s) + w_2(s) \geq T(m) + w_1(m) + w_2(m) \tag{3-BP}$$

for all $m \leq s$.

Proof. I first prove the 'if' statement for each scenario, and then the 'only if' statement.

Let (T, w_1, w_2) be a scheme, and let $(b(s), m(s))$ be a feasible bribe in some outcome s in scenario 1, with $m \neq s$. Then $(b(s), m(s))$ must strictly satisfy inequalities (1) and (2). Summing these gives $w_1(m(s)) + T(m(s)) > w_1(s) + T(s)$, which is the opposite of (1-BP). Therefore if (1-BP) holds then the scheme is 1-bribe proof, proving the "if" statement for scenario 1. For scenario 2, let $(b_i(s), m_i(s))$ be a feasible bribe with $m_i \neq s_i$. Then $(b_i(s), m_i(s))$ must satisfy inequalities (3) and (4), which sum to give $T(m_i(s), s_i) + w_i(m_i(s), s_{-i}) > T(s) + w_i(s)$. This is ruled out by (2-BP), proving the "if" statement for scenario 2. Finally, let $(b_1(s), b_2(s), m_1(s), m_2(s))$ be a feasible bribe in scenario 3, with either $m_1 \neq s_1$ or $m_2 \neq s_2$.

Then $(b_1(s), b_2(s), m_1(s), m_2(s))$ either satisfies (3) and (4) for some monitor i, in which case it is ruled out by (2-BP); or else it satisfies inequalities (5) and (4) for monitors $i = 1,2$. The latter inequalities sum to give $T(m_1(s), m_2(s)) + \sum_{i=1,2} w_i(m_1(s), m_2(s)) \geq T(s) + \sum_{i=1,2} w_i(s)$, which is ruled out by (3-BP). This proves the "if" statement for scenario 3.

Now suppose that one of the constraints is violated. Then some coalition of players can generate surplus by concealing evidence. Any bribe that splits this surplus between the members of the coalition will be feasible. For example, if (3-BP) is violated in a state s with report m, then there exists some $\epsilon > 0$ such that

$$T(m_1(s), m_2(s)) - \sum_{i=1,2} (w_i(m_1(s), m_2(s)) - w_i(s)) > T(s) + 2\epsilon$$

and $w_i(m_1(s), m_2(s)) + w_i(s) - w_i(m_1(s), m_2(s)) + \epsilon > w_i(s)$, so the bribes $b_i(s) = w_i(s) - w_i(m_1(s), m_2(s)) + \epsilon$ for monitors $i = 1,2$ are feasible. Therefore (T, w_1, w_2) is not 3-bribe proof, proving the "only if" part of the statement. The proofs for scenarios 1 and 2 are analogous. □

The scenario 2 constraints closely resemble the scenario 1 constraints, except with the addition of the i subscripts, and the lack of any constraint on transfers and wages in outcome (1,1) relative to outcome (0,0). This is because the agent cannot bribe both monitors, and so the principal does not have to worry about both pieces of evidence being concealed at the same time. In scenario 3, all but one of these constraints involves the agent bribing both monitors, even though only one monitor suppresses their evidence. It is necessary to account for such coalitions to rule out schemes that increase one monitor's wage when the other monitor reports a 0. In such schemes, the agent benefits from including the second monitor in the coalition because doing so increases the overall surplus that the coalition generates by concealing just one signal.

Lemma 2 shows that, in scenarios 1 and 3, the principal can restrict attention to schemes for which there are no (non-trivial) feasible bribes.

Lemma 2. *If a scheme (T, w_1, w_2) respects robust incentive compatibility and limited liability, then there exist 1- and 3-bribe proof schemes (T^1, w^1) and (T^3, w^3) that respect incentive compatibility and limited liability, and have weakly lower cost, $c(T^j, w^j) \leq c(T, w)$ for $j = 1,3$.*

Proof. The proof of lemma 2 follows a revelation principle type argument (Laffont and Martimort, 1997). I give the proof for scenario 3; the proof for scenario 1 is analogous. Suppose a scheme (T, w_1, w_2) satisfies (RIC), limited liability, and costs C, but violates 3-bribe proofness. Lemma 1 shows that there exists a feasible bribe for each state s where (3-BP) is violated. The set of feasible bribes is not compact (because of the strict inequalities in equations (1)–(6)), and therefore there may not exist a feasible bribe that maximises the agent's payoff. But the agent's payoff is bound above by $\max_s T(s) + w_1(s) + w_2(s)$, so it has a supremum, and this supremum is attained by a bribe $(b_1(s), b_2(s), m_1(s), m_2(s))$ in the closure of the feasible set.

Define a new scheme by $T^3(s) = T(m_1(s), m_2(s)) - b_1(s) - b_2(s)$ and $w_i^3(s) = w_i(m_1(s), m_2(s)) + b_i(s)$. Then (T^3, w_1^3, w_2^3) has the same ex post payoffs as the old scheme (T, w_1, w_2) together with the bribe. It satisfies limited liability because (T, w_1, w_2) satisfies limited liability and $(b_1(s), b_2(s), m_1(s), m_2(s))$ is in the closure of the set of feasible bribes, so $T(m_1(s), m_2(s)) - b_1(s) - b_2(s) \geq T(s) \geq 0$ and $w_i(m_1(s), m_2(s)) + b_i(s) \geq w_i \geq 0$. Similarly, (T^3, w_1^3, w_2^3) satisfies incentive compatibility because (T, w_1, w_2) satisfies (RIC) so $\sum_{s \in S} \pi_s T^3(s) = \sum_{s \in S} \pi_s T(s)(T(m_1(s), m_2(s)) - b_1(s) - b_2(s)) \geq 1$. Moreover, (T^3, w_1^3, w_2^3) must be bribe proof. If it were not, then there would be a bribe that strictly increases the agent's payoff $(b_1'(s), b_2'(s), m_1'(s), m_2'(s))$. But then the composition of the two bribes, $(b_1(s) + b_1'(s), b_2(s) + b_2'(s), m_1'(m_1(s)), m_2'(m_2(s)))$, is feasible under (T, w_1, w_2) and yields the agent a higher payoff than $(b_1(s), b_2(s), m_1(s), m_2(s))$, contradicting the fact that $(b_1(s), b_2(s), m_1(s), m_2(s))$ was assumed to attain the supremum of the agent's payoffs across all feasible bribes. Finally, the scheme (T^3, w_1^3, w_2^3) has a weakly lower cost because

$$\begin{aligned} c(T, w_1, w_2) &= \max_{\substack{(b_1(s), b_2(s), m_1(s), m_2(s)) \\ \text{is feasible}}} \sum_{s \in S} \pi_s (T(m'(s)) + w_1(m'(s)) + w_2(m'(s))) \\ &\geq \sum_{s \in S} \pi_s (T(m(s)) + w_1(m(s)) + w_2(m(s)) \\ &= \sum_{s \in S} \pi_s (T^3(s) + w_1^3(s) + w_2^3(s) \\ &= c(T^3, w_1^3, w_2^3), \end{aligned}$$

where the second equality holds because the bribes cancel out. □

The proof of lemma 2 does not work for scenario 2, because the composition of two bribes is not feasible if they each involve different monitors. But the purpose of studying scenario 2 is to show that the principal can potentially save money by hiring two monitors if they cannot be bribed simultaneously. If lemma 2 does not hold for scenario 2, then it only means that the cheapest feasible, bribe proof solution gives an upper bound on the cheapest feasible solution.

