The Use and Misuse of Coordinated Punishments

Daniel Barron and Yingni Guo*

July 29, 2020

Abstract

Communication facilitates cooperation by ensuring that deviators are collectively punished. We explore how players might misuse communication to threaten one another, and we identify ways that organizations can deter misuse and restore cooperation. In our model, a principal plays trust games with a sequence of short-run agents who communicate with one another. An agent can shirk and then extort pay by threatening to report that the principal deviated. We show that these threats can completely undermine cooperation. Investigations of agents' efforts, or dyadic relationships between the principal and each agent, can deter extortion and restore some cooperation. Investigations of the principal's action, on the other hand, typically do not help. Our analysis suggests that collective punishments are vulnerable to misuse unless they are designed with an eye towards discouraging it.

*Barron: Northwestern University, Kellogg School of Management, Evanston IL 60208; email: d-barron@kellogg.northwestern.edu. Guo: Northwestern University, Economics Department, Evanston IL 60208; email: yingni.guo@northwestern.edu. The authors would like to thank Nageeb Ali, Charles Angelucci, Nemanja Antic, Alessandro Bonatti, Renee Bowen, Joyee Deb, Wouter Dessein, Matthias Fahn, Benjamin Friedrich, George Georgiadis, Marina Halac, Peter Klibanov, Ilan Kremer, Nicolas Lambert, Stephan Lauermann, Jin Li, Elliot Lipnowski, Shuo Liu, Bentley MacLeod, David Miller, Joshua Mollner, Dilip Mookherjee, Arijit Mukherjee, Jacopo Perego, Michael Powell, Luis Rayo, Jonah Rockoff, Mark Satterthwaite, Andy Skrzypacz, Takuo Sugaya, Jeroen Swinkels, Joel Watson, and audiences at many conferences, workshops, and seminars. We thank the UCSD theory reading group for comments on a draft of this paper, and Andres Espitia for excellent research assistance.

1 Introduction

Productive relationships thrive on the enthusiastic cooperation of their participants. In many settings, individuals cooperate because they expect opportunistic behavior to be punished (Malcomson (2013)). Communication plays an essential role in coordinating these punishments, since it allows those who do not directly observe misbehavior to nevertheless punish the perpetrator. These coordinated punishments are central to cooperation among participants in online marketplaces (Hörner and Lambert (2018)), as well as between managers and workers (Levin (2002)), suppliers and customers (Greif et al. (1994); Bernstein (2015)), and members of communities (Ostrom (1990)).

Once armed with the power to trigger coordinated punishments, individuals face a grave temptation: they can extort concessions from their partners by threatening to *falsely* report opportunistic behavior (Gambetta (1993); Dixit (2003a, 2007)). In this paper, we explore how individuals might misuse coordinated punishments. We emphasize two overarching takeaways. First, we show that misuse is a serious vulnerability that can completely undermine cooperation. Second, we identify practical ways for organizations to restore cooperation in the face of this vulnerability.

The possibility of misuse is a serious concern for online platforms, where most interactions are short-lived and coordinated punishments are essential for ensuring cooperation. In an evocative recent example, an investigation of the lodging platform Airbnb uncovered a network of hosts who misused the platform's review system to extort guests (Conti (2019)). These hosts reneged on their obligations by altering guests' accommodations at the last minute. While Airbnb allows guests to request a refund in these circumstances, the hosts deterred refund requests by writing scathing reviews of guests who complained. By misusing Airbnb's review system in this way, hosts ensured that they received payment despite reneging on their end of the deal. The resulting scandal was serious enough to prompt action from the FBI.

This type of misuse is hardly unique to Airbnb. In the early days of eBay, buyers and

sellers could review each other based on the accuracy of product descriptions, the quality of delivery service, and the timeliness of payment. This review system led to a severe form of misuse, in which sellers would extort positive reviews from buyers by threatening to negatively review any buyer that complained. Klein et al. (2016) shows that these threats led to lower seller effort and lower buyer satisfaction. Review aggregators are similarly susceptible to misuse, as a wedding planning business learned when its reputation was sullied by a customer who posted vitriolic reviews in an attempt to extort services (Proctor (2018)). The phenomenon of extortionary reviews is widespread enough that platforms like TripAdvisor and Etsy have instituted explicit anti-extortion policies; Etsy's policy, for example, forbids a buyer from leaving "a negative review in an attempt to force the seller into providing a refund" or additional items.[1] The possibility of misuse has also spurred responses from regulatory authorities and lawmakers.[2]

To explore how coordinated punishments can be used and misused, we consider a model of a long-run principal who interacts with a sequence of short-run agents. Each agent exerts costly effort to benefit the principal, who can then choose to pay him. Agents observe only their own interactions but can communicate with one another. To capture the idea that extortion entails action-contingent threats – i.e., "pay me *or else* I will punish you" – we allow each agent to make a **threat** when he chooses his effort. This threat, which is observed by the principal but not by other agents, associates a message to each possible payment. Agents then follow through on their threats.

In this model, misuse completely undermines cooperation. The principal is willing to pay an agent only if she would otherwise be punished by future agents. Communication is therefore essential for cooperation. Once endowed with a message that triggers punishments, however, an agent can extort the principal by shirking and then threatening to send that

[1] TripAdvisor's anti-extortion policy is at https://www.tripadvisor.com/TripAdvisorInsights/w592; Etsy's policy is at https://www.etsy.com/legal/policy/extortion/239966959186. These policies note that the platform has limited ability to combat extortion that occurs outside its official messaging system.

[2] In the United Kingdom, the Competition and Markets Authority has noted that customers sometimes use the threat of negative reviews to demand discounts (Peachey (2015)). The Washington State legislature has considered a bill to deter extortionary online reviews.

message unless the principal pays him. Since this threat is enough to induce the principal to pay a hard-working agent, it is also enough to induce her to pay a shirking agent. Thus, the pay that an agent can demand is essentially independent of his effort. The stark implication of this logic is that agents do not exert any effort.

After establishing this impossibility result, we explore how organizations can deter extortion and encourage cooperation. We focus on two instruments that are available in many cooperative endeavors: *investigations*, which we model as public signals of either the agents' efforts or the principal's transfers, and *dyadic relationships*, which we model as a coordination game played by the principal and each agent.

The unifying idea of these instruments is that an agent is willing to exert effort only if doing so increases his **leverage** over the principal, which we define as the harshest punishment that the agent can trigger with a message. An agent can extort any transfer that is smaller than his leverage. If an agent's leverage is independent of his effort, as it is in any equilibrium of our baseline model, then he has no incentive to exert effort. If an agent's leverage is increasing in his effort, on the other hand, then he might exert effort in order to increase his leverage, so that he can demand higher pay. An instrument is valuable exactly when it *ties leverage to effort* in this way.

Building on this idea, we show that investigations into agents' efforts typically improve cooperation, whereas investigations into the principal's transfers typically do not. Effort signals are useful for deterring extortion, not because agents are directly rewarded or punished on the basis of these signals, but because these signals can tie leverage to effort. They do so by ensuring that harder-working agents can trigger harsher coordinated punishments. Agents are then willing to exert effort in order to obtain higher leverage and demand higher pay. In contrast, transfer signals can reveal whether the principal paid an agent but not whether that pay was *deserved* or not. Hence, such signals typically cannot tie leverage to effort and so cannot improve cooperation. The only exception is that, under stringent conditions, transfer signals can make the principal indifferent between transfers in a way

that leads to some effort. Even then, the extent of cooperation is limited by the need for occasional on-path punishments.

Next, we study how *dyadic relationships* between the principal and each agent can tie leverage to effort. Unlike online platforms, where short-lived interactions are the norm, manager-worker relationships are typically long-lived. As with short-lived relationships, institutions that coordinate punishments, such as unions, have the potential to improve cooperation and lead to exceptional productivity in long-lived relationships (Freeman and Medoff (1979); Levin (2002)). To do so, however, these institutions must first deter agents from misusing the coordinated punishments that they make possible.

We show that dyadic relationships can guard against misuse by rewarding the principal for *refusing* to pay a shirking agent. Dyadic relationships therefore complement coordinated punishments: organizations with strong dyadic relationships can implement severe coordinated punishments without opening the door to extortion, while those with weak dyadic relationships give their agents little leverage and so result in low effort. In the latter case, misuse remains a real threat to coordinated punishments, which is consistent with General Motors' experience at its plant in Fremont, California, in the 1980s. Workers at that plant misused the threat of grievances to get away with "shirking" behaviors like absenteeism and drug use during working hours, leading to low productivity that eventually resulted in the plant's closure (Glass and Langfitt (2015)).[3]

The premise of our analysis is that, while organizations can potentially benefit from coordinated punishments, they cannot perfectly control how their members use these punishments. Our model illustrates a stark version of this premise, in which misuse completely undermines cooperation. In practice, and as we explore in Section 6, we do not expect every agent in a particular context to engage in misuse. Rather, our point is that coordinated punishments are vulnerable to misuse, and that this vulnerability can seriously impair co-

[3]Another example comes from the recent Wells Fargo Scandal. During that scandal, Wells Fargo faced allegations that it punished employees who spoke up about fraudulent practices by falsely reporting them to FINRA for unethical behavior (Arnold and Smith (2016)). FINRA then shared this information with other prospective employers.

operation. Organizations ignore this vulnerability at their peril. Guarding against misuse demands a fundamentally different approach to designing incentive systems. In our setting, these systems are embedded in the rules or culture of an organization, as represented by an equilibrium of our game. Our lessons extend to other settings in which cheap-talk messages are used to motivate cooperation.

Related Literature

Our contribution is to explore how misuse undermines coordinated punishments and how organizations can combat it. Therefore, we build on the literature that studies how coordinated punishments support cooperation (Milgrom et al. (1990), Greif et al. (1994), Dixit (2003a,b)). Much of this literature has as its goal the identification of network structures or equilibrium strategies that are particularly conducive to cooperation (Lippert and Spagnolo (2011), Wolitzky (2013), Ali and Miller (2013, 2016), Ali et al. (2017)). Especially related is Ali and Miller (2016), which shows that players might not report deviations if doing so reveals that they are more willing to renege on their own promises. Extortion is a different but complementary obstacle to coordinated punishments.

