B Supplemental Appendix: Interpreting Commitment

In this appendix, we provide three alternative interpretations of the extortionary threats at the heart of our analysis. Appendix A. re-interprets threats as the outcome of bargaining between the principal and each agent. Appendix B. considers different types of behavioral preferences. Preferences for reciprocity and those for keeping one's word lead to extortion, while preferences for truth-telling lead to cooperation. Appendix C. re-interprets commitment to μ_t as an equilibrium refinement of the no-extortion game.

A. Misuse as a Bargaining Outcome

In the proof of Proposition 2, agent t's profitable deviation is to shirk and then demand payment by threatening the principal. In response, the principal pays and the message leads to her maximum continuation payoff, which resembles the outcome of a renegotiation in which agent t demands a bribe in exchange for sending the principal's preferred message.

We formalize this idea by studying a game with *interim renegotiation and bargaining.* Consider the following one-shot game between the principal and a single agent:

1. The agent chooses effort, $e \geq 0$.

2. The principal pays a transfer, $s \geq 0$.

3. The agent sends a message, $m \in M$.

We assume that the principal and the agent Nash bargain over $s \geq 0$ and $m \in M$, where the agent's bargaining weight is $\alpha \in [0, 1]$ and the disagreement point is message m^o and transfer s^o. The principal's and the agent's payoffs are $e - s + \frac{\delta}{1-\delta}\Pi(m)$ and $s - c(e)$, respectively, where $\Pi : M \to \mathbb{R}_+$ is an arbitrary function and $c(\cdot)$ is strictly increasing. We interpret this model as a single period of the no-extortion game, where $\Pi(m)$ is the principal's continuation payoff.

As in the extortion game, the agent exerts zero effort in every equilibrium of this game.

Proposition 7 *Consider the game with interim renegotiation and bargaining. In any equilibrium,* $e = 0$ *and* $m \in \arg\max_{\tilde{m}} \Pi(\tilde{m})$.

Proof of Proposition 7: The principal's and the agent's payoffs at the disagreement point are $e - s^o + \frac{\delta}{1-\delta}\Pi(m^o)$ and $s^o - c(e)$, respectively. Under Nash bargaining, the outcome, (s, m), satisfies $m \in \arg\max_{\tilde{m}} \Pi(\tilde{m})$.

Let $\overline{\Pi} \equiv \max_{\tilde{m}} \Pi(\tilde{m})$; then, the agent's Nash bargaining payoff is

$$s^o - c(e) + \alpha\frac{\delta}{1-\delta}\left(\overline{\Pi} - \Pi(m^o)\right),$$

so that

$$s = s^o + \alpha\frac{\delta}{1-\delta}\left(\overline{\Pi} - \Pi(m^o)\right).$$

This transfer is independent of e, so $e = 0$ in any equilibrium. ■

To see the connection between this result and Proposition 2, suppose that $s^o = 0$ and $m^o = \arg\min_{\tilde{m}} \Pi(\tilde{m})$, with the resulting payoff $\underline{\Pi} = \min_{\tilde{m}} \Pi(\tilde{m})$. Then, the on-path transfer equals

$$s = \alpha\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}),$$

which is a fraction α of agent t's leverage, $L \equiv \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi})$. That is, just like Proposition 2, Proposition 7 holds because agent t's leverage, and hence s, are independent of effort. Thus, we can re-interpret commitment as a form of bargaining between the principal and each agent. Note that this result holds for *any* bargaining parameter $\alpha > 0$.[19]

As in the main text, agent t is willing to exert effort only if doing so increases his leverage,

[19] In some settings, the principal might lose the proceeds from effort following a disagreement. In that case, parties bargain over both e and the message. Nash bargaining would then result in a transfer equal to

$$s = \alpha\left(e + \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi})\right).$$

This transfer is increasing in e, so the agent would be willing to exert some effort. However, so long as the agent's leverage, $\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi})$, is independent of e, equilibrium effort would not depend on the severity of coordinated punishments. That is, misuse would continue to undermine *coordinated* punishments in this setting.

so the remedies that we consider in the paper also lead to effort in the game with interim bargaining and renegotiation.[20] The bargaining protocol also suggests new remedies. For instance, agent t would be willing to exert effort if his bargaining weight, α, or breakdown transfer, s^o, were increasing in effort, or if $\Pi(m^o)$ was decreasing in effort.

This model assumes that the principal is willing to pay the bargained-upon transfer, s. She is indeed willing to do so, provided that she would otherwise earn $\underline{\Pi}$, since

$$-s + \frac{\delta}{1-\delta}\overline{\Pi} \geq \frac{\delta}{1-\delta}\underline{\Pi}.$$

B. Misuse as a Result of Behavioral Preferences

We now turn to behavioral preferences that lead to misuse. Since agents are otherwise indifferent across their messages, even weak behavioral preferences can have dramatic effects on equilibrium outcomes. We show that preferences for keeping one's word and preferences for reciprocity lead to extortion, while preferences for truth-telling restore cooperation.

Consider the following **cheap-talk game with behavioral preferences:**

1. The agent chooses $e_t \geq 0$ and $\mu_t : \mathbb{R}_+ \to M$.

2. The principal chooses $s_t \geq 0$.

3. The agent chooses $m_t \in M$.

The principal's payoff is $e_t - s_t + \frac{\delta}{1-\delta}\Pi(m_t)$. The agent's payoff is $s_t - c(e_t) + H(\mu_t, e_t, s_t, m_t)$, where $H(\cdot)$ captures agent t's behavioral preferences.

[20]In this alternative formulation, dyadic relationships (Section V.) are potentially a less plausible remedy, since we might expect parties to Nash bargain over the outcome of the dyadic relationship as well as the message.