Lemmas 1 and 2 imply that the principal's problem in each scenario can be formulated as follows:

$$\begin{aligned} \max_{T, w_1, w2} \; & -\sum_{s \in S} \pi_s (T(s) + w_1(s) + w_2(s)) & & \\ \text{st.} \; & \sum_{s \in S} \Delta_s T(s) \geq 1 & & \text{(IC)} \\ & T(s) \geq 0 & \forall s \in S \quad & \text{(LLA)} \\ & w_i(s) \geq 0 & \forall s \in S, i = 1, 2 \quad & \text{(LLM)} \\ & (j - \text{BP}) & & \end{aligned}$$

where $j = 1$ (in scenario 1), $j = 2$ (in scenario 2) or $j = 2, 3$ (in scenario 3).

4 Optimal schemes

I solve for the optimal schemes in two steps. First I take the transfer function T as given, and solve for the cheapest wages w_1 and w_2 that deter bribes in each scenario. Then, given the cost of the transfer T and the corresponding wages w_1 and w_2, I find the cheapest scheme for given parameters π and τ.

4.1 Incentivising the monitors

Proposition 1 says that the monitors are collectively paid the marginal value of their evidence in any optimal scheme. The only differences between the three scenarios occur in the outcome $(1, 1)$, because it is the only outcome where the marginal value of a piece of evidence depends on whether the other evidence is reported.

Proposition 1. *If (T^j, w_1^j, w_2^j) is a solution to the principal's problem in scenario j, then*

$$w_1^j(0,0) + w_2^j(0,0) = W^j(0,0; T^j) := 0 \tag{7}$$

$$w_1^j(1,0) + w_2^j(1,0) = W^j(1,0; T^j) := \max\{0, T^j(0,0) - T^j(1,0)\} \tag{8}$$

$$w_1^j(0,1) + w_2^j(0,1) = W^j(0,1; T^j) := \max\{0, T^j(0,0) - T^j(0,1)\} \tag{9}$$

for $j = 1,2,3$, and

$$w_1^1(1,1) + w_2^1(1,1) = W^j(1,1;T^j) := \max\{0, T^j(0,1) - T^j(1,1), T^j(1,0) - T^j(1,1), T^j(0,0) - T^j(1,1)\} \tag{10}$$

$$w_1^2(1,1) + w_2^2(1,1) = W^j(1,1;T^j) := \max\{0, T^j(0,1) - T^j(1,1), T^j(1,0) - T^j(1,1), T^j(0,1) + T^j(1,0) - 2T^j(1,1)\} \tag{11}$$

$$w_1^3(1,1) + w_2^3(1,1) = W^j(1,1;T^j) := \max\{w_1^1(1,1) + w_2^1(1,1), w_1^2(1,1) + w_2^2(1,1)\}. \tag{12}$$

Proof. Equality (7) says that neither monitor gets paid if neither has evidence. The wages $w_1(0,0)$ and $w_2(0,0)$ feature negatively in the principal's objective and no coalition can ever benefit from monitor i concealing evidence in these outcomes. So the corresponding wages only feature positively in monitor i's limited liability constraints, which means that these constraints must hold with equality in any optimal solution, giving (7).

Similarly, $w_1(1,0)$ and $w_2(1,0)$ have negative coefficient in the principal's objective and in some bribe constraints. The only places where they have positive coefficients are in the liability constraints, and in the bribe constraints for the outcome $(1,0)$. Substituting (7) into (1-BP) for outcome $(1,0)$ gives $w_1(1,0) \geq T(0,0) - T(1,0)$. The limited liability constraints give $w_1(1,0) \geq 0$ and $w_2(1,0) \geq 0$, so any optimal solution must have $w_2(1,0) = 0$ and $w_1(1,0) = \max\{0, T(0,0) - T(1,0)\}$, hence equation (8) holds in scenario 1. The same argument applies for scenario 2. Scenario 3, has an additional bribe constraint, (3-BP), which says that $w_1(1,0) + w_2(1,0) \geq T(0,0) - T(1,0)$, but this is already satisfied by (8). The same argument applies for outcome $(0,1)$.

The optimal wages differ between scenarios when both monitors have evidence. In scenario 1, $w_1(1,1)$ enters three bribe constraints, one for each of the possibility deviations to $(1,0), (0,1)$ and $(0,0)$. Substituting $w_1(1,0) = \max\{0, T(0,0) - T(1,0)\}$, $w_1(0,1) = \max\{0, T(0,0) - T(0,1)\}$ and $w_1(0,0) = 0$ into these constraints gives

$$\begin{aligned} w_1(1,1) &\geq T(1,0) - T(1,1) + \max\{0, T(0,0) - T(1,0)\} \\ w_1(1,1) &\geq T(0,1) - T(1,1) + \max\{0, T(0,0) - T(0,1)\} \\ w_1(1,1) &\geq T(0,0) - T(1,1). \end{aligned}$$

The upper envelope of these constraints gives $w_1(1,1) \geq \max\{0, T(0,0) - T(1,1), T(1,0) - T(1,1), T(0,1) - T(1,1)\}$. The wage $w_2(1,1)$ does not enter any bribe constraints but limited liability requires $w_2(1,1) \geq 0$. Both feature negatively in the principal's objective, so both must hold with equality. Summing gives (10).

In scenario 2, $w_1(1,1)$ only enters one bribe constraint, which says that $w_1(1,1) \geq T(0,1) - T(1,1)$, as well as monitor 1's liability constraint. Wage $w_1(1,1)$ has a negative coefficient in the objective, so one of these two constraints must hold with equality in any optimal solution, so $w_1^2(1,1) = \max\{0, T(0,1) - T(1,1)\}$. The same argument applies for $w_2^2(1,1)$. Summing gives (11).

In scenario 3, $w_1(1,1)$ and $w_2(1,1)$ enter five bribe constraints. Substituting (8) and (9) into them yields

$$\begin{aligned} w_1(1,1) &\geq T(0,1) - T(1,1) \\ w_2(1,1) &\geq T(0,1) - T(1,1) \\ w_1(1,1) + w_2(1,1) &\geq T(0,1) - T(1,1) + \max\{0, T(0,0) - T(0,1)\} \\ w_1(1,1) + w_2(1,1) &\geq T(1,0) - T(1,1) + \max\{0, T(0,0) - T(1,0)\} \\ w_1(1,1) + w_2(1,1) &\geq T(0,0) - T(1,1), \end{aligned}$$

as well as their liability constraints, which together imply that $w_1(1,1) + w_2(1,1) \geq 0$. The first two bribe constraints sum to give $w_1(1,1) + w_2(1,1) \geq T(0,1) + T(0,1) - 2T(1,1)$. The objective is decreasing in $w_1(1,1) + w_2(1,1)$, so at least one of these five constraints must hold with equality. If it is not the case that $w_1^3(1,1) + w_2^3(1,1) = T(0,0) - T(1,1)$, then $w_1^3(1,1) + w_2^3(1,1) = w_1^2(1,1) + w_2^2(1,1) > w_1^1(1,1) + w_2^1(1,1)$, so (12) holds. Otherwise, if $w_1^3(1,1) + w_2^3(1,1) = T(0,0) - T(1,1)$, then $w_1^3(1,1) + w_2^3(1,1) = w_1^1(1,1) + w_2^1(1,1) > w_1^2(1,1) + w_2^2(1,1)$, and (12) still holds. □

We can already see from equation (12) that the value of a second monitor is limited, because the principal is more constrained in scenario 3 than in scenario 1. Proposition 2 will show how this difference in the wages will affect the optimal choice of transfer.