Since extortion is inherently action-contingent – i.e., "pay me *or else* I will punish you" – our analysis is related to a growing literature on action-contingent threats and promises. Like us, some of these papers assume that players commit to threats in order to allow for action-contingent deviations (Wolitzky (2012), Chassang and Padro i Miquel (2018), Ortner and Chassang (2018)).[4] In our setting, we can also re-interpret commitment as an equilibrium refinement of the game without commitment, which is related to the approach taken in Zhu (2018, 2019). We contribute to this literature by studying how misuse undermines coordinated punishments and exploring new ways for organizations to deter it.

Most of the literature on cooperation focuses on the use of coordinated punishments

[4]Indeed, Ortner and Chassang (2018) have an appendix that studies extortion. However, that appendix assumes that reports lead to exogenous and fixed punishments, while the point of our analysis is to show how to optimally link messages to punishments.

rather than the potential for misuse. Dixit (2003a, 2007) are perhaps the first to formally model the misuse of coordinated punishments, albeit in a setting with centralized enforcers rather than decentralized communication. Bowen et al. (2013), which studies local adaptation in communities, considers a type of misuse that is not action-contingent. In contrast to that paper, our agents make action-contingent threats.

In our setting, an agent essentially threatens the principal with a bad "outside option" unless she pays him. Our paper is therefore connected to the literature on bargaining and renegotiation in repeated games. Particularly related are papers that allow players to bargain over surplus in equilibrium (Baker et al. (2002), Halac (2012, 2015), Miller and Watson (2013), Goldlücke and Kranz (2017), Miller et al. (2020)) and the literature on coalitional deviations (Ali and Liu (2018); Liu (2019)). By focusing on communication across agents, our paper studies a setting in which the principal's "outside option" depends on how messages affect future equilibrium play.[5]

More broadly, our framework builds on the relational contracting literature (Bull (1987), MacLeod and Malcomson (1989), Baker et al. (1994), Levin (2003)), especially those papers that study coordinated punishments (e.g., Levin (2002)). We study how extortion undermines such punishments. Recent papers have explored relational contracts in the presence of limited transfers (Fong and Li (2017), Barron et al. (2018)), asymmetric information (Halac (2012), Malcomson (2016)), or both (Li et al. (2017), Guo and Hörner (2018), Lipnowski and Ramos (2020)). We focus on a monitoring friction – agents do not observe one another's relationships – which implies that cooperation must rely on communication. Other papers that study relational contracts with bilateral monitoring, including Board (2011), Andrews and Barron (2016), and Barron and Powell (2018), do not allow agents to communicate. We complement these papers by identifying a reason why communication might be ineffective at sustaining cooperation.

[5]We formalize the connection between our model and the literatures on bargaining and coalitional deviations in Appendix B.

2 Model

Our baseline model is the following **extortion game.** A long-run principal ("she") interacts with a sequence of short-run agents (each "he"). In each period $t \in \{0, 1, 2, ...\}$, the principal and agent t play a trust game: agent t exerts effort, then the principal chooses how much to pay him. This interaction is observed only by the principal and agent t, but agent t can send a public message at the end of period t. Our key assumption is that before transfers are paid, agent t makes a **threat**, which is a mapping from the transfer he receives to the message he sends, and which is observed by the principal but not by other agents. Agent t then follows through on this threat.

Formally, the stage game in period t is:

1. Agent t chooses his effort $e_t \in \mathbb{R}_+$ and a threat $\mu_t : \mathbb{R} \to M$, where M is a large, finite message space.[6] Both e_t and μ_t are observed by the principal but not by any other agent.

2. The principal makes a transfer to agent t, $s_t \geq 0$, which is observed by agent t but not by other agents.[7]

3. The message $m_t = \mu_t(s_t)$ is realized and observed by all players.

The principal's period-t payoff and agent t's utility are $(e_t - s_t)$ and $(s_t - c(e_t))$, respectively, where $c(\cdot)$ is strictly increasing, strictly convex, and twice continuously differentiable, as well as satisfying $c(0) = c'(0) = 0$. We assume that there exists a first-best effort, e^{FB}, such that $c'(e^{FB}) = 1$. The principal has discount factor $\delta \in [0, 1)$, with corresponding normalized discounted payoffs $\Pi_t = (1 - \delta) \sum_{t'=t}^{\infty} \delta^{t'-t}(e_{t'} - s_{t'})$. Players observe a public randomization device (notation for which is suppressed) in every step of the stage game.

[6] The assumption that M is finite simplifies the proofs (by ensuring that various maxima and minima exist) but is not essential for the results.

[7] For almost all of our results, the assumption that agents do not pay the principal is without loss. The exception is Section 5; we allow agents to pay the principal in that section.

The principal observes everything, while agents observe only their own interactions with the principal and all messages. Our solution concept is Perfect Bayesian Equilibrium.[8] Some of our results focus on principal-optimal equilibria, which maximize the principal's *ex ante* expected payoff among all equilibria.

In the context of Airbnb, the agents are hosts who exert effort (e_t) to ensure that their properties are safe, comfortable, and described accurately. The principal is a guest who rewards such efforts by treating a property well, following house rules, and not demanding an undeserved refund (s_t). To give the guest an incentive to follow through on these rewards, the platform allows hosts to review guests (m_t), where negative reviews make it harder for the guest to rent from other hosts in the future. As Conti (2019) discovered, however, some hosts took advantage of this review system to "shirk" on quality, knowing that they could use the threat of a negative review (μ_t) to force guests to nevertheless pay them. In Section 3, we show that agents have an incentive to similarly misuse coordinated punishments in the extortion game. Sections 4 and 5 build on this result to explore how organizations can deter misuse.[9]

The threat, μ_t, is a transparent way to show how agents can misuse coordinated punishments, one that has precedent in the approaches taken by Dixit (2003a), Wolitzky (2012), Chassang and Padro i Miquel (2018), and Ortner and Chassang (2018). Other modeling approaches would result in a similar kind of misuse. We study several of these alternative models in Appendix B. There, we show that a similar kind of misuse arises when the principal and each agent Nash bargain over that agent's message. This bargaining model is similar in spirit to Halac (2012, 2015), Miller and Watson (2013), and Miller et al. (2020) and is related to the literature on coalitional deviations (Ali and Liu (2018), Liu (2019)). We also prove that we can re-interpret commitment to μ_t as the result of agents having preferences

[8]See the definition of plain PBE in Watson (2017).

[9]In some of our examples, e_t and s_t have slightly different interpretations than effort and transfer. In particular, e_t sometimes includes a transfer paid by agent t, while s_t sometimes includes a productive action taken by the principal. The model can be generalized to account for this interpretation. In particular, we could make e_t a payment and s_t a productive action, or even make both e_t and s_t productive, without changing our argument for why misuse undermines cooperation.

for either reciprocity (Fehr et al. (2020)) or keeping one's word (Vanberg (2008)), or as an equilibrium refinement similar to Dewatripont (1987), Tranaes (1998), and Zhu (2018, 2019). In each of these alternative models, the main lessons from our analysis hold: coordinated punishments are vulnerable to misuse, and tying an agent's leverage to his effort deters misuse.

Appendix C considers alternative communication structures, including models in which the principal can send messages or commit to threats, as well as ones in which agents can make repeated threats. In most of these variants, extortion continues to undermine cooperation. We also identify particular communication structures that can lead to cooperation in equilibrium, although these positive results typically come with substantial caveats.

We occasionally compare our results to a benchmark without extortion. Define the **no-extortion game** as identical to the extortion game, except that each agent t chooses m_t at the end of period t rather than being committed to μ_t. In the no-extortion game, agents cannot shirk and then make action-contingent threats, so they cannot misuse communication.

3 Threats Undermine Equilibrium Cooperation

This section shows how coordinated punishments are used and misused in equilibrium. We first illustrate how coordinated punishments sustain cooperation in the no-extortion game. Then, we show that misuse leads cooperation to completely unravel. This impossibility result uncovers the economics of misuse and forms the foundation for the rest of our analysis.

Cooperation requires agents to communicate with one another, since without communication an agent would have no way to punish the principal for deviating. In the no-extortion game, this type of communication is enough to sustain cooperation.

Proposition 1 *In the no-extortion game, $e_t = e^*$ and $s_t = c(e^*)$ in each $t \geq 0$ of every principal-optimal equilibrium, where e^* equals the minimum of e^{FB} and the largest e that satisfies $c(e) = \delta e$.*

Proof: We first argue that total equilibrium surplus is at most $e^* - c(e^*)$. By definition of e^{FB}, equilibrium surplus is at most $e^{FB} - c(e^{FB})$. If $c(e^{FB}) \leq \delta e^{FB}$, then $e^* = e^{FB}$ and the result follows. If $c(e^{FB}) > \delta e^{FB}$, then let $\overline{\Pi}$ be the principal's maximum *ex ante* equilibrium payoff. In any period $t \geq 0$ of any equilibrium, $(1-\delta)s_t \leq \delta\overline{\Pi}$ and $s_t - c(e_t) \geq 0$ must hold, since otherwise the principal or agent t could profitably deviate from s_t or e_t, respectively. Therefore, $(1-\delta)c(e_t) \leq \delta\overline{\Pi}$. Let $\bar{e}$ be the effort that maximizes $e - c(e)$ among those efforts that are attained in any period of any equilibrium. Then, $(1-\delta)c(\bar{e}) \leq \delta\overline{\Pi} \leq \delta(\bar{e} - c(\bar{e}))$ and so $c(\bar{e}) \leq \delta\bar{e}$. We conclude that $\bar{e} \leq e^* < e^{FB}$, so equilibrium surplus is at most $e^* - c(e^*)$.