B..1 Word-keeping preferences

First, we consider **word-keeping** preferences (e.g., Vanberg (2008)), defined by

$$H(\mu_t, e_t, s_t, m_t) \equiv \begin{cases} \epsilon & m_t = \mu_t(s_t) \\ 0 & \text{otherwise} \end{cases}$$

for some $\epsilon > 0$. That is, the agent earns extra utility ϵ from following through on his threat.

Word-keeping preferences lead to misuse and zero effort in equilibrium.

Proposition 8 *For any $\epsilon > 0$, every equilibrium of the cheap-talk game with word-keeping preferences entails $e_t = 0$.*

Proof of Proposition 8: At the end of the game, $m_t = \mu_t(s_t)$ is the agent's uniquely optimal message. This is identical to the extortion game, so the profitable deviation in the proof of Proposition 2 is also profitable in this setting. ■

B..2 Reciprocal preferences

Next, we study **reciprocal** preferences, which we model by setting

$$H(\mu_t, e_t, s_t, m_t) \equiv \begin{cases} \epsilon\Pi(m_t) & s_t \geq s_{REF} \\ -\epsilon\Pi(m_t) & s_t < s_{REF} \end{cases},$$

where $s_{REF} \geq 0$ is some reference transfer and $\epsilon > 0$. Note that the reference transfer, s_{REF}, is independent of the agent's effort; this type of reciprocity is consistent with Fehr *et al.* (2020).

For most s_{REF}, reciprocal preferences lead to zero effort. The sole exception is when the principal is exactly indifferent between paying s_{REF} and earning her maximum continuation payoff or paying 0 and earning her minimum continuation payoff. Under that condition, positive effort can be sustained in equilibrium for a reason similar to Section B.: the principal is indifferent between paying an agent who exerts effort and not paying an agent who shirks.

Proposition 9 *Fix any $\epsilon > 0$. Let $\overline{m} \in \arg\max_m \Pi(m)$ and $\underline{m} \in \arg\min_m \Pi(m)$. If*

$$s_{REF} \neq \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m})),$$

then every equilibrium of the cheap-talk game with reciprocal preferences entails $e_t = 0$. Otherwise, for any e_t satisfying $c(e_t) \leq \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m}))$, there exists an equilibrium in which the agent chooses e_t.

Proof of Proposition 9: If $s_t \geq s_{REF}$, $m_t = \overline{m}$ is a best response for the agent. If $s_t < s_{REF}$, $m_t = \underline{m}$ is a best response for the agent. Therefore, the principal's optimal payment is either s_{REF} or 0. Her uniquely optimal payment is s_{REF} if

$$s_{REF} < \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m})).$$

Her uniquely optimal payment is 0 if

$$s_{REF} > \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m})).$$

In both cases, transfers are independent of e_t, so $e_t = 0$.

If

$$s_{REF} = \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m})),$$

then the principal is indifferent between $s = 0$ and $s = s_{REF}$. Thus, she is willing to play the following strategy: pay $s = s_{REF}$ if $e_t = e^*$ for some e^* such that $c(e^*) \leq \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m}))$ and $s = 0$ otherwise. The agent is then willing to choose $e_t = e^*$.

B..3 Truth-telling preferences

Finally, we study **truth-telling** preferences (e.g., Abeler *et al.* (2019)). Define a "cooperate" message, $m_t = C$, and a "deviate" message, $m_t = D$, such that $\Pi(C) > \Pi(D)$. Define the "desired effort," e^*, as an effort level that satisfies

$$c(e^*) \leq \frac{\delta}{1-\delta}(\Pi(C) - \Pi(D)). \tag{20}$$

Let

$$H(\mu_t, e_t, s_t, m_t) = \begin{cases} \epsilon 1\{m_t = C\} & e_t = e^*, s_t = c(e^*) \\ \epsilon 1\{m_t = D\} & \text{otherwise} \end{cases}.$$

Intuitively, the agent earns $\epsilon > 0$ extra utility for truthfully reporting cooperation by choosing $m_t = C$ when $e_t = e^*$ and $s_t = c(e^*)$, as well as for truthfully reporting a deviation by choosing $m_t = D$ following any other actions.

Truth-telling preferences result in equilibrium effort e^*.

Proposition 10 *Consider the cheap-talk game with truth-telling preferences. For any $\epsilon > 0$ and e^* satisfying (20), there exists an equilibrium with $e_t = e^*$.*

Proof of Proposition 10: At the end of the game,

$$m_t = \begin{cases} C & e_t = e^*, s_t = c(e^*) \\ D & \text{otherwise} \end{cases}.$$

If the agent chooses $e_t = e^*$, the principal is willing to pay $s_t = c(e^*)$ because (20) holds. If the agent chooses $e_t \neq e^*$, then $m_t = D$, so the principal is willing to pay $s_t = 0$. The agent is willing to choose $e_t = e^*$, because $s_t = c(e^*)$ if he does so and $s_t = 0$ otherwise. ■

C. Misuse as an Equilibrium Refinement

Commitment is a straightforward way to make sure that agents' threats are more than just cheap talk. Crucially, however, agents do not ever send *ex post* suboptimal messages in the extortion game. Therefore, commitment refines the set of equilibria in the extortion game and in each remedy.

Proposition 11 *For any equilibrium of the extortion game or of the extortion game with effort signals, transfer signals, or dyadic relationships, there exists an equilibrium of the corresponding no-extortion game that induces the same distribution over* $(e_t, s_t, m_t)_{t=0}^{\infty}$.

Proof: In the extortion game, this result is immediate, because agents are indifferent among messages and so are willing to follow their threats, so that the equilibrium in Proposition 2 remains an equilibrium. In the games with effort signals or transfer signals, agents are again indifferent over messages and so a nearly identical argument proves the result.