The function $W^j(s;T)$ defined by proposition 1 allows us to remove the bribe constraints and the monitors' liability constraints from the principal's problem. Substituting it gives

$$\begin{aligned} &\max_T -\sum_{s\in S}\pi_s(T(s)+W^j(s;T)) \\ &\text{st.} \sum_{s\in S}\Delta_s T(s)\geq 1 && \text{(IC)} \\ &T(s)\geq 0 \qquad \forall s\in S. && \text{(LLA)} \end{aligned}$$

4.2 Incentivising the agent

The principal incentivises the agent by rewarding him for exonerating outcomes, i.e. outcomes s with $\Delta_s > 0$. She trades this off against the cost of rewarding to the agent, which is determined by the probability of realising the outcome s when the agent takes the right action; and the cost of deterring the monitors from falsely reporting the outcome s, which is determined by the probability of realising an outcome s' from which the agent can bribe the monitors to report s.

Lemma 3. *The optimal transfers for the agent are bang-bang. If (T^*, w_1^*, w_2^*) is a solution to the principal's problem, then there exists a set of 'reward' outcomes $R\subset S$ such that the agent is given a reward of size $\frac{1}{\sum_{s\in R}\Delta_s}$ if the monitors report an outcome in R. In all other outcomes, he receives no reward. I.e. $T^*(s)=\mathcal{I}(s\in R)/\sum_{s\in R}\Delta_s$.*

Proof. I first show that the solution to the principal's problem coincides with a solution to a linear programme, and then show that the solution to this linear programme must be bang-bang.

Any solution (T^*, w_1^*, w_2^*) to any of the three problems implies an ordering over rewards, e.g. $T^*(0,1)\geq T^*(1,0)\geq T^*(0,0)\geq T^*(1,1)$. This means that (T^*, w_1^*, w_2^*) also solves the problem where these three inequalities are imposed as constraints. This *order-constrained* problem has four choice variables, $T(0,0), T(1,0), T(0,1)$ and $T(1,1)$, and five linear constraints (the (IC) constraint, the three order constraints, and one active liability constraint). The role of the active liability constraint ($T(1,1)\geq 0$ in the example) is the same as the role of the order constraints, so I will refer to the order constraints and the liability constraint collectively as 'order constraints' for the remainder of the proof. The ordering determines which arms of the maximands in (8)–(12) are active, so the objective of the order constrained problem is linear. For example, if $T(0,1)\geq T(1,0)\geq T(0,0)\geq T(1,1)$, then proposition 1 implies that scenario 1 has

$$\begin{aligned} w_1^1(0,0)+w_2^1(0,0)&=0 \\ w_1^1(1,0)+w_2^1(1,0)&=0 \\ w_1^1(0,1)+w_2^1(0,1)&=0 \\ w_1^1(1,1)+w_2^1(1,1)&=T(1,0)-T(1,1), \end{aligned}$$

so the scenario 1 problem reduces to the linear programme

$$\begin{aligned} &\min_T \pi_{00}T(0,0)+(\pi_{10}+\pi_{11})T(1,0)+\pi_{01}T(0,1) \\ &\text{s.t.} \sum_{s\in S}\Delta_s T(s)\geq 1, && \text{(IC)} \\ &T(0,1)\geq T(1,0) \\ &T(1,0)\geq T(0,0) \\ &T(0,0)\geq T(1,1) \\ &T(1,1)\geq 0. && \text{(LLA)} \end{aligned}$$

We can solve this problem by studying the Khun-Tucker conditions. The first order conditions yield a system of four equations (one for each choice variable) in five Lagrange multipliers (one for each constraint). For example, if λ_{IC} denotes the multiplier on (IC) and λ_s denotes the multiplier on the order constraint for $T(s)$, then we get

$$\begin{aligned}
\lambda_{IC}\Delta_{11} + \lambda_{11} - \lambda_{00} &= 0 \\
-\pi_{00} + \lambda_{IC}\Delta_{00} + \lambda_{00} - \lambda_{10} &= 0 \\
-\pi_{10} - \pi_{11} + \lambda_{IC}\Delta_{10} + \lambda_{10} - \lambda_{01} &= 0 \\
-\pi_{01} + \lambda_{IC}\Delta_{01} + \lambda_{01} &= 0
\end{aligned}$$

If two multipliers equal zero then we have a system of three equations in two unknowns, which only has a solution in certain, knife-edge cases.[3] Therefore, at most one of the multipliers can be generically equal to 0. On the other hand, if the four order constraint multipliers are all positive, then complementary slackness implies that these constraints all hold with equality, so the transfers would all equal 0, violating (IC). Therefore, at least one of the order multipliers λ_s must equal 0.

Together, these conclusions imply that exactly one of the order multipliers equals zero, so complementary slackness implies that exactly one of the corresponding inequalities is strict. This implies that each of the transfers is either (i) equal to zero, if it lies below the strict inequality; or it is (ii) equal to the largest transfer, if it lies above the strict inequality. The set R is equal to the set of outcomes with transfers above the strict inequality. In the example above, if $\lambda_{10} = 0$, then we get $T(0,1) = T(1,0) > T(0,0) = 0$, so $R = \{(0,1),(1,0)\}$.

Let T^* denote the size of the strictly positive transfer. The principal wants to minimise T^* subject to an IC constraint which is increasing in T^*. Therefore the IC constraint must hold with equality in any optimal solution, giving $\sum_{s\in R}\Delta_s T^* = 1$, or $T^* = 1/\sum_{s\in R}\Delta_s$. □

Proposition 2 formally describes the easy, moderate and hard schemes described in the introduction. To aid intuition, I present these an matrices where entry (i,j) in the wage matrix denotes the value of the corresponding function in outcome $(i-1, j-1)$. For example, the easy, moderate and hard reward functions are given by

$$\begin{aligned}
T_E &:= \begin{bmatrix} \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} & \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} \\ \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} & 0 \end{bmatrix} \\
T_M &:= \begin{bmatrix} \frac{1}{\Delta_{00}+\Delta_{01}} & \frac{1}{\Delta_{00}+\Delta_{01}} \\ 0 & 0 \end{bmatrix} \\
T_H &:= \begin{bmatrix} \frac{1}{\Delta_{00}} & 0 \\ 0 & 0 \end{bmatrix}.
\end{aligned}$$

There is also a fourth type of scheme that may be optimal, which I refer to as the *cheeky* scheme with transfer function

$$T_C := \begin{bmatrix} 0 & \frac{1}{\Delta_{01}} \\ 0 & 0 \end{bmatrix}.$$

In this scheme, the principal gains by reducing the probability of paying any rewards by only paying rewards when monitor 2 reports incriminating evidence, which could be very rarely, given that the agent takes the right action in equilibrium. In section 6 I argue that this scheme is less robust than the others because it is susceptible to corruption by the principal.

[3] These cases admit a continuum of solutions, at least two of them will have the binary structure described in the statement of lemma 3.

Proposition 2. *One of the following four schemes is optimal in scenario 1 (monitor 1 receives both signals s_1 and s_2):*

$$S_E^1 := \left(T_E, \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right)$$
$$S_M := \left(T_M, \begin{bmatrix} 0 & 0 \\ \frac{1}{\Delta_{00}+\Delta_{01}} & \frac{1}{\Delta_{00}+\Delta_{01}} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right)$$
$$S_H^1 := \left(T_H, \begin{bmatrix} 0 & \frac{1}{\Delta_{00}} \\ \frac{1}{\Delta_{00}} & \frac{1}{\Delta_{00}} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right)$$
$$S_C := \left(T_C, \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Delta_{01}} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\right).$$

The costs of these schemes are $c_E^1(\pi,\Delta) := \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}}$, $c_M(\pi,\Delta) := \frac{1}{\Delta_{00}+\Delta_{01}}$, $c_H^1(\pi,\Delta) := \frac{1}{\Delta_{00}}$, and $c_C(\pi,\Delta) := \frac{\pi_{01}+\pi_{11}}{\Delta_{01}}$, respectively. The cost of the optimal scheme is equal to the lower envelope of these costs, $\min\{c_E^1, c_M, c_H^1, c_C\}$.