Consider the following strategy profile for each period $t \geq 0$: if $m_{t'} = C$ in all $t' < t$, then agent t chooses $e_t = e^*$; the principal chooses $s_t = c(e^*)$ if $e_t = e^*$ and $s_t = 0$ otherwise; and agent t chooses $m_t = C$ if neither player deviates and $m_t = D$ otherwise. If $m_{t'} \neq C$ in at least one $t' < t$, then $e_t = s_t = 0$ and $m_t = D$.

Once $m_{t'} \neq C$ in some $t' < t$, this strategy profile specifies the stage-game equilibrium and so players cannot profitably deviate. If $m_{t'} = C$ in all $t' < t$, then agent t has no profitable deviation because he earns 0 on-path and no more than 0 from deviating. The principal has no profitable deviation because $(1-\delta)s_t \leq \delta(e^* - c(e^*))$ is implied by $c(e^*) \leq \delta e^*$. This strategy is therefore an equilibrium. It is principal-optimal because it generates total surplus $e^* - c(e^*)$, which is the maximum equilibrium surplus, and it holds agents at their min-max payoffs. Moreover, every principal-optimal equilibrium gives the principal a payoff of $e^* - c(e^*)$ and so must entail $e_t = e^*$ in every period. ■

The proof of Proposition 1 relies on the following equilibrium construction. On the equilibrium path, each agent sends the message C if the principal pays him and D otherwise. Future agents min-max the principal if they observe the message D. Off the equilibrium path, a shirking agent sends a message that is independent of the principal's transfer, so the principal pays him nothing. The principal would rather pay a hard-working agent a transfer than be punished, and each agent would rather exert effort than shirk and forgo the transfer, so this construction can motivate effort.

Proposition 1 summarizes a core idea from much of the literature on coordinated punishments: the principal pays a hard-working agent because that agent would otherwise send a message that triggers future punishments. Implicit in this construction, and in much of the literature on coordinated punishments, is the requirement that shirking agents do not make similar threats, so that the principal *refrains* from paying a shirking agent. As our introduction makes clear, actual behavior does not always conform to this requirement. For instance, some Airbnb hosts shirk ($e_t = 0$) and then threaten to leave negative feedback ($m_t = D$) unless they are paid ($s_t > 0$).

The extortion game allows shirking agents to make exactly this type of threat. Our next result, which serves as the foundation for our analysis, shows that these threats destroy cooperation.

Proposition 2 *In the extortion game, every equilibrium entails* $e_t = s_t = 0$ *in every* $t \geq 0$.

Proof: Fix a history of messages, $m^{t-1} = (m_0, m_1, ..., m_{t-1})$, and let

$$\overline{\Pi} = \max_{m \in M} \left\{ \mathbb{E} \left[\Pi_{t+1} | m^{t-1}, m_t = m \right] \right\}$$

be the principal's maximum continuation surplus that can be induced by some message, which we denote $m_t = C$. Let $\underline{\Pi}$ be the similarly-defined minimum continuation payoff, with corresponding message $m_t = D$.

Suppose that agent t chooses some $e_t > 0$. He is willing to do so only if $s_t \geq c(e_t)$; the principal is willing to pay s_t only if

$$-(1-\delta)s_t + \delta\overline{\Pi} \geq \delta\underline{\Pi}. \tag{1}$$

For small $\epsilon > 0$, consider the following deviation by agent t. He chooses zero effort and

makes the threat:

$$\mu_t(s) = \begin{cases} C & s = s_t - \epsilon \\ D & \text{otherwise.} \end{cases} \tag{2}$$

Since (1) holds weakly at s_t, it holds strictly for $s_t - \epsilon$ and so the principal's unique best response to this deviation is to pay $s_t - \epsilon$. Agent t's payoff from this deviation is therefore $s_t - \epsilon$, which is strictly larger than $s_t - c(e_t)$ for sufficiently small ϵ. Hence, agent t can profitably deviate from any $e_t > 0$. Every equilibrium therefore has $e_t = 0$ for all $t \geq 0$, in which case $\overline{\Pi} = \underline{\Pi} = 0$ and so $s_t = 0$. ■

Whenever $s_t > 0$ on the equilibrium path, agent t can shirk and threaten to send a message that punishes the principal unless she pays him *slightly less* than s_t. Since the principal is willing to pay s_t to avoid this punishment, she strictly prefers to pay a smaller amount. Agent t can therefore shirk and still guarantee nearly the same transfer as if he had exerted effort. This deviation is so tempting that no agent will work.[10]

Before moving on, we reflect on what Proposition 2 reveals about the economics of misuse. Any equilibrium specifies a mapping from agent t's messages to the principal's continuation payoffs. Let $\overline{\Pi}$ and $\underline{\Pi}$ be, respectively, the largest and smallest continuation payoffs in the image of this mapping. Agent t's gain from extortion depends on his **leverage** over the principal, defined as the normalized difference between these continuation payoffs,

$$L \equiv \frac{\delta}{1-\delta}\left(\overline{\Pi} - \underline{\Pi}\right). \tag{3}$$

In the no-extortion game, the principal pays $s_t = 0$ following any deviation and pays some $s_t \leq L$ on the equilibrium path. Increasing agent t's leverage therefore unambiguously increases the scope for cooperation. In the extortion game, on the other hand, agent t is paid $s_t \approx L$ regardless of his effort. He therefore exerts zero effort, because his leverage, L,

[10] Proposition 2 would continue to hold even if the principal was protected by limited liability, which would impose the constraint that $s_t \leq e_t$ in each $t \geq 0$. Proving this result requires a slightly modified argument. In particular, one can show that unless $s_t = e_t$, agent t can profitably decrease his effort and make the threat (2). Thus, the principal's continuation payoff equals 0, which means that $s_t = 0$ and so $e_t = 0$ in every $t \geq 0$.

is independent of his effort.

This argument suggests that agent t *would* have the incentive to exert effort if doing so increased his leverage and hence the pay that he could demand. The next two sections explore this idea: to deter extortion, tie leverage to effort. As we will show, tying leverage to effort requires a fundamentally different approach to designing coordinated punishments.

4 Investigations

This section considers public signals of efforts or transfers. In the no-extortion game, such signals would be irrelevant; transfer signals would be redundant with the agents' messages, while effort signals would be redundant with what the principal, who is the only player that can directly punish shirking, already observes.[11]

In contrast, these signals *do* have the potential to deter misuse in the extortion game. We first show that effort signals can tie an agent's leverage to his effort, which can induce effort in equilibrium. However, deterring extortion in this way requires agents to earn rent, creating a tension between the surplus created in equilibrium and the surplus captured by the principal. Then, we show that transfer signals usually cannot tie leverage to effort. Therefore, transfer signals improve cooperation only under stringent conditions.

In the context of Airbnb, our analysis suggests that Airbnb should investigate the actions of hosts, rather than just those of guests. As we will show, negative reviews by a host should optimally trigger harsher punishments when that investigation suggests that he has exerted more effort. Tying a host's leverage to his effort in this way leads to better outcomes for both hosts and guests.[12] Moreover, we should observe hosts exerting more effort in settings where doing so improves their ability to trigger coordinated punishments.

[11] Formally, the effort level in Proposition 1 is the highest attainable effort even if we drop all equilibrium constraints except for the principal's dynamic enforcement constraint and the agents' participation constraints. Those two sets of constraints would be unaffected by signals.

[12] A practical caveat: this monitoring must be made immune to manipulation by the guest, since she has the incentive to fabricate evidence of shirking in order to ensure that the host's review is ignored.

4.1 Effort Investigations

The **extortion game with effort signals** is similar to the baseline extortion game, except that an effort-dependent signal, y_t, is publicly observed after s_t. We focus on a simple, binary signal structure: $y_t \in \{0,1\}$ with $\Pr\{y_t = 1|e_t\} = \gamma(e_t)$ for $\gamma(\cdot)$ strictly increasing and twice continuously differentiable. Agent t's threat can be any mapping from his pay *and* this signal to a message, so that (with an abuse of notation) $\mu_t : \mathbb{R}^2 \to M$ and $m_t = \mu_t(s_t, y_t)$. Payoffs are the same as in the extortion game.

The signal y_t can deter extortion by making an agent's expected leverage an increasing function of his effort. Because signals are noisy, however, a shirking agent typically retains some leverage and, hence, can extort some pay. Agents therefore refrain from extortion only if they earn an equilibrium rent. We show how the tension between total surplus and the agents' rents determines effort in a principal-optimal equilibrium.[13]

Proposition 3 *Consider an equilibrium of the game with effort signals. If $e_t = e$ on the equilibrium path, then agent t's equilibrium payoff is at least $\bar{u}(e)$, where*

$$\bar{u}(e) \equiv \max\left\{0, \frac{c'(e)}{\gamma'(e)}\gamma(e) - c(e)\right\}.$$

Suppose $\gamma(\cdot)$ is weakly concave. Then, $\bar{u}(\cdot)$ is strictly increasing, and in any $t \geq 0$ of any principal-optimal equilibrium, on-path effort solves

$$e_t \in \arg\max_e \{e - c(e) - \bar{u}(e)\}$$

subject to the constraint

$$\frac{c'(e)}{\gamma'(e)} \leq \frac{\delta}{1-\delta}(e - c(e) - \bar{u}(e)). \tag{4}$$

[13] Effort signals lead to higher effort by restoring a degree of *pairwise identifiability* to the signaling structure (see, e.g., Fudenberg et al. 1994). In particular, effort signals allow future agents to statistically distinguish between (i) the principal deviating in s_t, and (ii) an agent deviating in e_t and μ_t. Deviating in μ_t is profitable only if an agent also deviates in e_t, so effort signals are enough to restore at least some cooperation.

Proof: See Appendix A.

To prove Proposition 3, let $\overline{\Pi}(y)$ and $\underline{\Pi}(y)$ be the largest and smallest continuation payoffs induced by some message when the signal equals y. We can define an agent's leverage, $L(y)$, analogously to (3). Then, expected leverage, $\mathbb{E}\left[L(y)|e\right]$, depends on effort. As in Proposition 2, agent t can extort any transfer that is smaller than his expected leverage, so he chooses e_t to solve

$$e_t \in \arg\max_e \left\{\mathbb{E}\left[L(y)|e\right] - c(e)\right\}. \tag{5}$$

Since $L(\cdot) \geq 0$, this incentive constraint is identical to that of a static moral-hazard problem with limited liability; agent t's leverage is the analogue of the contractual payment, which can depend on y. As is typical in such models, agent t earns a rent, which equals $\bar{u}(e_t)$ for this signal structure.