Consider the extortion game with dyadic relationships. Let σ^* be an equilibrium, and consider the following strategy profile of the game: in each period $t \geqslant 0$, the players play as in σ^* unless agent t deviates from $m_t = \mu_t(s_t)$. If he deviates, then $a_t = (l, l)$.

No player can profitably deviate from a_t, because a_t is always an equilibrium of the coordination game. By the choice of a_t following a deviation in m_t, agent t is willing to choose $m_t = \mu_t(s_t)$. Thus, the principal and agent t have no profitable deviation from e_t, μ_t, or s_t, since continuation play is exactly as in σ^*. So this strategy profile is an equilibrium of the no-extortion game. ■

Since agents are indifferent among messages, they are always willing to follow through on their threats. If they do, then the resulting mapping from transfer to message is identical to the corresponding mapping in the extortion game, leading to identical equilibrium outcomes.

C Supplemental Appendix: Principal Communication and Multiple Extortion Opportunities

This appendix explores alternative ways to model communication. In Appendix A., we show that extortion remains a problem even if the principal can send a public message at the end of each period. Intuitively, if the principal could lessen her punishment by reporting extortion, then she would always do so regardless of whether or not extortion actually occurred. Appendix B. then shows that extortion *can* be eliminated if the principal can commit to threats as a function of each period's transfer, provided that she makes her threat *weakly before* the agent makes his threat. This positive result should be interpreted with skepticism, however, since unlike the agents, the principal sometimes has an incentive to deviate from her threat.[21] Finally, once the principal pays an extorting agent, that payment is sunk and so the agent has an incentive to extort again. Appendix C. explores cooperation when agents have multiple opportunities to extort.

A. The Principal Can Send Messages

Let M_p be the set of messages for the principal, and m_p a typical message. In each period $t \geq 0$, the principal chooses a message $m_{p,t}$, and this message is publicly observed. We consider two different stage games, where the principal might choose $m_{p,t} \in M_p$ either before or after agent t chooses m_t. If the principal chooses $m_{p,t}$ before m_t is realized, we assume that μ_t is a function of s_t only (and so does not depend on $m_{p,t}$).

The principal talks after agent t. Consider some period t. We let $\Pi(m, m_p)$ be the principal's continuation payoff if (m, m_p) realizes. Given agent t's message m, the principal always chooses m_p to maximize $\Pi(m, m_p)$. We let $\Pi(m) := \max_{m_p} \Pi(m, m_p)$, so $\Pi(m)$ is the principal's continuation payoff after agent t's message m. We let $\overline{\Pi}$ and $\underline{\Pi}$ be the highest and

[21] That is, unlike its role for agents, commitment forces the principal to send messages that are *ex post* suboptimal. Hence, allowing the principal to commit does *not* refine the equilibrium set of the game without commitment.

lowest continuation payoffs that agent t's message can induce. Then, incentive constraints are identical to the extortion game (i.e., Proposition 2). The principal's message does not mitigate extortion at all, so our impossibility result still holds.

Proposition 12 *Suppose that in each period t the principal sends $m_p \in M_p$ after agent t sends m. The principal-optimal equilibrium is outcome-equivalent to that in Proposition 2.*

The principal talks before agent t. Consider some period t. Define $\Pi(m_p, m)$ as the principal's continuation payoff if $m_t = m$ and $m_{p,t} = m_p$. Once the principal chooses s_t, she knows $m_t = \mu_t(s_t)$. The principal therefore chooses $m_{p,t}$ to maximize her continuation payoff given agent t's message.[22] The same argument as in the previous case applies, so every equilibrium involves zero effort in each period.

B. The Principal Can Make Threats

In this appendix, we modify the extortion game by allowing the principal to choose a threat at the same time as each agent. We first show that Proposition 1 holds in this game, which means that allowing the principal to commit to messages as a function of transfers eliminates extortion. We then give two reasons why this result should be treated with skepticism.

Formally, suppose that in each $t \geq 0$, the principal chooses a threat $\nu_t : \mathbb{R} \to M$ at the same time that agent t chooses e_t and μ_t. At the end of t, message $m_t^P = \nu_t(s_t)$ is realized and publicly observed (along with agent t's message m_t). We can adapt the proof of Proposition 1 to show that the principal can earn no more than $e^* - c(e^*)$ in this game, where e^* is defined as in Proposition 1. It suffices to construct an equilibrium in which she earns that payoff.

Consider the following strategy profile. Play starts in the cooperation phase. In this

[22]This intuition would not change if agents could commit to a mixture over M, in which case the principal would choose $m_{p,t}$ to maximize her continuation payoff given the mixture. The key is that agent t can use her message to implement the same punishment regardless of whether he works or shirks.

phase,

$$\nu_t(s_t) = \mu_t(s_t) = \begin{cases} C & s_t \geq c(e^*) \\ D & \text{otherwise} \end{cases}$$

and $e_t = e^*$. If neither player deviates, then $s_t = c(e^*)$; if only agent t deviates, then $s_t = 0$; if the principal or both players deviate, then the principal best-responds given the threats. The game stays in the cooperative phase if $m_t = m_t^P = C$. Otherwise, it switches to the punishment phase with probability $\gamma \in [0,1]$. In the punishment phase, agents exert no effort and the principal pays no transfers.