One of the following four schemes is optimal in scenario 2 (monitor i receives signal s_i; the agent can bribe either one but not both monitors): S_M, S_C, or

$$S_E^2 := \left(\begin{bmatrix} \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} & \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} \\ \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} \end{bmatrix}\right)$$
$$S_H^2 := \left(\begin{bmatrix} \frac{1}{\Delta_{00}} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ \frac{1}{\Delta_{00}} & 0 \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{\Delta_{00}} \\ 0 & 0 \end{bmatrix}\right)$$

The costs of the schemes S_E^2 and S_H^2 are $c_E^2(\pi,\Delta) := \frac{1+\pi_{11}}{\Delta_{00}+\Delta_{01}+\Delta_{10}}$ and $c_H^2(\pi,\Delta) := \frac{1-\pi_{11}}{\Delta_{00}}$ respectively. The cost of the optimal scheme is equal to the lower envelope of the costs, $\min\{c_E^2, c_M, c_H^2, c_C\}$.

One of the following four schemes is optimal in scenario 3 (monitors i receives signal s_i; the agent can bribe one, other or both of them): S_E^2, S_M, S_H^1 or S_C. The cost of the optimal scheme is equal to the lower envelope of the costs, $\min\{c_E^2, c_M, c_H^1, c_C\}$.

Proof. Every optimal scheme has a transfer function defined by the set R in lemma 3. The set R can take $4! = 24$ possible values — one for every subset of S. For each subset, proposition 1 tells us how to calculate the corresponding wages, so we can easily calculate and compare the costs of the corresponding schemes.

There are four subsets of size 1. The sets $\{(0,0)\}$ and $\{(0,1)\}$ yield the hard and cheeky reward functions T_H and T_C respectively. The set $\{(1,0)\}$ is symmetric to T_C, but not as cheap because it rewards the superior evidence, which creates worse incentives than rewarding the inferior evidence (as T_C does). The set $\{(1,1)\}$ is not feasible because it the assumption that $\Delta_{11} < 0$ would yield a negative transfer $T(1,1) = 1/\Delta_{11} < 0$, violating limited liability.

There are six subsets of size 2. The set $\{(0,0),(0,1)\}$ yields the moderate reward function T_M. The set $\{(0,0),(1,0)\}$ is symmetric but more expensive because it rewards superior evidence (because $\Delta_{01} > \Delta_{10}$ implies $\frac{1}{\Delta_{00}+\Delta_{01}} < \frac{1}{\Delta_{00}+\Delta_{10}}$). The set $\{(1,1),(1,0)\}$ yields a transfer function that costs $\frac{\pi_{11}+\pi_{10}}{\Delta_{11}+\Delta_{10}}$, which is strictly more than the cost of the cheeky scheme because $\Delta_{11} < 0$. So it cannot be optimal. A similar argument applies to $\{(1,1),(0,1)\}$. The 'diagonal' function that rewards in outcomes $(0,1)$ and $(1,0)$ may be feasible, but in all cases it is weakly more expensive than the easy reward scheme. Both are implemented by the same wages, but the easy scheme offers better incentives because it rewards the most exonerating outcome $(0,0)$. The 'diagonal' function that rewards in outcomes $(0,0)$ and $(1,1)$ may be feasible, but in all cases it is strictly more expensive than the hard reward scheme because the latter does pay rewards in the incriminating outcome $(1,1)$.

There are four subsets of size 3. One of them yields the easy reward function. The set that rewards all outcomes except $(0,0)$ cannot satisfy (IC). The set that rewards all outcomes except $(0,1)$ may be feasible if $\Delta_{01} < 0$, but is always more expensive than the moderate reward scheme. The same is true for the set $\{(0,1),(1,0),(1,1)\}$.

The subsets of size 0 and size 6 entail all transfers being equal, which does not satisfy (IC). Therefore these cannot give solutions.

The costs are easily calculated using proposition 1. It remains to give examples to show that there exist parameters for which each scheme is optimal. The easy scheme is optimal under the parameters $\pi_E := \begin{bmatrix} 0.3 & 0.3 \\ 0.3 & 0.1 \end{bmatrix}$ and $\tau_E := \begin{bmatrix} 0 & 0.1 \\ 0.1 & 0.8 \end{bmatrix}$, which yield the differences $\Delta_E := \begin{bmatrix} 0.4 & 0.2 \\ 0.2 & -0.7 \end{bmatrix}$. These parameters yield costs

$$c_E^1 = \frac{10}{7} < c_E^2 = \frac{11}{7} < c_M = c_C = 2 < c_H^2 = 3 < c_H^1 = \frac{10}{3}.$$

The moderate scheme is optimal under the parameters $\pi_M := \begin{bmatrix} 0.8 & 0 \\ 0.2 & 0 \end{bmatrix}$ and $\tau_M := \begin{bmatrix} 0 & 0.1 \\ 0.1 & 0.8 \end{bmatrix}$, which yield the differences $\Delta_M := \begin{bmatrix} 0.8 & -0.1 \\ 0.1 & -0.8 \end{bmatrix}$. These parameters yield costs

$$c_M = \frac{10}{9} < c_E^1 = c_E^2 = c_H^2 = c_H^1 = \frac{5}{4} < c_C = 2.$$

The hard scheme is optimal under the parameters $\pi_H := \begin{bmatrix} 0.9 & 0 \\ 0 & 0.1 \end{bmatrix}$ and $\tau_H := \begin{bmatrix} 0 & 0.1 \\ 0.1 & 0.8 \end{bmatrix}$, which yield the differences $\Delta_H := \begin{bmatrix} 0.9 & -0.1 \\ -0.1 & -0.7 \end{bmatrix}$. These parameters yield costs

$$c_H^2 = 1 < c_H^1 = \frac{10}{9} < c_M = \frac{10}{8} < c_E^1 = \frac{10}{7} < c_E^2 = \frac{11}{7} < c_C = \infty.$$

The cheeky scheme is optimal under the parameters $\pi_C := \begin{bmatrix} 0.6 & 0.4 \\ 0 & 0 \end{bmatrix}$ and $\tau_C := \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix}$, which yield the differences $\Delta_C := \begin{bmatrix} 0.5 & 0.0 \\ 0.4 & -0.9 \end{bmatrix}$. These parameters yield costs

$$c_C = 1 < c_E^1 = c_E^2 = c_M = \frac{10}{9} < c_H^2 = \frac{9}{5} < c_H^1 = \frac{10}{5}.$$

□

5 One or two monitors?

Theorem 1.

1. *If the agent cannot collude with both monitors simultaneously, then the principal may be strictly better or strictly worse off when each signal is received by a different monitor.*
2. *Otherwise, the principal is always weakly, and sometimes strictly, better off when both pieces of evidence are accessed by the same monitor.*

Proof. The key point to note from proposition 2 is that $c_E^1 \leq c_E^2$ and $c_H^2 \leq c_H^3$. All that remains to prove theorem 1 is to find parameters π and τ where these inequalities are strict.

The principal is strictly better off in scenario 2 than in scenario 1 under the parameters (π_H, τ_H), because then $c_H^2 < \min\{c_E^1, c_M, c_H^1, c_C\}$. But the principal is strictly better off in scenario 1 than in scenario 2 under the parameters (π_E, τ_E), because then $c_E^1 < \min\{c_E^2, c_M, c_H^2, c_C\}$. This proves the first statement.