As in a static moral-hazard problem with limited liability, it is optimal to set $L(0) = 0$; that is, agent t's message affects the principal's continuation payoff only if $y_t = 1$. If $\gamma(\cdot)$ is concave, then we can replace (5) with its first-order condition, $L(1) = c'(e)/\gamma'(e)$. Calculating the principal-optimal equilibrium payoff therefore reduces to maximizing total surplus minus the agent's rent, subject to the constraint that $L(1)$ cannot exceed the principal's continuation payoff. Since the principal's on-path continuation payoff optimally equals her maximum equilibrium payoff, $L(1) = c'(e)/\gamma'(e)$ must satisfy the dynamic enforcement constraint, (4).

One immediate consequence of Proposition 3 is that there exists a principal-optimal equilibrium that is stationary on the equilibrium path. A second consequence is that agent t's maximum leverage is limited by the fact that *future* agents earn rent in equilibrium. That is, the right-hand side of (4) is decreasing in $\bar{u}(\cdot)$, which implies that each agent's rent-seeking behavior imposes a negative externality on the principal's relationships with other agents.

In practice, agents might have some sway over the signal distribution, as, for instance, when a union decides how to investigate grievances. Both the principal and agents prefer

some kind of investigation to none, but they disagree on the optimal signal structure. In particular, agents would like the signal to maximize their rent, while the principal would like the signal to maximize total surplus net of that rent. Since an agent's rent in a principal-optimal equilibrium, $\bar{u}(\cdot)$, is equal to his rent from an optimal limited-liability contract, both his and the principal's preferences over signal structures are similar to those in a static contracting environment with limited liability.[14] For fixed effort e, $\bar{u}(e)$ is increasing in $\frac{\gamma(e)}{\gamma'(e)}$, so agents tend to prefer a signal distribution that puts weight on "false positives:" $y_t = 1$ occurs frequently and with a probability that is (locally) not very responsive to effort.

4.2 Transfer Investigations

We now turn to public signals of transfers. In contrast to Section 4.1, transfer signals are not a reliable remedy to extortion. The reason is that such signals reveal nothing about effort, so they usually cannot tie leverage to effort. The only exception is that certain signal distributions can be used to make the principal exactly indifferent between two different transfers when faced with an agent's optimal threat. Only under the stringent conditions that allow this indifference do equilibria with positive effort exist.

The **extortion game with transfer signals** is identical to the extortion game except that in each period $t \geq 0$, a public signal $x_t \in \mathbb{R}$ is realized after s_t and observed by everyone. Agent t's threat maps each (s_t, x_t) to a message m_t, so $\mu_t : \mathbb{R}^2 \to M$ with $\mu_t(s_t, x_t) = m_t$. We again focus on binary signals, so that $x_t \in \{0, 1\}$ with $\Pr\{x_t = 1 | s_t\} = \phi(s_t)$ for some strictly increasing and twice continuously differentiable $\phi(\cdot)$.

Our main result in this section is a set of conditions on $\phi(\cdot)$ that must hold for an equilibrium with positive effort to exist. To understand these conditions, consider play in some period t. Define $\Pi(m_t, x_t)$ as the principal's continuation payoff if agent t's message is m_t and the signal is x_t. After agent t chooses his threat μ_t, the principal chooses s_t to

[14] For an analysis of optimal signal structures in that static setting, see, for example, Starmans (2017). The sole difference between our analysis and the static contracting environment is the presence of a dynamic enforcement constraint, (4), which becomes slack as $\delta \to 1$.

maximize her payoff:

$$\max_{s} -(1-\delta)s + \delta \mathbb{E}\left[\Pi(\mu_t(s,x),x)|s\right]. \tag{6}$$

Note that (6) is independent of agent t's effort. Therefore, if a unique transfer maximizes (6), then the principal will pay that transfer regardless of agent t's effort. This leads to our first necessary condition: agent t exerts positive effort only if the principal is exactly indifferent between at least two transfers when she faces the equilibrium threat. The second necessary condition requires that no alternative threat would induce the principal to pay agent t more than his equilibrium payoff. Only under these two conditions is agent t willing to exert effort, and even then, the effort cost cannot exceed the difference between the on-path transfer and the largest amount that a shirking agent t can extort.

These two requirements imply a set of stringent necessary conditions on $\phi(\cdot)$.

Proposition 4 *Consider an equilibrium of the game with transfer signals. If $e_t > 0$ on the equilibrium path, then there exists $s^* > 0$ and $\hat{s} \in [0, s^*)$ such that (i) $c(e_t) \leqslant s^* - \hat{s}$, (ii) $\phi''(s^*) \leqslant 0$, and (iii)*

$$\phi'(s^*) = \frac{\phi(s^*) - \phi(\hat{s})}{s^* - \hat{s}}. \tag{7}$$

In particular, if $\phi(\cdot)$ is strictly concave on $\mathbb{R}_+$, then $e_t = 0$ in each $t \geq 0$ of every equilibrium.

Equation (7) combines the two conditions for $e_t > 0$ described above. First, the principal must be indifferent between paying the on-path transfer, s^*, and some other amount that is no less than $\hat{s}$, when faced with the equilibrium threat. Second, *no* threat can induce the principal to pay a transfer near s^*. The first of these conditions pins down the average slope of $\phi(\cdot)$ between $\hat{s}$ and s^*, while the second condition says that the derivative of $\phi(\cdot)$ near s^* equals the same number. Therefore, the average slope between $\hat{s}$ and s^* must equal the tangent slope at s^*, implying (7). Period-t effort must then satisfy $s^* - c(e_t) \geq \hat{s}$, since otherwise agent t could profitably shirk and extort $\hat{s}$.

Condition (7) cannot hold if $\phi(\cdot)$ is strictly concave, in which case every equilibrium entails $e_t = s_t = 0$ in each $t \geq 0$, just as in the extortion game without transfer signals.

Thus, positive equilibrium effort is possible only if $\phi(\cdot)$ has both convex and concave regions. For particular examples of such signal structures, we can construct equilibria with positive effort. Such equilibria require the principal to be punished on the equilibrium path. For these reasons, we view transfer investigations as unreliable, in the sense that they do not improve cooperation for a wide variety of signal distributions, and inefficient, because even when they can motivate effort, the resulting equilibrium entails occasional on-path punishments.

5 Dyadic Relationships

In the extortion game, the principal can punish an agent only by withholding pay, while an agent can punish the principal only by communicating with future agents. While this is a reasonable model of online platforms and other settings with short-lived interactions, other relationships, including those between managers and workers, are long-lived.

In this section, we explore how ongoing *dyadic* interactions between the principal and each individual agent can deter extortion. As is familiar from the literature on repeated games, dyadic relationships can be used to punish an agent for shirking or the principal for reneging on a hard-working agent. We now emphasize a third effect that is new to our setting: dyadic relationships can be used to punish the principal for acquiescing to extortion, which decreases the leverage of a shirking agent. By tying leverage to effort in this way, dyadic relationships facilitate coordinated punishments.

Consider the **extortion game with dyadic relationships**, which makes two changes to the extortion game. The first is minor: when the principal chooses s_t, we allow agent t to simultaneously make a transfer to the principal, $s_t^A \geq 0$, which is observed by the principal but not by other agents. Note that allowing such transfers would not change any of our other results. The second, more substantial change is that *after* agent t sends his message, m_t, the principal and agent t play a symmetric, simultaneous-move coordination game. The actions of this coordination game are observed by the two participants but not by any other agents.

While our analysis can be readily extended to general, asymmetric coordination games, we focus on the following simple game:

$$\begin{array}{ccc} & h & l \\ h & (v_H, v_H) & (v_L, v_L) \\ l & (v_L, v_L) & (v_L, v_L) \end{array},$$

where $v_H > v_L$. Letting v_t be the realized payoff from this coordination game, the principal's and agent t's payoffs are $e_t - s_t + s_t^A + v_t$ and $s_t - s_t^A - c(e_t) + v_t$, respectively.

The coordination game represents, in a simple way, any future interactions between the principal and an agent. We use this simple approach to demonstrate two lessons. First, dyadic relationships can deter misuse by tying leverage to effort. Second, unless these dyadic relationships are strong, the possibility of misuse still undermines cooperation.

In Appendix D, we show that these two lessons also hold in a setting with truly long-lived relationships. This appendix studies a repeated game, where in each period, one of a finite number of long-lived agents is randomly chosen to play the extortion game with the principal. We prove two results in this game. On the one hand, long-lived relationships can deter extortion by tying leverage to effort. On the other hand, unless those relationships are strong, extortion continues to undermine effort in equilibrium. This latter finding is consistent with the experience of the GM-Fremont plant, where long-lived relationships were not strong enough to prevent extortionary grievances from severely curtailing productivity.

Now, we show that positive effort can be sustained in the extortion game with dyadic relationships. However, effort is constrained by the strength of each dyadic relationship, as measured by $(v_H - v_L)$.

Proposition 5 *In the extortion game with dyadic relationships, $c(e_t) \leq 3(v_H - v_L)$ in every $t \geq 0$ of any equilibrium. If e^* is the minimum of e^{FB} and the solution to $c(e^*) = 3(v_H - v_L)$, then there exists a $\bar{\delta} < 1$ such that for any $\delta \geq \bar{\delta}$, $e_t = e^*$ in every $t \geq 0$ on the equilibrium path in any principal-optimal equilibrium.*

Proof: See Appendix A.

The constraint $c(e_t) \leq 3(v_H - v_L)$ reflects the fact that dyadic relationships optimally encourage cooperation via three channels: (i) they punish agents for shirking, (ii) they punish the principal for refusing to pay a hard-working agent, and (iii) they *reward* the principal for refusing to pay a shirking agent. The first two of these channels are familiar. The third channel is new and shows how dyadic relationships enable coordinated punishments.