Choosing γ to solve

$$c(e^*) = \frac{\delta}{1-\delta}\gamma\left(e^* - c(e^*)\right) \tag{21}$$

implies that the principal is willing to pay $s_t = c(e^*)$ on the equilibrium path. If agent t deviates, then the principal's continuation payoff cannot exceed $e^* - c(e^*)$ if she pays $s_t = c(e^*)$ and equals $(1-\gamma)(e^* - c(e^*))$ if she pays any other amount. Condition (21) implies that she is willing to pay $s_t = 0$ in that case. Agent t therefore has no profitable deviation from e_t or μ_t. The principal has no profitable deviation from ν_t, since given μ_t, she earns no more than $e^* - c(e^*)$ for paying $s_t = c(e^*)$ and no more than $(1-\gamma)(e^* - c(e^*))$ for paying any other amount. This strategy profile is therefore an equilibrium. It is principal-optimal because it maximizes total equilibrium surplus and gives all of that surplus to the principal.

This argument shows that allowing the principal to commit to a threat eliminates extortion. Essentially, the principal's and each agent's threats can be used to "cross-check" one another. If the principal is punished whenever messages disagree, then agents cannot extort any *smaller* amount than the amount that the principal pays a hard-working agent on-path. As in the proof of Proposition 4, the principal can then be made indifferent between paying $s_t = c(e^*)$ and $s_t = 0$, so that she is willing to pay a hard-working agent but not one that shirks.

While allowing the principal to commit to a threat can in principle restore cooperation, this result should be treated with skepticism for two reasons. First, while agents are indifferent across messages, the principal is not. Indeed, Appendix A. shows that she has a strict incentive to send the message that maximizes her continuation payoff. Commitment therefore forces the principal to send messages that she strictly prefers not to send, which stands in contrast to the agents, for whom commitment simply breaks indifference across messages. Consequently, we cannot treat the principal's threat as an equilibrium refinement; no analogue to Proposition 11 exists for the game with principal commitment.

Second, as Appendix A. illustrates, this result requires the principal to choose ν_t *(weakly) before* agent t chooses μ_t and e_t. If agent t chooses μ_t first, then he can shirk and extort the principal, in which case her unique best-response is to pay that agent and then send a message that guarantees a high continuation payoff. The conclusion that principal commitment eliminates extortion therefore depends on a particular assumption about *the sequence in which* each player makes threats.

C. Each Agent has Multiple Extortion Opportunities

This section considers equilibria if agents have multiple opportunities to make threats in the extortion game. Once the principal gives in to an extortion attempt, an agent has every incentive to repeat the same threat in the hope of extracting yet more money. In equilibrium, the principal should anticipate that each payment might not be the final one. What is the effect on equilibrium cooperation?

We introduce the **extortion game with repeated threats** to address this question. In each period $t \in \{0, 1, ...\}$, the principal and agent t play the following stage game:

1. Agent t chooses $e_t \in \mathbb{R}_+$.

2. The following payment subgame is played repeatedly. At the end of each repetition,

the stage game moves to the next stage with probability $\rho \in (0,1)$, and otherwise the payment subgame repeats. In repetition $k \in \{1, 2, ...\}$ of the payment subgame:

(a) Agent t chooses $\mu_t^k : \mathbb{R} \to \mathcal{M}$;

(b) The principal chooses a transfer $s_t^k \in \mathbb{R}_+$.

3. Let $K \in \mathbb{R}$ be the final iteration of the payment subgame. Then $s_t = \sum_{k=1}^{K} s_t^k$ and $m_t = \mu_t^K(s_t^K)$.

The principal and agent t's payoffs are identical to the extortion game. In particular, the principal does not discount between iterations of the payment subgame.

This model assumes that agents can threaten the principal an unknown number of times, and that only the message associated with the final threat is observed by other agents. If each agent could threaten the principal a known, finite number of times, then only their final threats would matter in equilibrium, so the analysis from the baseline extortion model would apply.

We show that our results from the extortion game hold even if each agent can make an uncertain number of threats, provided that the probability of being able to make one additional threat, $1 - \rho$, is not too large. To prove this result, we show that the amount that an agent can extort is his leverage, regardless of his effort. Hence, the equilibrium effort is constantly zero.

Proposition 13 *Consider an equilibrium of the payment subgame, and let* $\overline{\Pi}$ *and* $\underline{\Pi}$ *equal the principal's highest and lowest continuation payoffs, respectively, with corresponding messages* $\overline{m}$ *and* $\underline{m}$*. If* $\rho > \frac{1}{2}$*, then*

$$\mathbb{E}[s_t] = \frac{\delta}{1-\delta}\left(\overline{\Pi} - \underline{\Pi}\right)$$

in any equilibrium. Hence, $e_t = 0$ *for any* t *in any equilibrium.*

Proof of Proposition 13

Let U_M and U_m be the agent's largest and smallest equilibrium payoffs in the payment subgame. Our goal is to show that for any $\rho > \frac{1}{2}$, $U_M = U_m = \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}) \equiv L$.

The following equilibrium strategy gives agent t a payoff of L: in each k,

$$\mu_t^k = \begin{cases} \overline{m} & s_t^k \geq \rho L \\ \underline{m} & \text{otherwise.} \end{cases}$$

Facing this threat, the principal is indeed willing to pay $s_t^k = \rho L$, since with probability ρ this message is the final message. The probability of the payment subgame surviving to iteration k equals $(1-\rho)^{k-1}$, so this strategy profile gives the agent an expected payoff

$$\rho L \sum_{k=1}^{\infty} (1-\rho)^{k-1} = L.$$

The principal's expected payoff equals $\frac{\delta}{1-\delta}\underline{\Pi}$. Agent t has no profitable deviation from μ_t^k, since the principal would be unwilling to pay any amount larger than ρL. Thus, this strategy is an equilibrium, so $U_M \geq L$. Moreover, $U_M \leq L$, because the principal cannot earn a payoff lower than $\frac{\delta}{1-\delta}\underline{\Pi}$ in equilibrium, and total surplus cannot exceed $\frac{\delta}{1-\delta}\overline{\Pi}$.

