Under Suspicion: Trust Dynamics with Secret Undermining*

Aaron Kolb† Erik Madsen‡

March 29, 2022

Abstract

We study how an organization should dynamically screen an agent of uncertain loyalty whom it suspects of committing damaging acts of undermining. The organization controls the stakes of the relationship, while the agent strategically times undermining, which can occur repeatedly and is detected only stochastically. The optimal commitment stakes policy exhibits both discreteness and gradualism, with distinct "untrusted" and "trusted" phases featuring gradually rising stakes during the untrusted phase and a discrete gap in stakes between phases. This policy is also the equilibrium outcome when the organization cannot commit, and the agent's equilibrium undermining policy exhibits variable, non-monotonic intensity.

JEL Classification: C70, D82, D83, D86, M51

Keywords: Principal-agent model, undermining, variable stakes, imperfect monitoring

*An earlier version of this paper circulated under the title "Leaks, Sabotage, and Information Design". The authors thank Alessandro Bonatti, Anton Kolotilin, Elliot Lipnowski, Laurent Mathevet, Eran Shmaya, Alex Wolitzky, and numerous seminar audiences for helpful conversations and comments.

†Department of Business Economics and Public Policy, Kelley School of Business, Indiana University (Email: kolba@indiana.edu).

‡Department of Economics, New York University (Email: emadsen@nyu.edu).

1 Introduction

In many principal-agent relationships, loyalty is critical to the success of the relationship. While incentives can sometimes be aligned by monetary rewards or direct monitoring, these tools are of limited use when an agent's role entails complex responsibilities or significant autonomy. In such settings, the agent cannot be trusted to faithfully execute his employer's goals unless he shares those goals. In this paper we study the dynamics of principal-agent relationships when an agent's loyalty is uncertain.

Disloyalty can arise in a variety of agency settings. A leading example is espionage, which may occur whenever agents have access to privileged or secret information valued by a foreign power or industrial competitor.[1] Disloyalty can also arise in contexts as diverse as office politics and government bureaucracy: An ambitious subordinate may discredit his supervisor by sabotaging her projects or reporting embarrassing mistakes to her superiors, and an ideological civil servant may disrupt implementation of a policy by quietly disobeying its mandates or leaking sensitive internal discussions to the press.[2] Whatever the setting, if a principal suspects one of her agents of disloyal acts, she must screen him to determine whether the relationship is worth continuing.

We focus on screening via two key instruments: variable responsibility and imperfect monitoring of the agent's behavior. Ideally, the principal would like to give the agent minimal responsibility while closely monitoring his actions, until enough time has passed that a disloyal agent would have been caught in the act. However, screening is complicated by the fact that agents are strategic, and a disloyal agent under suspicion may forego subversive acts today to build trust and gain better opportunities tomorrow. As a result, the principal must weigh how much responsibility to give the agent, as well as how quickly to ramp up responsibility as the agent gains her trust. We build a dynamic model of this strategic interaction between principal and agent.

[1] Historically, espionage has been common during wartime and other periods of heightened geopolitical tension, such as the Cold War. It is also of significant contemporary relevance, as the United States government has accused China of substantial espionage activity, particularly targeted at the acquisition of sensitive technology: https://www.fbi.gov/news/speeches/responding-effectively-to-the-chinese-economic-espionage-threat. For a list of recent Chinese-linked espionage cases including commercial and military technologies, see https://www.csis.org/programs/technology-policy-program/survey-chinese-linked-espionage-united-states-2000.

[2] Disloyalty by civil servants has been widely discussed in US politics in the context of a "deep state" of White House officials accused of working to undermine the president. Miles Taylor, a former Homeland Security chief of staff within the Trump administration, has openly admitted to such behavior, and has described a group of colleagues with similar motives: https://www.nytimes.com/2018/09/05/opinion/trump-white-house-anonymous-resistance.html. The Biden administration has reportedly faced similar disloyalty by Immigrations and Customs Enforcement officers: https://www.nytimes.com/2021/02/03/us/politics/biden-trump-immigration.html.