Adopting the notation from Section 3, an agent's on-path leverage equals

$$L^* = \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}) + (v_H - v_L),$$

reflecting the fact that a principal who reneges is punished in both the period-t coordination game and the continuation equilibrium. If agent t shirks, then his leverage decreases to

$$\hat{L} = \max\left\{\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}) - (v_H - v_L), 0\right\},$$

since play in the coordination game rewards the principal for not paying a shirking agent. An agent can extort any transfer that is strictly less than his leverage, so his on-path transfer is $L^* - \hat{L}$ larger than the maximum amount he can extort. This difference is maximized if

$$\frac{\delta}{1-\delta}\left(\overline{\Pi} - \underline{\Pi}\right) = v_H - v_L, \tag{8}$$

in which case $L^* - \hat{L} = 2(v_H - v_L)$. Combining this difference in transfers with the fact that a shirking agent faces a direct punishment of $(v_H - v_L)$, we conclude that equilibrium effort must satisfy $c(e^*) \leq 3(v_H - v_L)$.

Other papers that focus on extortionary incentives in dyadic relationships, including Basu (2003), Dixit (2003a), and Myerson (2004), focus on how those relationships can encourage cooperation by increasing on-path leverage, L^*, and by directly punishing a shirking agent. In contrast, we focus on a third channel: dyadic relationships can complement coordinated

punishments by decreasing the leverage of a shirking agent, $\hat{L}$. Decreasing $\hat{L}$ would be irrelevant in a setting without extortion, since in that case, the value of coordinated punishments depends only on how they affect on-path leverage, L^*. In the extortion game, however, what matters is how leverage varies with effort, $L^* - \hat{L}$. Decreasing $\hat{L}$ means that each agent can be given access to coordinated punishments without misusing them. Consequently, as represented by (8), stronger dyadic relationships (measured by $v_H - v_L$) optimally expand the scope for coordinated punishments (measured by $\overline{\Pi} - \underline{\Pi}$).

Stepping back from the formal analysis, how might a firm cultivate dyadic relationships that deter extortion? The first step is to create manager-worker relationships with multiple equilibrium payoffs, so that $(v_H - v_L)$ is large. The firm's formal contracts must be structured in a way that supports this multiplicity (see, e.g., Che and Yoo (2001)). This was not the case at GM-Fremont, where managerial incentives were based heavily on formal contracts that left little room for relational contracts (Glass and Langfitt (2015)).[15] The firm's culture must then select among these equilibria in a way that deters extortion. Our analysis suggests that, by implementing the right formal incentives and fostering the right culture, an organization can encourage strong dyadic relationships that support effective coordinated punishments.

6 Extortion on the Equilibrium Path

This section enriches our baseline model so that some, but not all, agents commit to their threats. We show that in the resulting principal-optimal equilibria, cooperation and extortion coexist on the equilibrium path. Extortion spills over onto non-extorting relationships in two ways. First, the expectation of future extortion makes the principal less willing to pay transfers today. Second, to make extortion less attractive, principal-optimal equilibria entail weaker coordinated punishments that lead to less effort from non-extorting agents.

Consider the following **costly extortion game.** Suppose that, at the start of every

[15] Both managers and workers were incentivized to keep the production line running at all times, so they had little incentive to cooperate on, for example, fixing production mistakes or improving quality.

period $t \in \{0, 1, ...\}$, agent t privately observes a cost $k_t \geq 0$, $k_t \sim G(\cdot)$, and then chooses whether or not to invest. If he invests, then his payoff decreases by k_t and he plays the extortion game with the principal; otherwise, he plays the no-extortion game with the principal. Only the principal observes agent t's investment decision; other agents observe only m_t.

We interpret k_t as agent t's cost of committing to his threat. An agent might incur this cost by signaling that he is willing to follow through on extortionary threats.[16] The extortion and the no-extortion games are special cases of this game where $k_t = 0$ or k_t is large, respectively. In this section, we focus on distributions over k_t such that agents invest with an interior probability.[17]

We characterize principal-optimal equilibria in the costly extortion game in terms of the leverage given to each agent.

Proposition 6 *Consider the costly extortion game. In every period t of any principal-optimal equilibrium, on-path play is determined by an L_t that solves*

$$L_t \in \arg\max_{L \geq 0} \left\{ (1 - G(L))\, c^{-1}(L) - L \right\}$$

subject to the constraint

$$L \leq \frac{\delta}{1-\delta} \left((1 - G(L))\, c^{-1}(L) - L \right). \tag{9}$$

Agent t invests whenever $k_t < G(L_t)$. If agent t invests, then he chooses $e_t = 0$ and is paid $s_t = L_t$, while if he does not invest, then he chooses $e_t = c^{-1}(L_t)$ and is paid $s_t = L_t$.

Proof: See Appendix A.

[16] In the context of online platforms, an agent might incur this cost by using a non-official communication system in order to hide attempted extortion from the platforms, or it might represent the expected cost associated with the tail risk of getting caught.

[17] Another special case is a model in which each agent can commit to threats with a fixed probability, λ. This corresponds to a cost distribution where $k_t = 0$ with probability λ and otherwise k_t is large.

Using the notation from Section 3, define

$$L \equiv \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi})$$

as agent t's leverage. If agent t invests, then as in the proof of Proposition 2, his unique equilibrium strategy is to shirk and extort as much as possible, so $s_t = L$ and $e_t = 0$. If agent t does not invest, then as in the proof of Proposition 1, he is willing to choose e_t only if $c(e_t) \leq s_t$, while the principal is willing to pay s_t only if $s_t \leq L$. Agent t invests whenever the costs of doing so, k_t, are smaller than the gains, $L - (s_t - c(e_t))$.

As in Proposition 3, principal-optimal equilibria are sequentially principal-optimal. In each period of such an equilibrium, L ensures that the principal is exactly willing to compensate each agent for his effort, given that (i) an agent who invests exerts zero effort and is paid L, and (ii) L is no more than the principal's equilibrium continuation payoff. These two conditions lead to (9).

Increasing an agent's leverage increases both his temptation to invest and the effort he is willing to exert if he does not. In a principal-optimal equilibrium, agent t extorts with probability $G(L)$ and otherwise exerts effort $e_t = c^{-1}(L)$. Thus, higher L has opposing effects on equilibrium cooperation: it leads to a higher prevalence of extortion and higher payments to extorting agents, but it also leads to higher effort among those agents who do not extort. The optimal L balances these forces and so is typically lower than it would be without the possibility of misuse.

The dynamic enforcement constraint, (9), illustrates a further negative spillover from extorting to non-extorting relationships. The right-hand side of this constraint equals the principal's on-path continuation payoff. Future non-extorting agents contribute $c^{-1}(L) - L > 0$ to this payoff, while future extorting agents contribute $-L < 0$. Thus, even in non-extorting relationships, the expectation that *future* agents will extort undermines cooperation.

As Proposition 6 shows, an agent who invests in extortion disproportionately benefits

from severe coordinated punishments. Consequently, we might expect extorting agents to be disproportionately attracted to platforms and organizations that rely on coordinated punishments but fail to guard against misuse. Conti (2019) makes this point in the context of Airbnb, arguing that it was susceptible to this kind of negative selection because users could easily make multiple accounts. Conversely, organizations with restricted and stable memberships would be less susceptible to negative selection and misuse.

As in Sections 4 and 5, organizations can deter misuse in the costly extortion game by tying leverage to effort. We have seen that effort investigations and dyadic relationships can limit misuse when agents can costlessly commit to threats, which means that they can *a fortiori* do so in the costly extortion game. The principal might do even better by using these instruments in ways that would not be possible when extortion is costless. For instance, she might use these instruments to discourage investment in extortion without eliminating it entirely, in which case extortion would still occur on the equilibrium path.

7 Conclusion

This paper exposes a vulnerability in coordinated punishments: agents can misuse messages intended to report deviations. We also explore practical ways to restore cooperation, all of which build on the same core intuition: to deter misuse, tie an agent's leverage over the principal to his effort.

Fundamentally, misuse undermines cooperation because it severs the link from effort to transfer and message. Other modeling approaches that similarly sever this link would lead to the similar takeaways as this paper. Appendix B analyzes several such alternatives, including: (i) allowing the principal and each agent to bargain over the message (Halac (2012, 2015); Goldlücke and Kranz (2017)), (ii) endowing the agents with preferences for keeping their word (Vanberg (2008)) or preferences for reciprocity (similar to those documented in Fehr et al. (2020)), and (iii) imposing an equilibrium refinement in the no-extortion game

(Zhu (2018, 2019)). In each of these alternatives, just as in our main analysis, agents exert effort only if doing so increases their leverage.

All of these approaches, together with our applications, point to the conclusion that misuse is a crucial obstacle to cooperation across a variety of contexts. We do not claim that misuse arises *whenever* coordinated punishments are used; indeed, Appendix B shows that agents who have strict preferences for telling the truth (including revealing their own deviations) would not engage in misuse. Rather, our main point is that agents have a powerful incentive to misuse coordinated punishments in a way that undermines their value. Our applications illustrate, and our analysis demonstrates, that coordinated punishments are vulnerable to this type of misuse, which can severely undermine cooperation. Organizations ignore this vulnerability at their peril.

While the principal and agents are asymmetric in our model, extortionary threats are also a feature in more symmetric interactions, as in, for example, communal enforcement (e.g., Dixit (2007), Ali and Miller (2016)). In such settings, *both* sides can potentially extort one another. How do players cooperate in the presence of two-sided extortion? What networks best facilitate cooperation, and how are rents shared within those networks? How should business associations, communities, and firms structure communication channels to support strong relational contracts? We hope that our analysis provides a foundation for analyzing such questions.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. *Econometrica 87*(4), 1115–1153.

Ali, S. N. and C. Liu (2018). Conventions and coalitions in repeated games. Working Paper.