Now, we bound U_m from below. The principal's minimum equilibrium payoff equals $\frac{\delta}{1-\delta}\underline{\Pi}$; let $\frac{\delta}{1-\delta}\Pi_M$ equal her maximum equilibrium payoff. Then, the principal's **unique** best response to

$$\mu_t^k = \begin{cases} \overline{m} & s_t^k \geq s \\ \underline{m} & \text{otherwise} \end{cases} \tag{22}$$

equals $s_t^k = s$, so long as

$$-s + \rho \frac{\delta}{1-\delta}\overline{\Pi} + (1-\rho)\frac{\delta}{1-\delta}\underline{\Pi} > \rho \frac{\delta}{1-\delta}\underline{\Pi} + (1-\rho)\frac{\delta}{1-\delta}\Pi_M,$$

or

$$s < (1-2\rho)\frac{\delta}{1-\delta}\underline{\Pi} + \rho\frac{\delta}{1-\delta}\overline{\Pi} - (1-\rho)\frac{\delta}{1-\delta}\Pi_M.$$

Let s_M equal the *supremum* transfer that satisfies this constraint, with $s_M = 0$ if no transfer does. Then,

$$\frac{\delta}{1-\delta}\Pi_M \leq \frac{\delta}{1-\delta}\overline{\Pi} - s_M \sum_{k=1}^{\infty}(1-\rho)^{k-1},$$

since if agent t earns less than $s_M \sum_{k=1}^{\infty}(1-\rho)^{k-1} = \frac{s_M}{\rho}$, he can profitably deviate to (22) in each k, with $s = s_M - \epsilon$ for $\epsilon > 0$ arbitrarily small.

By definition of s_M, we must have

$$s_M \geq \max\left\{0, (1-2\rho)\frac{\delta}{1-\delta}\underline{\Pi} + \rho\frac{\delta}{1-\delta}\overline{\Pi} - (1-\rho)\left(\frac{\delta}{1-\delta}\overline{\Pi} - \frac{s_M}{\rho}\right)\right\}.$$

Simplifying, we have

$$s_M \geq \max\left\{0, (2\rho-1)\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}) + \frac{1-\rho}{\rho}s_M\right\}.$$

For $\rho > \frac{1}{2}$, the right-hand side of this equality is strictly positive. In that case, we can gather terms to yield

$$\frac{2\rho-1}{\rho}s_M \geq (2\rho-1)\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}).$$

Cancelling $2\rho - 1$ from both sides of this expression yields

$$s_M \geq \rho L,$$

in which case $U_m \geq \rho L \sum_{k=1}^{\infty}(1-\rho)^{k-1} = L$.

We conclude that if $\rho > \frac{1}{2}$, agent t's unique equilibrium payoff equals L, so $\mathbb{E}[s_t] = L$, as desired.■

What happens if $\rho < \frac{1}{2}$?

The payment subgame resembles a repeated game, where the probability of continuing to another iteration, $1 - \rho$, corresponds to the discount factor. This subgame is also positive-sum; feasible total surplus can be as low as $\frac{\delta}{1-\delta}\underline{\Pi}$ or as high as $\frac{\delta}{1-\delta}\overline{\Pi}$. Consequently, for $\rho < \frac{1}{2}$, we can use repeated-game incentives to deter agent t from extorting. One way to construct these incentives is familiar from Section V.: the principal is punished after she gives in to an extortion attempt. The principal therefore refuses to pay anything following a deviation, so the agent refrains from extortion.

D Supplemental Appendix: Long-run Agents

This appendix studies coordinated punishments with long-run agents. Appendix A. presents the model, which builds on Section V.. Appendix B. shows that we can use the repeated interactions between the principal and each agent to deter extortion and facilitate coordinated punishments. In general, however, the possibility of misuse remains a serious problem; Appendix C. shows that equilibrium effort falls to 0 as the number of agents increases.

A. A Model with Long-Run Agents

Consider a repeated game with a single principal and N agents with a shared discount factor $\delta \in [0, 1)$. In each period, the following stage game is played:

1. One agent $x_t \in \{1, ..., N\}$ is chosen uniformly at random as the active agent.

2. The active agent chooses $e_t \in [0, \bar{e}]$ and $\mu_t : \mathbb{R} \to M$, which are observed by the principal but not by other agents.

3. The principal and the active agent exchange transfers $s_t \in [0, \bar{s}]$ and $s_t^A \in [0, \bar{s}]$, respectively, with resulting net transfer to the active agent $s_t^n = s_t - s_t^A \in \mathbb{R}$. These transfers are observed only by the principal and the active agent.

4. The message $m_t = \mu_t(s_t)$ is realized and publicly observed.

The principal's and agent i's payoffs in each period t are $\pi_t = e_t - s_t^n$ and

$$u_{i,t} = \begin{cases} 0 & x_t \neq i \\ s_t^n - c(e_t) & x_t = i \end{cases},$$

respectively, with corresponding expected discounted payoffs $\Pi_t = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\pi_t$ and $U_{i,t} = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)u_{i,t}$. Our solution concept is weak Perfect Bayesian Equilibrium.

In this formulation of the game, μ_t determines the period-t message but not agent x_t's future actions. Consequently, and similarly to Section V., agent x_t's future interactions with the principal can be used to deter extortion. If we instead assumed that threats also determine agent x_t's future actions—or, in the bargaining re-interpretation from Section A., that players bargained over the entire continuation play—then long-run relationships would be a less plausible remedy.