Lemma 1 shows that the principal is more constrained in scenario 3 than in scenario 1, so if the agent can collude with both monitors then she is weakly better off when both pieces of evidence are accessed by the same monitor. The proof of proposition 2 shows that the principal is strictly better off in scenario 1 than in scenario 3 under the parameters (π_E, τ_E) because then $c_E^1 < \min\{c_E^2, c_M, c_H^3, c_C\}$. □

It follows from the proof of theorem 1 that one monitor is strictly better when $c_E^1 = \frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}} < \min\{c_H^2, c_M, c_C\}$, no matter whether or the agent can bribe both monitors. This holds only when $\pi_{01} > \tau_{01}$, $\pi_{10} > \tau_{10}$ and π_{11} is small, which means that $\pi_{01}\pi_{10} > \tau_{01}\tau_{10}$. Thus one monitor tends to be preferred if the signals are more positively correlated when the agent takes the bad action than when the agent takes the good action, because this implies that $\pi_{11}\pi_{00} - \pi_{01}\pi_{10}$ is small and $\tau_{11}\tau_{00} - \tau_{01}\tau_{10}$ is big. Conversely, two monitors are strictly better than one only if the agent cannot bribe both monitors, and if $c_H^2 = \frac{1-\pi_{11}}{\Delta_{00}} < \min\{c_E^1, c_M, c_C\}$, which tends to be true in the reverse case.

In the special case where both monitors observe exactly the same information, we get that $\pi_{01} = \pi_{10} = \tau_{01} = \tau_{10} = 0$ which implies that $\Delta_{10} = \Delta_{01} = 0$. Hence $c_H^2 < c_E^1 = c_M = c_H^1 < c_H^2 < c_C$, so the principal strictly prefers two monitors if the agent cannot bribe them both; otherwise, she is indifferent.

6 Robustness

In this section, I show that the main result (theorem 1) continues to hold under various alternative assumptions.

6.1 Coalitions without the Agent

In the main text, I have considered schemes that are robust to collusion between the agent and one or both monitors. There is also a risk that that the principal might bribe the monitors to conceal evidence. This might seem counter-intuitive, because I have been focussing on the problem of incentivising the monitors to truthfully report their evidence. But the reason for having the monitors truthfully report their evidence is not that the principal inherently cares about the information, but rather that the principal has to provide credible incentives for the agent to take the right action. By the time the monitor has received the evidence, the agent has already taken his action, so the principal could gain by bribing the monitors to conceal evidence that would otherwise require the principal to reward the agent. If the agent anticipates this then she has no incentive to take the right action.

For example, the cheeky scheme S_C only pays rewards when monitor 2 reports evidence: it rewards the agent in outcome $(0,1)$ and monitor 1 in outcome $(1,1)$. Therefore the principal would be willing to pay monitor 2 any bribe $b < T^C(0,1)$ to conceal their evidence, and monitor 2 would be willing to accept any bribe $b > 0$. So the cheeky scheme is not principal-bribe proof (as opposed to agent-bribe proof). The only way to make the cheeky scheme bribe proof is for the principal to commit to pay the same total quantity rewards in the outcomes $(0,0)$ and $(1,0)$ (when monitor 2 does not report evidence) as in the outcomes $(0,1)$ and $(1,1)$ (when they do). But this undermines the advantage of the cheeky scheme, which was to avoid paying rewards in these outcomes. If the principal is forced to respect principal-bribe proofness then the cost of the scheme increases from $\frac{\pi_{01}+\pi_{11}}{\Delta_{01}}$ to $\frac{1}{\Delta_{01}}$, which is strictly more than the cost of the moderate scheme, $\frac{1}{\Delta_{01}+\Delta_{00}}$.

In general, a scheme is robust to collusion between the principal and a single monitor if they can never increase their joint surplus by concealing evidence:

$$
\begin{aligned}
& w_i(s) - (T(s) + \sum_{j=1,2} w_j(s)) \geq w_i(m_i, s_{-i}) - (T(m) + \sum_{j=1,2} w_j(m_i, s_{-i})) && \forall m_i \leq s_i \text{ and } i = 1,2 \\
\iff & T(m) + w_{-i}(m_i, s_{-i}) \geq T(s) + w_{-i}(s) && \forall m_i \leq s_i \text{ and } i = 1,2.
\end{aligned}
$$

These constraints only rule out the cheeky scheme. For instance, if $i = 2$, $s = (0,1)$ and $m_2 = 0$ then the left side equals 0 and the right side equals $\frac{1}{\pi_{01}}$, so the constraint is violated. A scheme is robust to collusion between the principal and both monitors if:

$$
\begin{aligned}
\sum_{j=1,2} w_j(s) - (T(s) + \sum_{j=1,2} w_j(s)) \geq \sum_{j=1,2} w_j(m) - (T(m) + \sum_{j=1,2} w_j(m)) && \forall m \leq s \\
\iff T(m) \geq T(s) && \forall m \leq s.
\end{aligned}
$$

This says that the agent's reward has to increase when the monitors fail to provide hard, incriminating evidence. Like the single monitor constraints, they rule out the cheeky scheme. However, the easy, moderate and hard schemes respect all of these constraints, so there is no change to the main results. The only impact is to expand the regions where the moderate and easy schemes are optimal.

Another possibility is that the monitors might collude with each other, without the agent or the principal. For example, in the scenario 2 easy scheme S_E^2, the monitors get a joint payoff of 0 in the outcome $(1,1)$, but they get a collective payoff of $\frac{1}{\Delta_{00}+\Delta_{01}+\Delta_{10}}$ in the outcomes $(0,1)$ and $(1,0)$. Each monitor would be happy to bribe the other monitor to suppress their evidence, and each monitor would be happy to accept the other's bribe. Therefore, this scheme is not robust to monitor-monitor collusion, and can only be made robust by increasing the monitor's collective payoff in $(1,1)$. But doing so yields the scenario 1 easy scheme, S_E^1, which undermines the benefits of the second monitor. In general, a scheme is monitor-monitor bribe proof if

$$
\sum_{i=1,2} w_i(s) \geq \sum_{i=1,2} w_i(m)
$$

for all $m \leq s$. The easy, moderate and hard schemes respect these constraints. Thus, even if the agent cannot bribe both monitors, a single monitor is still weakly better if there is a risk that the two monitors collude with one another.

6.2 Limited liability

The assumption the players have limited liability rules out the use of punishments, so the principal has to use rewards to provide incentives. This implies that the agent must earn rent from any feasible scheme. In this section, I discuss the implications of replacing the players' limited liability constraints with voluntary participation constraints. Whereas limited liability constraints require that a player's ex post payoff is weakly positive, a voluntary participation constraint instead requires that their ex ante payoff is positive. This means that punishments are permitted, so long as they are balanced by enough rewards to ensure that, on average, the player is indifferent between the scheme and the status quo.

I begin by considering the relaxed problem in which the agent's (LLA) constraint is replaced with a voluntary participation constraint (VPA) requiring that $\sum_{s \in S} \pi_s T(s) \geq 0$. In this case, the principal can reclaim the agent's rent by charging him a fine (negative transfer) in outcomes where she does not reward him. Proposition 2 shows that one of four schemes solves the principal's problem in the (LLA) case. Proposition 3 says that one of the same four schemes can be modified to solve the principal's problem in the (VPA) case, by subtracting the agent's expected reward from his ex post reward. Formally, if $S = (T, w_1, w_2)$ then define $\mathrm{VP}_A(S) := (T - \mathop{\mathbb{E}}_{s \sim \pi}[T], w_1, w_2)$.