Ali, S. N. and D. Miller (2013). Enforcing cooperation in networked societies. Working Paper.

Ali, S. N. and D. Miller (2016). Ostracism and forgiveness. *American Economic Review 106*(8), 2329–2348.

Ali, S. N., D. Miller, and D. Yang (2017). Renegotiation-proof multilateral enforcement.

Andrews, I. and D. Barron (2016). The allocation of future business: Dynamic relational contracts with multiple agents. *American Economic Review 106*(9), 2742–2759.

Arnold, C. and R. Smith (2016, 10). Bad form, wells fargo. NPR.

Baker, G., R. Gibbons, and K. Murphy (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics 109*(4), 1125–1156.

Baker, G., R. Gibbons, and K. J. Murphy (2002). Relational contracts and the theory of the firm. *The Quarterly Journal of Economics 117*(1), 39–84.

Barron, D., J. Li, and M. Zator (2018). Productivity and debt in relational contracts. Working Paper.

Barron, D. and M. Powell (2018). Policies in relational contracts. Forthcoming, American Economic Journal: Microeconomics.

Basu, K. (2003). *Analytical Development Economics: the Less Developed Economy Revisited*. MIT Press.

Bernstein, L. (2015). Beyond relational contracts: Social capital and network governance in procurement contracts. *Journal of Legal Analysis 7*(2), 561–621.

Board, S. (2011). Relational contracts and the value of loyalty. *American Economic Review 101*(7), 3349–3367.

Bowen, T. R., D. M. Kreps, and A. Skrzypacz (2013). Rules with discretion and local information. *The Quarterly Journal of Economics 128*(3), 1273–1320.

Bull, C. (1987). The existence of self-enforcing implicit contracts. *The Quarterly Journal of Economics 102*(1), 147–159.

Chassang, S. and G. Padro i Miquel (2018). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. Forthcoming, Review of Economic Studies.

Che, Y.-K. and S.-W. Yoo (2001). Optimal incentives for teams. *The American Economic Review 91*(3), 525–541.

Conti, A. (2019). I accidentally uncovered a nationwide scam on airbnb. *Vice*.

Dewatripont, M. (1987). The role of indifference in sequential models of spatial competition: an example. *Economics Letters 23*(4), 323–328.

Dixit, A. (2003a). On modes of economic governance. *Econometrica 71*(2), 449–481.

Dixit, A. (2003b). Trade expansion and contract enforcement. *Journal of Political Economy 111*(6), 1293–1317.

Dixit, A. (2007). *Lawlessness and Economics: Alternative Modes of Governance*. Princeton University Press.

Fehr, E., M. Powell, and T. Wilkening (2020). Behavioral constraints on the design of subgame-perfect implementation mechanisms.

Fong, Y.-F. and J. Li (2017). Relational contracts, limited liability, and employment dynamics.

Freeman, R. and J. Medoff (1979). The two faces of unionism. *The Public Interest 57*, 69–93.

Fudenberg, D., D. Levine, and E. Maskin (1994). The folk theorem with imperfect public monitoring. *Econometrica 62*(5), 997–1039.

Gambetta, D. (1993). *The Sicilian Mafia: The Business of Private Protection*. Harvard University Press.

Glass, I. and F. Langfitt (2015). Nummi 2015.

Goldlücke, S. and S. Kranz (2017). Reconciliating relational contracting and hold-up: A model of repeated negotiations. Working Paper.

Greif, A., P. Milgrom, and B. Weingast (1994). Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy 102*(4), 745–776.

Guo, Y. and J. Hörner (2018). Dynamic allocation without money. Working Paper.

Halac, M. (2012). Relational contracts and the value of relationships. *American Economic Review 102*(2), 750–779.

Halac, M. (2015). Investing in a relationship. *RAND Journal of Economics 46*(1), 165–186.

Hörner, J. and N. Lambert (2018). Motivational ratings. *Review of Economic Studies*. Forthcoming.

Klein, T., C. Lambertz, and K. Stahl (2016). Market transparency, adverse selection, and moral hazard. *Journal of Political Economy 124*(6), 1677–1713.

Levin, J. (2002). Multilateral contracting and the employment relationship. *The Quarterly Journal of Economics 117*(3), 1075–1103.

Levin, J. (2003). Relational incentive contracts. *The American Economic Review 93*(3), 835–857.

Li, J., N. Matouschek, and M. Powell (2017, February). Power dynamics in organizations. *American Economic Journal: Microeconomics 9*(1), 217–41.

Lipnowski, E. and J. Ramos (2020). Repeated delegation. *Forthcoming, Journal of Economic Theory*.

Lippert, S. and G. Spagnolo (2011). Networks of relations and word-of-mouth communication. *Games and Economic Behavior 72*(1), 202–217.

Liu, C. (2019). Stability in repeated matching markets. Working Paper.

MacLeod, B. and J. Malcomson (1989). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica 57*(2), 447–480.

Malcomson, J. (2013). Relational incentive contracts. In R. Gibbons and J. Roberts (Eds.), *Handbook of Organizational Economics*, pp. 1014–1065.

Malcomson, J. (2016). Relational contracts with private information. *Econometrica 84*(1), 317–346.

Milgrom, P., D. North, and B. Weingast (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics and Politics 2*(1), 1–23.

Miller, D., T. Olsen, and J. Watson (2020). Relational contracting, negotiation, and external enforcement. *American Economic Review 110*(7), 2153–2197. Working Paper.

Miller, D. and J. Watson (2013). A theory of disagreement in repeated games with bargaining. *Econometrica 81*(6), 2303–2350.

Myerson, R. B. (2004). Justice, institutions, and multiple equilibria. *Chicago Journal of International Law 5*(1), 91–108.

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

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Peachey, K. (2015). Online reviews 'used as blackmail'. *BBC News*.

Proctor, J. (2018). Disgruntled bride ordered to pay 115k after defamatory posts ruin chinese wedding-photo business.

Starmans, J. (2017). Optimal agents.

Tranaes, T. (1998). Tie-breaking in games of perfect information. *Games and Economic Behavior 22*(1), 148–161.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica 76*(6), 1467–1480.

Watson, J. (2017). A general, practicable definition of perfect bayesian equilibrium.

Wolitzky, A. (2012). Career concerns and performance reporting in optimal incentive contracts. *B.E. Journal of Theoretical Economics (Contributions) 12*(1).

Wolitzky, A. (2013). Cooperation with network monitoring. *The Review of Economic Studies 80*(1), 395–427.

Zhu, J. Y. (2018). A foundation for efficiency wage contracts. *American Economic Journal: Microeconomics 10*(4), 248–288.

Zhu, J. Y. (2019). Better monitoring...worse productivity? Working Paper.

A Omitted Proofs

A.1 Proof of Proposition 3

Consider an equilibrium. Suppose $e_t = e$ at some on-path, period-t history, and let $\overline{\Pi}(y)$ and $\underline{\Pi}(y)$ be the principal's largest and smallest continuation payoffs following signal realization y, with corresponding messages $\overline{m}(y)$ and $\underline{m}(y)$. Define $L(y) \equiv \frac{\delta}{1-\delta}(\overline{\Pi}(y) - \underline{\Pi}(y))$.

For each effort e_t, agent t can choose

$$\mu_t(s, y) = \begin{cases} \overline{m}(y) & s_t \geq \hat{s} \\ \underline{m}(y) & \text{otherwise.} \end{cases}$$

Whenever

$$\hat{s} < \hat{s}(e_t) \equiv L(0) + \gamma(e_t)(L(1) - L(0)),$$

the principal's unique best response to this μ_t is to pay $\hat{s}$. On the other hand $s_t = 0$ is a best response to any $\hat{s} \geq \hat{s}(e_t)$. Thus, agent t's equilibrium effort, e, must satisfy

$$e \in \arg\max_{e'} \left\{ \hat{s}(e') - c(e') \right\}.$$

If $e > 0$, then a necessary condition for agent t to choose $e_t = e$ is that

$$c'(e) = \hat{s}'(e) = \gamma'(e)(L(1) - L(0)). \tag{10}$$

Since $\gamma'(e) > 0$, we can solve for $L(1) - L(0)$ in (10) and plug into the definition of $\hat{s}(e_t)$ to yield

$$\hat{s}(e) = L(0) + \gamma(e)\frac{c'(e)}{\gamma'(e)}.$$

Agent t earns at least 0, so

$$s_t - c(e) \geq \max\{0, \hat{s}(e) - c(e)\} \geq \max\left\{0, \gamma(e)\frac{c'(e)}{\gamma'(e)} - c(e)\right\} \equiv \bar{u}(e),$$

as desired.

Now, suppose $\gamma(\cdot)$ is concave. Since $c'(0) = c(0) = 0$, $\bar{u}(0) = 0$, and

$$\frac{d}{de}\left\{\gamma(e)\frac{c'(e)}{\gamma'(e)} - c(e)\right\} > 0,$$

so that $\bar{u}(\cdot)$ is strictly increasing. Moreover, the first-order condition (10) is both necessary and sufficient for agent t to exert effort $e_t = e$.

We now characterize principal-optimal equilibrium. Let Π^* be the principal's payoff in such an equilibrium. Note that on the equilibrium path, the principal's continuation payoff equals $\overline{\Pi}(y)$ following realization y, since otherwise agent t could demand a higher transfer using the promise of $\overline{\Pi}(y)$.

Suppose that $\overline{\Pi}(y) < \Pi^*$ for some $y \in \{0, 1\}$. In that case, we can increase both $\overline{\Pi}(y)$ and $\underline{\Pi}(y)$ by the same constant to keep $L(y)$, and hence agent t's incentives, unchanged. Doing so strictly increases the principal's payoff. So the principal's on-path continuation payoff equals Π^* in each $t \geq 0$ of any principal-optimal equilibrium. Then $\Pi^* = (1-\delta)(e_t - s_t) + \delta\Pi^*$, so $\Pi^* = e_t - s_t$ in any $t \geq 0$ on the equilibrium path.