B. Long-Run Relationships Deter Extortion ($N = 2$)

This section illustrates how long-run relationships can deter misuse. We make this point for the case with two agents, using an equilibrium construction that is conceptually similar to the one used to prove Proposition 5. In this construction, each bilateral relationship is used to punish the principal if she either reneges on paying a hard-working agent *or* pays an agent who has shirked. For two agents, this construction is enough to attain the no-extortion benchmark effort.

Proposition 14 *Suppose $N = 2$. The principal's maximum equilibrium payoff is*

$$\begin{aligned} \Pi^* \quad &\equiv \quad \max_e \{e - c(e)\} \\ s.t. \quad &\quad c(e) \leq \delta e \end{aligned} \tag{23}$$

The constraint $c(e) \leq \delta e$ is a necessary condition in the no-extortion game. Thus, Π^* equals the maximum equilibrium payoff that could be attained without extortion. Thus, with two agents, long-lived relationships eliminate the inefficiency associated with misuse.

To see the proof of this result, note that when δ is high, each relationship can be treated separately and still achieve first-best effort, so coordinated punishments are not needed and misuse does not arise. For lower δ, we construct an equilibrium with coordinated punishments. On the equilibrium path, the principal pays agent i for two reasons. First, not paying leads to the breakdown of her relationship with i. Second, agent i sends a "defect" message that leads to sanctions in the *other* relationship.

To stop agents from misusing the "defect" message, the equilibrium punishes the principal for paying an agent who shirks and extorts. Suppose that agent i shirks and then threatens to send the "defect" message unless he is paid. If the principal refuses to pay him, then agent i exerts effort in the future. If the principal pays him, then their relationship breaks down, and moreover, agent i attempts to extort the principal in future interactions. We show that this punishment is severe enough to stop the principal from giving in to extortion. Similar to Section V., this construction uses long-run relationships to punish the principal for acquiescing to extortion, which in turn deters the agents from extorting in the first place.

Proof of Proposition 14

The payoff Π^* is the highest payoff subject to the principal's dynamic enforcement constraint, so we only need to construct an equilibrium that achieves this payoff.

Let e^* be the effort that solves (23). First, suppose that $(1-\delta)c(e^*) \leq \frac{\delta}{2}(e^* - c(e^*))$, and consider the following strategy. All messages are ignored; in each period t, $e_t = e^*$, $s_t = c(e^*)$, $s_t^A = 0$, and $\mu_t = C$ if no deviation has occurred in periods with agent x_t active, and otherwise $e_t = s_t = s_t^A = 0$ and $\mu_t = C$. Agents are clearly willing to follow this strategy.

The principal is willing to follow it if

$$(1-\delta)c(e^*) \leq \frac{\delta}{2}(e^* - c(e^*)),$$

which holds by assumption. Thus, this strategy is an equilibrium that attains the desired payoff.

Now, suppose that $(1-\delta)c(e^*) > \frac{\delta}{2}(e^* - c(e^*))$, and consider the following strategy. Define

$$T_i(t) = \max\{t' < t | x_{t'} = i\}$$

be the most recent period prior to t in which $x_{t'} = i$. For each agent, there are two possible public states, labeled SB and OB. If $m_{T_i(t)} = D$ (C), then agent $-i$ is in state SB (OB). The public state SB includes two private states: S and B. Similarly, the public state OB includes two private states: O and B. The states O, S, B correspond to the meanings of "on-path", "sanction", and "breakdown." Note that all players know an agent's public state, but only the principal and agent i know i's private state.

We define the strategy in each state:

- An agent in public state SB exerts no effort and always chooses $\mu_t = C$. The principal pays such an agent $s_t = 0$.
- If agent i is in public state OB:
  - If agent i is in private state O and believes the other agent is in O or S, then he chooses $e_t = c(e^*)$ and

$$\mu_t(s_t) = \begin{cases} C & s_t \geq c(e^*) \\ D & s_t < c(e^*) \end{cases}.$$

  - If agent i is in private state O and believes the other agent is in B with positive probability, then he chooses $e_t = 0$ and

$$\mu_t(s_t) = \begin{cases} C & s_t \geq c(e^*) \\ D & s_t < c(e^*) \end{cases}.$$

– If agent i is in private state B, then he chooses $e_t = 0$ and

$$\mu_t(s_t) = \begin{cases} C & s_t \geq \frac{\delta e^*}{2} \\ D & s_t < \frac{\delta e^*}{2} \end{cases}.$$

- If agent i is in public state OB, then the principal's transfers are:
  - If agent i is in O and agent $-i$ is in O or S, then whenever $x_t = i$:
    - $s_t = c(e^*)$ if agent i does not deviate in t;
    - If agent i deviates in t, then $s_t = 0$.
  - If agent i is in O and agent $-i$ is in B, then whenever $x_t = i$, the principal pays $s_t = 0$.
  - If agent i is in B and agent $-i$ is in O or S, then whenever $x_t = i$, the principal pays:
    - $s_t = \frac{\delta e^*}{2}$ if agent i does not deviate in t;
    - If agent i deviates in t, then s_t is the smallest transfer that induces $m_t = C$ so long as that transfer is less than $\frac{\delta e^*}{2}$, and otherwise $s_t = 0$.
  - If both agents are in B, then the principal pays $s_t = 0$.

Finally, we describe how the strategy transitions between states.

- Both agents start in public state OB and private state O.
- Agent $-i$'s past message determines agent i's public state. If $m_{T_{-i}(t)} = D$, then agent i is in SB; if $m_{T_{-i}(t)} = C$, then agent i is in OB.

- If agent i observes a deviation by the principal, then he transitions to B. Otherwise, agent i stays in O or S, depending on his public state. So long as agent i is in O, he assigns zero probability to the other agent being in B. Note that it is never the case that both agents are in state SB.