Proposition 3. *Suppose the agent's limited liability constraints are replaced with a voluntary participation constraint*

$$\sum_{s\in S} \pi_s T(s) \geq 0. \tag{VPA}$$

1. *One of the following four schemes is optimal in scenario 1 (monitor 1 receives both signals s_1 and s_2): $VP_A(S_E^1), VP_A(S_M), VP_A(S_H^1)$, or $VP_A(S_C)$. The costs of these schemes are*

$$\begin{aligned}\tilde{c}_E^1(\pi,\Delta) &:= \frac{1-\pi_{00}-\pi_{01}-\pi_{10}}{\Delta_{00}+\Delta_{01}+\Delta_{10}}\\ \tilde{c}_M(\pi,\Delta) &:= \frac{1-\pi_{00}-\pi_{01}}{\Delta_{00}+\Delta_{01}}\\ \tilde{c}_H^1(\pi,\Delta) &:= \frac{1-\pi_{00}}{\Delta_{00}}\\ \tilde{c}_C(\pi,\Delta) &:= \frac{\pi_{11}}{\Delta_{01}}\end{aligned}$$

   respectively. The cost of the optimal scheme is equal to the lower envelope of these costs, $\min\{\tilde{c}_E^1, \tilde{c}_M, \tilde{c}_H^1, \tilde{c}_C\}$*.*

2. *One of the following four schemes is optimal in scenario 2 (monitor i receives signal s_i; the agent can bribe either one but not both monitors): $VP_A(S_M)$, $VP_A(S_C)$, $VP_A(S_E^2)$ or $VP_A(S_H^2)$. The costs of schemes $VP_A(S_E^2)$ and $VP_A(S_H^2)$ are*

$$\begin{aligned}\tilde{c}_E^2 &:= \frac{1+\pi_{11}-\pi_{00}-\pi_{01}-\pi_{10}}{\Delta_{00}+\Delta_{01}+\Delta_{10}}\\ \tilde{c}_H^2 &:= \frac{1-\pi_{11}-\pi_{00}}{\Delta_{00}}.\end{aligned}$$

   The cost of the optimal scheme is equal to the lower envelope of the costs, $\min\{\tilde{c}_E^2, \tilde{c}_M, \tilde{c}_H^2, \tilde{c}_C\}$*.*

3. *One of the following four schemes is optimal in scenario 3 (monitors i receives signal s_i; the agent can bribe one, other or both of them): $VP_A(S_E^2), VP_A(S_M), VP_A(S_H^1)$ or $VP_A(S_C)$. The cost of the optimal scheme is equal to the lower envelope of the costs,* $\min\{\tilde{c}_E^2, \tilde{c}_M, \tilde{c}_H^1, \tilde{c}_C\}$*.*

Proof. I use the fact that the principal's objective depends on the levels of the transfers, whilst the constraints depend on their differences.

First, note that the voluntary participation constraint has to bind; otherwise, the principal could strictly improve her payoff, without violating any constraints, by reducing all of the agent's transfers by a small enough amount. This means that the agent earns no rent in any optimal solution.

Now define a new variable $\tilde{T}(s) := T(s) - T(1,1)$ to be the difference between the agent's reward in outcome s and his reward in outcome $(1,1)$. Note that $\sum_{s\in S} \pi_s \tilde{T}(s) = \sum_{s\in S} \pi_s T(s) - T(1,1)$ because π_s sum to 1; that $\sum_{s\in S} \Delta_s \tilde{T}(s) = \sum_{s\in S} \Delta_s T(s)$ because the Δ_s sum to 0; and that $\tilde{T}(s) - \tilde{T}(s') = T(s) - T(s')$ for all $s, s' \in S$. Therefore, we can restate the principal's problem as

$$\begin{aligned}\max_{\tilde{T},T(1,1)} \; & -\sum_{s\in S} \pi_s(\tilde{T}(s) + W^j(s;\tilde{T}))\\ \text{st.} \; & \sum_{s\in S} \Delta_s \tilde{T}(s) \geq 1 && \text{(IC)}\\ & \sum_{s\in S} \pi_s \tilde{T}(s) = T(1,1), && \text{(VPA)}\end{aligned}$$

where W^j is defined by proposition 1 (like before), but with T replaced by $\tilde{T}$. Thus the proof of lemma 3 tells us that $\tilde{T}(s) = \frac{\mathcal{I}(s\in R)}{\sum_{s\in R}\Delta_s}$. This implies that $T(1,1) = \frac{\sum_{s\in R}\pi_s}{\sum_{s\in R}\Delta_s}$, so $T^*(s) = \frac{\mathcal{I}(s\in R) - \sum_{s\in R}\pi_s}{\sum_{s\in R}\Delta_s}$.

The proof of proposition 2 then tells us that the schemes listed in the statement are putative optimal schemes. Their costs are easily derived by subtracting off the rent that the agent receives in the corresponding schemes in proposition 2.

□

Thus relaxing the agent's liability constraint does not affect the solution in any important way. Comparing the costs of the schemes in proposition 3 with those in proposition 2 show that the cost of the easy scheme decreases the most when (LLA) is replaced with (VPA). The easy schemes are cheaper in the one monitor scenario than in the two monitor scenarios, so relaxing the (LLA) constraints only reinforces the conclusion of theorem 1 that one monitor is better than two.

If the monitors' liability constraints are both relaxed to voluntary participation constraints, then neither of them earns any rent. In this case, there is no difference between a single monitor and two monitors — their collective surplus is zero in both cases. Instead, I assume that the monitors have some liability limit $L > 0$, where L is small enough to be binding. For $S = (T, w_1, w_2)$ define $\mathrm{VP}_M(S) := (T, w_1 - L, w_2 - L)$.

Proposition 4. *Suppose the monitor's limited liability constraints are replaced by weaker liability limits, $w_i(s) \geq -L$, for some exogenous liability limit $L > 0$, and voluntary participation constraints*

$$\sum_{s \in S} \pi_s w_i(s) \geq 0. \tag{VPM}$$

One of the following four schemes is optimal in scenario 1 (monitor 1 receives both signals s_1 and s_2): $VP_M(S_E^1), VP_M(S_M), VP_M(S_H^1)$, or $VP_M(S_C)$. The costs of these schemes are

$$\begin{aligned}
\check{c}_E^1(\pi, \Delta) &:= \frac{1}{\Delta_{00} + \Delta_{01} + \Delta_{10}} - \min\{L, \frac{\pi_{11}}{\Delta_{00} + \Delta_{01} + \Delta_{10}}\} \\
\check{c}_M(\pi, \Delta) &:= \frac{1}{\Delta_{00} + \Delta_{01}} - \min\{L, \frac{\pi_{11} + \pi_{10}}{\Delta_{00} + \Delta_{01}}\} \\
\check{c}_H^1(\pi, \Delta) &:= \frac{1}{\Delta_{00}} - \min\{L, \frac{\pi_{11} + \pi_{10} + \pi_{01}}{\Delta_{00}}\} \\
\check{c}_C(\pi, \Delta) &:= \frac{\pi_{01} + \pi_{11}}{\Delta_{01}} - \min\{L, \frac{\pi_{11}}{\Delta_{01}}\}
\end{aligned}$$

respectively. The cost of the optimal scheme is equal to the lower envelope of these costs, $\min\{\check{c}_E^1, \check{c}_M, \check{c}_H^1, \check{c}_C\}$.

One of the following four schemes is optimal in scenario 2 (monitor i receives signal s_i; the agent can bribe either one but not both monitors): $VP_M(S_M)$, $VP_M(S_C)$, $VP_M(S_E^2)$ or $VP_M(S_H^2)$. The costs of schemes $VP_M(S_E^2)$ and $VP_M(S_H^2)$ are

$$\begin{aligned}
\check{c}_E^2 &:= \frac{1 + \pi_{11}}{\Delta_{00} + \Delta_{01} + \Delta_{10}} - 2\pi_{11} L \\
\check{c}_H^2 &:= \frac{1}{\Delta_{00}} - (\pi_{01} + \pi_{10}) L
\end{aligned}$$

The cost of the optimal scheme is equal to the lower envelope of the costs, $\min\{\check{c}_E^2, \check{c}_M, \check{c}_H^2, \check{c}_C\}$.