In a principal-optimal equilibrium with $\gamma''(e) \leq 0$, $s_t = \mathbb{E}[L(y)|e]$, where e_t solves

$$\max_{L(\cdot)\geq 0, e} e - \mathbb{E}[L(y)|e]$$

subject to (10) and

$$L(y) \leq \frac{\delta}{1-\delta}\Pi^*.$$

Thus, $L(0) = 0$, in which case $L(1) = \frac{c'(e)}{\gamma'(e)}$ and so $\mathbb{E}[L(y)|e] = \gamma(e)\frac{c'(e)}{\gamma'(e)} = \bar{u}(e) + c(e)$. Substituting these simplifications into this constrained maximization problem yields the

constrained maximization problem in the statement of the Proposition. ■

A.2 Proof of Proposition 4

Fix a period t. Let $\Pi(m,x)$ be the principal's continuation payoff following message m and signal x. Let $\overline{\Pi}(x) = \max_m \Pi(m,x)$ and $\underline{\Pi}(x) = \min_m \Pi(m,x)$ with $\overline{m}(x)$ and $\underline{m}(x)$ being the corresponding maximizer and minimizer. We let π^D be the smallest payoff that the principal can guarantee herself,

$$\pi^D = \max_s -(1-\delta)s + \delta \mathbb{E}\left[\underline{\Pi}(x)|s\right]. \tag{11}$$

Define s_A as the smallest maximizer of (11). We argue that agent t's payoff is at least s_A. He can always choose $e_t = 0$ and

$$\mu_t(s,x) = \begin{cases} \overline{m}(x), & \text{if } s = s_A \\ \underline{m}(x), & \text{if } s \neq s_A. \end{cases}$$

Faced with this threat, the principal earns π^D from paying s_A and strictly less than π^D from paying $s < s_A$. Therefore, the principal will pay at least s_A.

Consider the set of transfers that can give the principal a higher payoff than π^D:

$$\{s : -(1-\delta)s + \delta \mathbb{E}\left[\overline{\Pi}(x)|s\right] > \pi^D\}. \tag{12}$$

If this set is nonempty, we let s_B be the supremum of this set. We argue that agent t can get a payoff arbitrarily close to s_B. In particular, he can choose $e_t = 0$ and

$$\mu_t(s,x) = \begin{cases} \overline{m}(x), & \text{if } s = s_B - \epsilon \\ \underline{m}(x), & \text{if } s \neq s_B - \epsilon. \end{cases}$$

Since $\phi(\cdot)$ is continuous, the principal's unique best response is to pay $s_t = s_B - \epsilon$ for small enough $\epsilon > 0$.

Now, define $\hat{s} = \max\{s_A, s_B\}$ if the set (12) is nonempty, and $\hat{s} = s_A$ otherwise. Agent t can guarantee a payoff arbitrarily close to $\hat{s}$ if he shirks, so he chooses $e_t = e^*$ only if $s^* - c(e^*) \geqslant \hat{s}$, which is our first necessary condition. Moreover, we can show that

$$-(1-\delta)s^* + \delta\mathbb{E}\left[\overline{\Pi}(x)|s^*\right] = s^D \tag{13}$$

$$-(1-\delta)\hat{s} + \delta\mathbb{E}\left[\overline{\Pi}(x)|\hat{s}\right] = s^D. \tag{14}$$

To see why (13) holds, note that the principal is willing to pay s^* so the left-hand side of (13) must be weakly higher than π^D. But either s_B does not exist, in which case (13) must hold with equality, or the supremum of the set (12) must be strictly below s^*, so that again (13) holds with equality. Equality (14) follows from the continuity of $\phi(\cdot)$ and the definition of $\hat{s}$.

Combining (13) and (14), we have

$$s^* - \hat{s} = \frac{\delta}{1-\delta}\left(\phi(s^*) - \phi(\hat{s})\right)\left(\overline{\Pi}(1) - \overline{\Pi}(0)\right). \tag{15}$$

Given (13) , $-(1-\delta)s + \delta\mathbb{E}\left[\overline{\Pi}(x)|s\right]$ must attain a local maximum at $s = s^*$, since otherwise (12) would contain elements arbitrarily close to s^* and so $s^* \leq \hat{s}$. Thus,

$$\phi'(s^*)\left(\overline{\Pi}(1) - \overline{\Pi}(0)\right) = \frac{1-\delta}{\delta} \tag{16}$$

and $\phi''(s) \leqslant 0$. Combining (15) and (16) yields our final necessary condition:

$$\phi'(s^*) = \frac{\phi(s^*) - \phi(\hat{s})}{s^* - \hat{s}}.$$

If $\phi(\cdot)$ is strictly concave, it cannot satisfy this condition for $s^* > \hat{s}$. ■

A.3 Proof of Proposition 5

Consider period t of an equilibrium. Define $\overline{\Pi}$ and $\underline{\Pi}$ as the principal's largest and smallest continuation payoffs, respectively, with corresponding messages $\overline{m}$ and $\underline{m}$. Agent t can always deviate to $e_t = s_t^A = 0$ and

$$\mu_t(s) = \begin{cases} \overline{m} & s = \hat{s} \\ \underline{m} & \text{otherwise.} \end{cases}$$

Following this deviation, the principal's unique best response is $s_t = \hat{s}$ if

$$\hat{s} < v_L - v_H + \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}). \tag{17}$$

Similarly, if agent t does not deviate, the principal is willing to pay $s_t = s^*$ only if

$$s^* \leq v_H - v_L + \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}). \tag{18}$$

Agent t is willing to choose $e_t = e^*$ only if $s^* - c(e^*) + (v_H - v_L) \geq \hat{s}$ for *any* $\hat{s}$ satisfying (17). Given the bound (18) on s^*, we conclude that $e_t = e^*$ in equilibrium only if $3(v_H - v_L) \geq c(e^*)$.

Each agent must earn at least v_L, so the principal's equilibrium payoff cannot exceed $e^* - c(e^*) + 2v_H - v_L$, where $e^* = e^{FB}$ if $c(e^{FB}) \leq 3(v_H - v_L)$ and e^* satisfies $c(e^*) = 3(v_H - v_L)$ otherwise. To complete the proof, we construct an equilibrium that attains this bound. Play starts in the cooperative phase: in each $t \geq 0$, agent t chooses $e_t = e^*$ and

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

Transfers equal $s_t = max\{0, c(e^*) - (v_H - v_L)\}$, $s_t^A = max\{0, (v_H - v_L) - c(e^*)\}$ if agent t does not deviate and $s_t = 0$, $s_t^A = (v_H - v_L)$ if he does. If either nobody deviates or agent t deviates from (e_t, μ_t) but then nobody deviates from (s_t, s_t^A), then $v_t = v_H$; otherwise,

$v_t = v_L$. Play continues in the cooperative phase until $m_t = D$, at which point it transitions to the punishment phase with probability α. In the punishment phase, $e_t = s_t = 0$ in each period. Let α satisfy

$$\max\{0, c(e^*) - 2(v_H - v_L)\} = \frac{\delta}{1-\delta}\alpha\left(e^* - c(e^*) + 2v_H - v_L\right).$$

For $\delta < 1$ sufficiently close to 1, $\alpha \in [0, 1]$.

The principal earns $e^* - c(e^*) + v_H + (v_H - v_L)$ surplus in each period of the cooperative phase. If agent t deviates in (e_t, μ_t), then he earns v_L by paying $s_t^A = (v_H - v_L)$ and $-s_t^A + v_L$ from deviating, so he has no profitable deviation from s_t^A. Regardless of μ_t, the principal has no profitable deviation from $s_t = 0$ following a deviation in (e_t, μ_t) if

$$v_H - v_L \geq \frac{\delta}{1-\delta}\alpha(e^* - c(e^*) + 2v_H - v_L) = \max\{0, c(e^*) - 2(v_H - v_L)\},$$

which holds because $c(e^*) \leq 3(v_H - v_L)$. On the equilibrium path, if $c(e^*) - (v_H - v_L) \geq 0$, then the principal has no profitable deviation from s_t because

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

This is because, by definition of α,

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

If $c(e^*) - (v_H - v_L) < 0$, then agent t has no profitable deviation from s_t^A because $c(e^*) - (v_H - v_L) + v_H \geq v_L$.

Given these transfers, agent t earns v_L from choosing the equilibrium (e_t, μ_t) and no more than v_L from deviating. So this strategy profile is an equilibrium. It is principal-optimal because it attains the upper bound on the principal's equilibrium payoff. ■

A.4 Proof of Proposition 6

Consider an equilibrium and a history at the start of period t. Define $\overline{\Pi}$ and $\underline{\Pi}$ as in the proof of Proposition 2, with corresponding messages $\overline{m}$ and $\underline{m}$, and let

$$L_t \equiv \frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi}).$$

Suppose agent t invests. If $e_t > 0$ or $s_t < L_t$, then the deviation from the proof of Proposition 2 is profitable for $\epsilon > 0$ sufficiently small. Consequently, $e_t = 0$ and $s_t = L_t$ whenever agent t invests.

Suppose agent t does not invest. He must earn at least a payoff of zero in equilibrium, so $s_t - c(e_t) \geq 0$. The principal must be willing to pay s_t, so $s_t \leq L_t$. For any e_t and s_t that satisfy these two constraints, consider the following strategy profile:

1. Agent t chooses e_t.

2. The principal pays s_t if agent t has not deviated and pays nothing otherwise.

3. Agent t sends $\overline{m}$ if no deviation has occurred and $\underline{m}$ otherwise.

If agent t chooses e_t, the principal is willing to pay s_t because $s_t \leq L_t$. If agent t deviates, then $m_t = \underline{m}$ regardless of the principal's action, so she pays nothing. Agent t is willing to choose e_t because $s_t \geq c(e_t)$. Thus, neither player has a profitable deviation from this strategy profile. We conclude that any (e_t, s_t) with $c(e_t) \leqslant s_t \leqslant L_t$ can be implemented in an equilibrium, as desired.