- An agent in state B remains in state B forever. An agent in state B initially assigns probability 0 to the other agent being in state B, and then updates according to the equilibrium strategy.

We argue that this strategy is an equilibrium. The first step in this argument is to calculate payoffs at each combination of states. If the state is (B, B) or the public state is (SB, SB), then all players earn 0. It remains to consider the states that correspond to the public states (OB, OB) or (SB, OB). Routine but tedious calculations yield the following payoffs in these states:

$(i, -i)$	U_i	U_{-i}	Π
(O, O)	0	0	$e^* - c(e^*)$
(S, O)	0	0	$\frac{1}{2-\delta}\left(e^* - c(e^*)\right)$.
(O, B)	$-\frac{1-\delta}{2-\delta}c(e^*)$	$\frac{1-\delta}{2-\delta}\frac{\delta e^*}{2}$	$\frac{(1-\delta)e^*}{2}$
(S, B)	$-\frac{(1-\delta)\delta}{(2-\delta)^2}c(e^*)$	$\frac{2(1-\delta)}{(2-\delta)^2}\frac{\delta e^*}{2}$	0

Next, we argue that players cannot profitably deviate from this strategy.

1. An agent in state S believes that his actions do not affect continuation play, so he has no profitable deviation from $s_t^A = e_t = 0$ and $\mu_t = C$. Similarly, the principal pays $s_t = 0$ to an agent in state S.

2. Consider state (O, O). If an agent deviates, then $s_t = 0$ and play either stays in (O, O) or moves to (O, S), both of which yield continuation payoffs 0. Thus, agents earn 0 both on- and off-path. For the principal:

(a) If agent i has not deviated, paying $s_t = c(e^*)$ leads to state (O, O), with payoff

$$-(1-\delta)c(e^*) + \delta(e^* - c(e^*)).$$

Any deviation to $s_t < c(e^*)$ leads to state (B, S), resulting in payoff 0. The inequality $\delta e^* \geq c(e^*)$ implies that the principal cannot profitably deviate.

(b) If agent i has deviated, then $s_t > 0$ leads to state (B, S) or (B, O), while $s_t = 0$ leads to state (O, S), or (O, O). Thus, the principal has no profitable deviation from $s_t = 0$ so long as the payoff from (O, S) is larger than the payoff from (B, O), or

$$\frac{1}{2-\delta}(e^* - c(e^*)) \geq \frac{(1-\delta)e^*}{2}.$$

This inequality is implied by $\delta e^* \geq c(e^*)$.

3. In state (O, S), we have already shown that there is no profitable deviation whenever the agent in state S is active. When the agent in state O is active, the argument is identical to the case of (O, O).

4. In state (O, B), for the agent in state B, $\frac{\delta e^*}{2}$ is the maximum that the principal is willing to pay to stay in (O, B) rather than move to (S, B). Thus, this agent has no profitable deviation, nor does the principal have a profitable deviation when this agent is active. The agent in state O believes that the state is (O, O) with probability 1, so the argument from state (O, O) applies. When the agent in state O is active, the principal is willing to choose $s_t = 0$, because

$$(1-\delta)c(e^*) > \frac{\delta}{2}(e^* - c(e^*)) \geq \frac{\delta}{2}(1-\delta)e^*,$$

where the first inequality holds by assumption, and the second inequality holds because $c(e^*) \leq \delta e^*$. Thus, the principal cannot profitably deviate when the active agent is in state O.

5. In state (S, B), the argument is the same as in (O, B) for the agent in state B.

We conclude that players cannot profitably deviate from this strategy, which is therefore an equilibrium that delivers the desired payoff. ■

C. Extortion Remains a Problem ($N \to \infty$)

This section shows that long-lived relationships are not always enough to eliminate misuse. We make this point by proving a limiting result: as the number of agents grows without bound, the maximum equilibrium effort shrinks to zero. Here, more agents imply that each agent has a "weaker" relationship with the principal, so this result corresponds to the implication of Proposition 5 that effort shrinks to zero as $v_H - v_L$ does.

Proposition 15 *In the game with N long-lived agents, let e_N equal the maximum effort attained at any on-path history of any equilibrium. Then,*

$$\lim_{N \to \infty} e_N = 0.$$

Proof of Proposition 15

Fix N, and consider an on-path history in period t. Let $x_t = i$, and suppose e_t is the equilibrium effort. Define U^* as agent i's expected on-path continuation payoff given his private history, and note that agent i believes that his continuation payoff is at least 0 following a deviation. Hence, e_t satisfies

$$s_t - c(e_t) + \frac{\delta}{1-\delta} U^* \geq \hat{s}_t, \tag{24}$$

where $\hat{s}_t$ is the supremum of the set of transfers for which there exists some μ_t such that the principal's unique best response to μ_t is to pay $\hat{s}_t$.