One of the following four schemes is optimal in scenario 3 (monitors i receives signal s_i; the agent can bribe one, other or both of them): $VP_M(S_E^2), VP_M(S_M), VP_M(S_H^1)$ or $VP_M(S_C)$. The cost of the optimal scheme is equal to the lower envelope of the costs $\min\{\check{c}_E^2, \check{c}_M, \check{c}_H^1, \check{c}_C\}$.

This result is similar to proposition 3, but now the cost saving for each scheme is proportional to the probability of rewarding the monitors, rather than to the probability of rewarding the agent. The monitors are most likely to be rewarded in the hard scheme (when the agent is least likely to be rewarded), so the hard schemes become relatively more attractive when the monitors have greater liability limits. Since the hard scheme is weakly cheaper in the two monitor scenarios, and strictly so in scenario 2, this suggests that two monitors are relatively more attractive than one when they have relaxed liability constraints.

If monitor i has some positive liability limit $L_i > 0$, then it is always optimal for the principal to reward the agent when monitor i reports exonerating evidence, because she can reclaim the monitor i's wages. Hence the signal s_i will be used even though it may not be very accurate. Finally, if the agent's and at least one of the monitor's liability constraints are relaxed to voluntary participation constraints, then the principal can incentivise the right action for free with a single monitor.

6.3 More than two monitors

Suppose there is a set $I = \{1, \ldots, n\}$ of signals, each received by a different monitor $i \in I$. Redefine the signal space $S := \{0,1\}^n$, the distribution of signals when the right action is taken, $\pi \in \Delta\{0,1\}^n$ and the distribution of signals when the wrong action is taken, $\tau \in \Delta\{0,1\}^n$. For subset $J \subseteq I$, define $s_J := (s_j)_{j \in J}$, $s_{-J} := (s_i)_{i \notin J}$, and $\pi_J := \sum_{s|s_J=1} \pi_s$. If the agent can bribe all n-monitors, then the principal is better off with a single monitor, for the same reason that she is in the two monitor case. But if the agent cannot bribe all the monitors, then the principal may be better off with more monitors. Conjecture 1 considers the case where the agent can bribe a maximum of one monitor.

Conjecture 1.

1. *If all pieces of evidence are accessed by monitor 1, so that the principal must respect the agent-monitor bribery constraints*
$$T(s) + \sum_{i \in I} w_i(s) \geq T(m) + \sum_{i \in I} w_i(m),$$
for all $m \leq s$, then
   (a) *any feasible scheme $(T, (w_i)_{i \in I})$ must satisfy*
   $$w_1(s) = \max\{T(s) - T(m_J, s_{-J}) \mid m_J \leq s_J\};$$
   (b) *the optimal scheme has a transfer function of the form $T(s) = \mathcal{I}(s \in R)/\sum_{s \in R} \Delta_s$, where $R \subset S$ contains an outcome s only if it contains the lower contour set of s;*
   (c) *the cost of the optimal mechanism is $\frac{1}{\sum_{s \in R} \Delta_s}$.*
2. *If each signal is received by a different monitor and the agent can only bribe a single monitor, so that the principal must respect the agent-monitor bribery constraints*
$$T(s) + w_i(s) \geq T(0, s_{-i}) + w_i(0, s_{-i}),$$
for all $i \in I$, then
   (a) *any feasible scheme $(T, (w_i)_{i \in I})$ must satisfy*
   $$w_i(s) \geq \max\{0, T(s) - T(0, s_{-i})\}$$
   for all $s \in S$;
   (b) *the optimal scheme has a transfer function of the form as in (1a).*
   (c) *the cost of the optimal mechanism is $\frac{\sum_{s \in R} \pi_s}{\sum_{s \in R} \Delta_s}$.*

The conjecture says that the optimal mechanism may be easy (if $R = S \setminus \{(1, \ldots, 1)\}$), hard (if $R = \{(0, \ldots, 0)\}$), or somewhere in between (if, up to reordering, R contains $(1, 0, \ldots, 0)$ but not $(1, \ldots, 1, 0)$). Like the two monitor case, each monitor gets rewarded when they alone report evidence, but not when others report evidence. The cost savings from hiring more monitors are potentially large if the optimal set R is small. E.g. if the optimal scheme is a 'hard' one that only rewards the outcome $\mathbf{0} = (0, \ldots, 0)$where all n monitors report exonerating evidence, then the cost of the n-monitor scheme is $\pi_{\mathbf{0}} < 1$ times the cost of the one-monitor scheme.

6.4 Signal structures

I have so far restricted attention to the simplest possible information structure with which to cleanly compare one and two monitor scenarios. If the monitors can receive more than two signals, and can pay to manipulate signals, then the result is largely unchanged — the principal rewards the agent when the monitors report exonerating evidence, and reports the monitors when they report hard, incriminating evidence. If there is a chance of receiving hard, exonerating evidence and principal-monitor collusion is not possible, then the principal need not reward the monitors at all because they have no conflict of interest. But if principal-monitor collusion is possible then the principal must reward the agent and the monitors when they report hard exonerating evidence, in order to provide credible incentives for the agent.

The case where both monitors observe both signals is isomorphic to the case where they each observe one signal, and the signals are perfectly correlated. The case where monitor 1 observes both signals and monitor 2 observes a single signal is isomorphic to the case where monitor 1 observes a signal with four realisations, $s_1 = (0,0), (0,1), (1,0)$ and $(1,1)$, monitor 2 observes a binary signal $s_2 = 0$ or 1, and $\mathbb{P}[s_1 \in \{(0,0),(1,0)\}, s_2 = 1] = \mathbb{P}[s_1 \in \{(0,1),(1,1)\}, s_2 = 0] = 0$.

7 Applications

In this section I discuss how the previous results apply to three different contexts: journalism, whistle-blowing and joint financial auditing.

Journalists (the 'monitors') relate verifiable evidence to the public (the 'principal') about the behaviour of firms, politicians or public institutions (the 'agent'). The public pay journalists for publicising useful information, either through direct purchase of media or through increased advertising revenue. The information relayed by journalists can be used to punish or reward firms and politicians.[4] However, journalists are susceptible to corruption: they may be bribed to suppress evidence or to create fake stories, or they may be threatened for publicising evidence. Although journalists may be punished for reporting fake news, they are not punished for failing to report real news, so in this respect, they have limited liability.

These stylised facts match the main features of the model. Although I do not attempt to model fake evidence, we can think of 'fake' news as noisy, soft evidence. If there are many monitors, and the agent cannot bribe them all, then proposition 3 and conjecture 1 predict that the public will punish the agent whenever at least one journalist reports incriminating evidence, and journalists will be rewarded only when they alone report evidence. Moreover, the size of each monitor's reward should be proportional to the size of the agent's fine. Reality is not quite so stark as this, but it does seem to be true that a journalist's payoff (including revenue from the story and future earnings from improved career prospects) is increasing in the agent's fine (the size of the story), and decreasing in the number of other journalists who cover the story.