Since an investing agent exerts no effort and obtains a pay of L_t, from now on we use e_t, s_t for the effort exerted by, and the pay received by, agent t who didn't invest. Given this continuation play, agent t is willing to invest if and only if

$$L_t - (s_t - c(e_t)) \geq k_t,$$

where $L_t - (s_t - c(e_t))$ and k_t represent the gain from, and cost of, investment, respectively.

Now, consider a principal-optimal equilibrium, and let Π^* equal the principal's maximum equilibrium payoff. We must have $\overline{\Pi} = \Pi^*$ in each period t, since agent t's incentive depends only on L_t so we can increase $\overline{\Pi}, \underline{\Pi}$ while keeping L_t fixed. The principal's payoff is

$$\max_{L_t, s_t, e_t} G(L_t - (s_t - c(e_t)))\left\{\delta\Pi^* - (1-\delta)L_t\right\} + (1 - G(L_t - (s_t - c(e_t))))\left\{(1-\delta)(e_t - s_t) + \delta\Pi^*\right\} \tag{19}$$

subject to the constraint that $c(e_t) \leqslant s_t \leqslant L_t$. The constraint $s_t \leqslant L_t$ must bind, since the principal would like L_t to be as small as possible. Substituting $L_t = s_t$ into (19), the objective in (19) becomes:

$$G(c(e_t))\left\{\delta\Pi^* - (1-\delta)s_t\right\} + (1 - G(c(e_t)))\left\{(1-\delta)(e_t - s_t) + \delta\Pi^*\right\}$$

The derivative of this objective with respect to s_t is $-1 + \delta$. Hence, it is optimal to choose $s_t = c(e_t)$. The objective in (19) becomes

$$\delta\Pi^* + (1-\delta)\left((1 - G(c(e_t))e_t - c(e_t)\right).$$

Therefore, the optimal effort maximizes $(1 - G(c(e_t))e_t - c(e_t)$ and Π^* is given by this maximum:

$$\Pi^* = \max_{e_t} (1 - G(c(e_t))e_t - c(e_t).$$

The formula for Π^* is quite clear. The principal has to pay $c(e_t)$ to both an extorting agent and a nonextorting one. However, she only obtains e_t from the nonexorting agent, which occurs with probability $1 - G(c(e_t))$. ■

B Appendix: Interpreting Commitment

In this appendix, we provide three alternative interpretations of the extortionary threats at the heart of our analysis. Appendix B.1 re-interprets threats as the outcome of bargaining between the principal and each agent. Appendix B.2 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 B.3 re-interprets commitment to μ_t as an equilibrium refinement of the no-extortion game.

B.1 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. The outcome of this deviation is that the principal pays and the message leads to her maximum possible 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 and the agent Nash bargain over a transfer that is paid to the agent, $s \geq 0$, and a message, $m \in \mathcal{M}$. If bargaining breaks down, then the disagreement point is message m^o and transfer s^o. The agent's bargaining weight is $\alpha \in [0, 1]$.

The principal's and the agent's payoffs are $e - s + \frac{\delta}{1-\delta}\Pi(m)$ and $s - c(e)$, respectively, where $\Pi : \mathcal{M} \to \mathbb{R}_+$ is an arbitrary function and $c(\cdot)$ is strictly increasing. We interpret this model as a single period of the no-extortion game. Under this interpretation, $\Pi(m)$ represents the principal's continuation payoff from interactions with future agents.

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. The Nash bargaining solution requires the agreed-upon outcome, (s, m), to satisfy $m \in \arg\max_{\tilde{m}} \Pi(\tilde{m})$.

Let $\overline{\Pi} \equiv \max_{\tilde{m}} \Pi(\tilde{m})$; then, s must give the agent a payoff of

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

so that

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

Note that this transfer is independent of e. Since $c(\cdot)$ is strictly increasing, we conclude that $e = 0$ in any equilibrium. ■

To see the connection between this result and Proposition 2, suppose that $s^o = 0$ and m^o minimizes the principal's continuation payoff: $m^o = \arg\min_{\tilde{m}} \Pi(\tilde{m})$, with 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. It is in this sense that 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 $\alpha > 0$.[18]

[18] 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

How can organizations encourage cooperation in this setting? As in the main text, agent t is willing to exert effort only if doing so increases his leverage. Therefore, the same remedies that we consider in the paper can also lead to effort in the game with interim bargaining and renegotiation.[19] 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.

As written, this model assumes that the principal commits to the transfer, s, at the time of bargaining. However, this assumption is not essential. The principal prefers to pay $s = \alpha\frac{\delta}{1-\delta}(\overline{\Pi} - \underline{\Pi})$ rather than getting punished in the continuation game, since

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

Thus, the outcome of the bargaining process can be sustained in an equilibrium without commitment.

B.2 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 eliminate misuse and restore cooperation.

For simplicity, we consider a one-period version of the game between a principal and a single agent, which we call the **cheap-talk game with behavioral preferences:**

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

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.

[19] In this alternative formulation, dyadic relationships (Section 5) 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.

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

3. The agent chooses $m_t \in \mathcal{M}$.

The principal's payoff is $e_t - s_t + \frac{\delta}{1-\delta}\Pi(m_t)$, where we interpret $\Pi(\cdot)$ as the principal's continuation payoff from interactions with future agents. The agent's payoff is $s_t - c(e_t) + H(\mu_t, e_t, s_t, m_t)$, where the function $H(\cdot)$ captures agent t's behavioral preferences.

B.2.1 Word-keeping preferences

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

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

for some $\epsilon > 0$. That is, we interpret the threat μ_t as a promise of the form "if you pay me s_t, I will send message m_t." The agent earns extra utility $\epsilon > 0$ from following this promise.

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.2.2 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).

We show that for most reference transfers, 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 and paying 0 and earning her minimum continuation payoff. Under that condition, positive effort can be sustained in equilibrium for a reason similar to Section 5: the principal is willing both to pay an agent who exerts effort and not to pay an agent who shirks.

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

1. *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 $e_t = 0$.

2. *If*

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

then there exists an equilibrium in which e_t satisfies $c(e_t) \leq \frac{\delta}{1-\delta}(\Pi(\overline{m}) - \Pi(\underline{m}))$.

Proof of Proposition 9: If $s_t \geq s_{REF}$, the agent prefers to maximize $\Pi(m_t)$, so $m_t = \overline{m}$ is a best response for the agent. If $s_t < s_{REF}$, the agent prefers to minimize $\Pi(m_t)$, so $m_t = \underline{m}$ is a best response for the agent.

Given this mapping from transfers to messages, 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 therefore 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. Faced with this strategy, the agent is willing to choose $e_t = e^*$.

B.2.3 Truth-telling preferences

Finally, we study **truth-telling** preferences (e.g., Abeler et al. (2019)). We model truth-telling preferences by defining 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 some effort level that satisfies

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

Then, define

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

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.

These truth-telling preferences are enough to eliminate extortion, so that positive effort can be sustained in equilibrium.

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. ■

B.3 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, both in the extortion game and in each remedy.

Recall that the no-extortion game is identical to the extortion game, except that each agent t chooses m_t freely at the end of period t rather than being committed to μ_t.

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 follows immediately from the fact that agents are indifferent among messages and so are willing to follow their threats. Proposition 2 shows such an equilibrium exists, which completes the proof. 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 $t \geqslant 0$,

1. Agent t chooses e_t, μ_t as in σ^*.
2. The principal chooses s_t as in σ^*.
3. Agent t chooses $m_t = \mu_t(s_t)$.
4. Let a_t denote the action profile in the coordination game. If agent t follows this message strategy, a_t is as in σ^*; otherwise, $a_t = (l, l)$.

No player has a profitable deviation from a_t because a_t is always an equilibrium of the simultaneous move game at the end of the period. By the choice of a_t following a deviation in m_t, agent t has a weak incentive to follow the specified message strategy m_t. But then 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, as desired. ■

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. The only complication to this argument arises in the extortion game with dyadic relationships, since an agent's payoff in the coordination game can potentially respond to his message. However, we can always find an equilibrium in which agents are punished in the dyadic relationship if they deviate from their threats, in which case agents are willing to follow through on μ_t.

C Appendix: Communication by the Principal and Multiple Extortion Opportunities

This appendix explores alternative ways to model communication. In Appendix C.1, 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 C.2 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.[20] Finally, once the principal pays an extorting agent, that payment is sunk and so the agent has an incentive to extort again. Appendix C.3 explores cooperation when agents have multiple opportunities to extort.

C.1 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

[20] 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.

principal's continuation payoff after agent t's message m. We let $\overline{\Pi}$ and $\underline{\Pi}$ be the highest and 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.[21] The same argument as in the previous case applies, so every equilibrium involves zero effort in each period.

C.2 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.

[21] 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.

Consider the following strategy profile. Play starts in the cooperation phase. In this 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 C.1 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 C.1 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.3 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 5: 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 Appendix: Long-run Agents

This appendix studies coordinated punishments with long-run agents. Appendix D.1 presents the model, which builds on Section 5. Appendix D.2 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 D.3 shows that equilibrium effort falls to 0 as the number of agents increases.

D.1 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.

D.2 Long-Run Relationships can Deter Extortion (for $N = 2$)

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*

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

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. It is in this sense that with two agents, long-lived relationships are enough to eliminate the inefficiency associated with misuse.

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:

$$\begin{array}{ccccc} (i,-i) & & U_i & U_{-i} & \Pi \\ (O,O) & & 0 & 0 & e^* - c(e^*) \\ (S,O) & & 0 & 0 & \frac{1}{2-\delta}\left(e^* - c(e^*)\right) \\ (O,B) & & -\frac{1-\delta}{2-\delta}c(e^*) & \frac{1-\delta}{2-\delta}\frac{\delta e^*}{2} & \frac{(1-\delta)e^*}{2} \\ (S,B) & & -\frac{(1-\delta)\delta}{(2-\delta)^2}c(e^*) & \frac{2(1-\delta)}{(2-\delta)^2}\frac{\delta e^*}{2} & 0 \end{array} \cdot$$

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. ■

D.3 Extortion Remains a Serious Problem (as $N \to \infty$)

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. ■