We can bound s_t from above and $\hat{s}_t$ from below. Define $\overline{\Pi}^*$ as the principal's on-path continuation payoff, with corresponding message $m_t = C$. Define $\underline{\Pi}^P$ as her *worst* continuation payoff, with corresponding message $m_t = D$. Then, the principal earns no less than $\underline{\Pi}^P$ from deviating to $s_t = 0$, so

$$s_t \leq \frac{\delta}{1-\delta}(\overline{\Pi}^* - \underline{\Pi}^P). \tag{25}$$

To bound $\hat{s}_t$ from below, define $\underline{\Pi}^*$ as the principal's *best* continuation payoff following message $m_t = D$. Define $\overline{\Pi}^P$ as her *worst* continuation payoff following message $m_t = C$. For any $\tilde{s} \geq 0$, agent x_t can deviate by choosing $e_t = 0$ and

$$m_t = \begin{cases} D & s_t < \tilde{s} \\ C & s_t \geq \tilde{s} \end{cases}.$$

The principal's unique best response to this deviation is to pay $\tilde{s}$, provided that

$$-\tilde{s} + \frac{\delta}{1-\delta}\overline{\Pi}^P > \frac{\delta}{1-\delta}\underline{\Pi}^*.$$

Therefore, a lower bound on $\hat{s}$ is

$$\hat{s} \geq \frac{\delta}{1-\delta}\left(\overline{\Pi}^P - \underline{\Pi}^*.\right) \tag{26}$$

Applying (25) and (26) to (24), we derive the following necessary condition for e_t:

$$\frac{\delta}{1-\delta}\left(\overline{\Pi}^* + U^* - \underline{\Pi}^P\right) - c(e_t) \geq \frac{\delta}{1-\delta}\left(\overline{\Pi}^P - \underline{\Pi}^*\right),$$

or

$$c(e_t) \leq \frac{\delta}{1-\delta}\left(\overline{\Pi}^* - \overline{\Pi}^P + U^* + \underline{\Pi}^* - \underline{\Pi}^P\right). \tag{27}$$

Our goal is to show that the right-hand side of (27) converges to 0 as $N \to \infty$. Note that

$$U^* \leq \frac{1}{N}\bar{s},$$

since agent i earns a stage-game payoff of 0 whenever $x_t \neq i$. Therefore, $U^* \to 0$ as $N \to \infty$.

Consider $\overline{\Pi}^* - \overline{\Pi}^P$. Define $\bar{h}^*$ as the history at the end of period $t+1$ that leads to continuation payoff $\overline{\Pi}^*$, and similarly define $\bar{h}^P$ as the history corresponding to $\overline{\Pi}^P$. Since $m_t = C$ in both of these histories, all agents $j \neq i$ cannot distinguish between $\bar{h}^*$ and $\bar{h}^P$. Therefore, suppose the principal plays the following strategy at $\bar{h}^P$: in every $t' > t$,

1. If agent x_t has not been the active agent since period t, then play the same actions as in the corresponding history following $\bar{h}^*$.

2. If agent x_t has been active in some period following t (including the current period), then play $s_t = 0$.

This strategy is feasible, so the principal's payoff from it bounds $\overline{\Pi}^P$ from below.

Denoting an arbitrary history at the start of period t by h^t, define

$$\mathcal{B}^{t'} = \left\{ h^{t'} | h^{t'} \text{ follows } \bar{h}^P \text{ and there exists no } \hat{t} \in (t, t'] \text{ such that } x_{\hat{t}} = x_t \right\}.$$

That is, $\mathcal{B}^{t'}$ is the set of period-t' histories such that agents $j \neq i$ cannot distinguish $\bar{h}^P$ and $\bar{h}^*$. If the principal plays the strategy specified above after $\bar{h}^P$, then she earns the same payoff that she would earn following $\bar{h}^*$ at every history in $\mathcal{B}^{t'}$. Since the principal earns no less than 0 at any history,

$$\overline{\Pi}^P \geq \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\mathbb{E}\left[\pi_{t'} | \mathcal{B}^{t'}\right] \Pr\{\mathcal{B}^{t'}\}.$$

Denoting the complement of $\mathcal{B}^{t'}$ by $\neg\mathcal{B}^{t'}$, we can write

$$\overline{\Pi}^* = \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\mathbb{E}\left[\pi_{t'}|\mathcal{B}^{t'}\right]\Pr\{\mathcal{B}^{t'}\} + \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\mathbb{E}\left[\pi_{t'}|\neg\mathcal{B}^{t'}\right]\Pr\{\neg\mathcal{B}^{t'}\},$$

so that

$$\overline{\Pi}^* - \overline{\Pi}^P \leq \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\mathbb{E}\left[\pi_{t'}|\neg\mathcal{B}^{t'}\right]\Pr\{\neg\mathcal{B}^{t'}\}.$$

Noting that $\Pr\{\neg\mathcal{B}^{t'}\} = \left(1 - \left(\frac{N-1}{N}\right)^{t'-t}\right)$ and $\pi_{t'} \leq \bar{e} + \bar{s}$, we conclude that

$$\overline{\Pi}^* - \overline{\Pi}^P \leq \sum_{t'=t}^{\infty} \delta^{t'-t}(1-\delta)\left(1 - \left(\frac{N-1}{N}\right)^{t'-t}\right)(\bar{e}+\bar{s}).$$

The right-hand side of this expression converges to 0 as $N \to \infty$. Thus, $\lim_{N\to\infty}\left(\overline{\Pi}^* - \overline{\Pi}^P\right) = 0$.

We can make a very similar argument for the difference $\underline{\Pi}^* - \underline{\Pi}^P$, since $m_t = D$ in both of the histories leading to $\underline{\Pi}^*$ and $\underline{\Pi}^P$. Therefore, at the history leading to payoff $\underline{\Pi}^P$, the principal can play as in the history $\underline{\Pi}^*$ until agent x_t is again the active agent, and thereafter play $s_t = 0$. As $N \to \infty$, this strategy leads to an expected payoff that is arbitrarily close to $\underline{\Pi}^*$. So $\lim_{N\to\infty}\left(\underline{\Pi}^* - \underline{\Pi}^P\right) = 0$.

Putting the above arguments together, we conclude that

$$\lim_{N\to\infty} e_N \leq \lim_{N\to\infty}\left(\overline{\Pi}^* - \overline{\Pi}^P + U^* + \underline{\Pi}^* - \underline{\Pi}^P\right) = 0,$$

as desired. ■