Many regulators offer rewards for hard evidence to whistle-blowers. For instance, in the US the False Claims Act (the 'principal') offers a reward to whistle-blowers ('monitors') who provide evidence that can be used to convict a party (the 'agent') of fraud. The whistle-blower's reward is usually between 15 and 25% of the fine enacted on the convicted party. Similarly, various U.S. acts, including the Clean Water and Clean Air Acts allow any U.S. citizen to file a lawsuit against an offending party. In many cases, the citizens bringing these cases receive a de facto reward for doing so since the amount given as compensation for litigation costs often exceeds the actual litigation. Competition regulation also rewards whistle-blowers for obtaining evidence of breaches of competition law because anti-competitive behaviour is generally harmful to other firms and consumers, so it is in the interests of other firms and consumers to enforce the law. Additionally, some regulators offer explicit financial rewards, for example, the UK Competition and Markets Authority offers rewards of up to a hundred thousand pounds to whistle-blowers who supply evidence of anti-competitive behaviour.

[4] For example, facts uncovered by journalists at the Guardian and Channel 4 have been used as evidence to fine the Brexit Campaign for breaking electoral rules.

In these situations there coexist a team of ‘public inspectors’ who obtain low quality evidence with high probability, and a set of ‘whistle-blowers’ who receive highly accurate evidence with low probability. The baseline model of this paper would suggest that only the whistle-blowers are hired and that they are only rewarded when they obtain incriminating evidence. It seems likely that the public inspectors may have greater liability limits than whistle-blowers, in which case section 6.2 predicts that the public inspectors will always be hired and that whistle-blowers will only be hired if the public inspector’s liability limit is lower than the size of reward required to satisfy the agent’s (IC) constraint. However, the feature that no whistle-blowers are rewarded when they all report incriminating evidence is not explained by heterogeneous monitor liability constraints, rather it is a feature of otherwise optimal mechanisms which do not respect the agent-multi-monitor-coalition constraints (scenario 2). This seems to be a likely explanation because it is in some contexts difficult to imagine an agent colluding with all the possible whistle-blowers (e.g. local citizens of a polluted lake), so it may not be necessary for the mechanism to respect collusion constraints for certain, large coalitions.

Many countries including France, Germany, and Switzerland, require legal entities to undergo joint financial audits Vanstraelen et al. (2009). Joint financial audits involve two independent accountancy firms working together to produce an audit. Both firms are bear responsibility for the final report, and are liable to be sued for damages arising from inaccurate reports, otherwise they are paid a fixed wage. Legal entities may be fined, sued or otherwise punished in proportion to the amount of incriminating evidence reported. Possible collusion between the auditee and auditors is well known to be a problem since the auditee generally chooses the auditors themselves. Indeed, part of the rationale for joint audits is to reduce the costs of deterring collusion. This outcome is consistent with the predictions of section 6.2 when the liability limit for both monitors is strictly positive, but not large enough to rely on a single monitor.

8 Conclusion

Is it better to receive information from a single monitor, or from two strategically independent monitors? The answer depends on the optimal reward structure, which is, in turn, a function of the (primitive) distribution of evidence. If both types of hard evidence are incriminating enough then the “easy” reward scheme is optimal. But then the agent can receive a reward by bribing either one of the two monitors. The principal can only deter bribes by rewarding both monitors, so she is no better off than when both signals are received by a single monitor. If one signal is more incriminating than the other then the “moderate” scheme is optimal. But this scheme only relies on the superior evidence, so the principal is indifferent between a monitor who receives both types of evidence and a monitor who receives superior evidence only. If neither type of evidence is incriminating enough then the “hard” scheme is optimal. If the agent cannot bribe both monitors, then the principal can save money in the two monitor case by only rewarding each monitor for reporting their evidence when the other fails to do so. This is consistent with real world institutions such as journalism and whistleblowing. But it is not robust to larger collusive coalitions. If the agent can bribe both monitors, then this scheme is vulnerable to larger scale cover-ups because neither monitor stands to be rewarded when they both report evidence. The fact that the agent can act collectively with both monitors as easily as he can with one monitor, means that the principal is weakly better off with a single monitor who receives both signals.

Is it reasonable to assume that the agent can bribe two monitors as easily as a single one? A key assumption in this paper is that bribery negotiations take place under common knowledge of the realised signal s and of the player’s respective utility functions. But in reality, there may be a degree of uncertainty about precisely what evidence monitors hold, about the legal implications of this evidence, or about each player’s “moral cost” of engaging in bribery. The need to negotiate with differing private evaluations of these factors may not be conducive to the formation of larger side-coalitions. If this is the case, then the principal may prefer information to be received by a larger number of monitors so that she can rely more on their pre-existing asymmetric information to deter bribes (which is free for her), and less on rewards (which are expensive). Another possibility is that the principal can endogenously create asymmetric information with the specific objective of undermining coalition formation (e.g. Ortner and Chassang, 2018). I consider this question in chapter 2.

References

Aoyagi, M. (2005). Collusion through mediated communication in repeated games with imperfect private monitoring. *Economic Theory 25*(2), 455–475.

Beck, J. (2000). The false claims act and the english eradication of qui tam legislation. *North Carolina Law Review 78*(3), 539–642.

Ben-Porath, E. and M. Kahneman (1996). Communication in repeated games with private monitoring. *Journal of Economic Theory 70*(2), 281–297.

Ben-Porath, E. and B. Lipman (2012). Implementation with partial provability. *Journal of Economic Theory 147*(5), 1689–1724.

Che, Y.-K. and J. Kim (2006). Robustly collusion-proof implementation. *Econometrica 74*(4), 1063–1107.

Crémer, J. (1996). Manipulations by coalitions under asymmetric information: The case of groves mechanisms. *Games and Economic Behavior 13*(1), 39–73.

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013, 09). Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*. *The Quarterly Journal of Economics 128*(4), 1499–1545.

Felli, L. and R. Hortala-Vallve (2016, 11). Collusion, blackmail and whistle-blowing *. *Quarterly Journal of Political Science 11*.

Green, J. and J.-J. Laffont (1979, 04). On Coalition Incentive Compatibility. *The Review of Economic Studies 46*(2), 243–254.

Green, J. R. and J.-J. Laffont (1986). Partially verifiable information and mechanism design. *The Review of Economic Studies 53*(3), 447–456.

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

Koessler, F. and E. Perez-Richet (2019). Evidence reading mechanisms. *Social Choice and Welfare 53*(3), 375–397.

Kofman, F. and J. Lawarrée (1993). Collusion in hierarchical agency. *Econometrica 61*(3), 629–656.

Laffont, J.-J. and D. Martimort (1997). Collusion under asymmetric information. *Econometrica 65*(4), 875–912.

Laffont, J.-J. and D. Martimort (2000). Mechanism design with collusion and correlation. *Econometrica 68*(2), 309–342.

Langpap, C. and J. P. Shimshack (2010). Private citizen suits and public enforcement: Substitutes or complements? *Journal of Environmental Economics and Management 59*, 235–249.

Liu, Z. J., Z. Wang, and Z. Yin (2022). When is duplication of effort a good thing in law enforcement? *Journal of Public Economic Theory n/a*(n/a).

McAfee, R. P., H. M. Mialon, and S. H. Mialon (2008). Private v. public antitrust enforcement: A strategic analysis. *Journal of Public Economics 92*(10), 1863 – 1875.

Ortner, J. and S. Chassang (2018). Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy 126*(5), 2108–2133.

Polinsky, A. M. (1980). Private versus public enforcement of fines. *The Journal of Legal Studies 9*(1), 105–127.

Rahman, D. (2012, May). But who will monitor the monitor? *American Economic Review 102*(6), 2767–97.

Strausz, R. (1997). Delegation of monitoring in a principal-agent relationship. *Review of Economic Studies 64*(3), 337–357.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, & Organization 2*(2), 181–214.

Vafaï, K. (2005). Collusion and organization design. *Economica 72*(285), 17–37.

Vanstraelen, A., C. Richard, and J. R. Francis (2009, 11). Assessing france's joint audit requirement: Are two heads better than one? *Auditing A Journal of Practice & Theory 28*.