In our model, a principal employs an agent to perform a task repeatedly over time. The principal may decide how much responsibility to give the agent to complete his task, which we quantify by a one-dimensional *stakes* variable. Higher stakes increase the agent's task productivity, and so the principal would prefer to give a loyal agent maximum responsibility. However, at any point in time the agent may choose to secretly *undermine* the principal, and the damage inflicted by an act of undermining is also increasing in the stakes. Any act of undermining has a chance of going undetected by the principal, giving the agent the opportunity to undermine multiple times.

The agent may be either loyal or disloyal—a loyal agent has no incentive to undermine, while a disloyal agent wishes to inflict maximum cumulative damage on the principal. The agent's loyalty is initially private, and must be learned by the principal by watching for evidence of undermining. The principal can attempt to screen the agent by varying the stakes of the task over time, and she may also fire the agent at any time.

Our main result is a characterization of the principal's optimal stakes policy under commitment. Stakes evolve through three distinct phases. First, during an initial *probationary* phase stakes are frozen at a minimal level. Next, stakes are raised gradually at a constant rate during an *escalation* phase. Finally, the agent is *cleared* and stakes jump upward to their maximal level, where they remain thereafter in a *trusted* phase. This stakes policy induces undermining by the disloyal agent throughout the probationary and trusted phases, while during the escalation phase the disloyal agent is indifferent between undermining and not, with any choice yielding the same expected profits for the principal. These dynamics are broadly consistent with evidence from real-world espionage investigations, as we document in Section 4.

A key prediction of our model is that when one party can repeatedly and privately inflict harm on the other, relationships may exhibit abrupt transitions between low-stakes "untrusted" and high-stakes "trusted" phases. Our analysis therefore qualifies the prediction of gradualism commonly made in the existing literature on long-term bilateral relationships with variable stakes, for instance in Sobel (1985), Watson (1999, 2002), and Kreps (2018) (when screening out uncooperative agents), and in Thomas and Worrall (1994), Albuquerque and Hopenhayn (2004), Rayo and Garicano (2017), Fudenberg and Rayo (2019), Fudenberg et al. (2021), and Atakan et al. (2020) (when aligning incentives for realizing gains from trade). The main innovation in our paper relative to this literature, and the driving force behind our novel finding of discrete stakes transitions, is imperfect monitoring of a repeated action.

Optimal stakes dynamics are shaped by the interaction between the principal's desire to backload a disloyal agent's payoffs and her need to deter him from delaying undermining.

When designing an optimal contract, the principal maximizes her payoff from a loyal agent subject to delivering a target utility level to a disloyal one. Crucially, a disloyal agent applies a higher effective discount rate to future payoffs than does a loyal one, due to the possibility of detection and termination. As a result, the principal optimally delivers a disloyal agent's utility through a contract which backloads his payoffs as much as possible.

In our setting, maximum backloading corresponds to a bang-bang outcome in which stakes are held low for a time and then abruptly raised to their maximal level. This motive generates a discrete clearance event as highlighted above. However, the amount of backloading the principal can implement is limited by a disloyal agent's ability to strategically delay undermining if continuation utilities rise too quickly. To deter such delay, the principal must gradually raise stakes early in the contract to moderate the abruptness of the stakes rise at the time of clearance. This incentive constraint generates a smooth rise of stakes during the escalation phase.

This mechanism is robust to a number of important extensions to our model. We show that the loyal agent's optimal stakes curve is shaped by the same forces, and exhibits the same combination of gradual and abrupt stakes adjustments, in environments with endogenous monitoring; punishments for undermining; and agent preferences incorporating factors beyond the principal's payoff. A new feature of these environments is that the principal may benefit from offering a menu of employment contracts, with the disloyal agent screened into a separate contract with reduced stakes in return for reduced monitoring or punishment following detection.

Optimal stakes dynamics are also robust to a lack of commitment power. We show that in a dynamic game between the principal and agent, in equilibrium the principal chooses the same stakes path, and achieves the same payoff, as under commitment. Furthermore, the requirement of sequential rationality yields sharp predictions about the dynamics of undermining, resolving the degeneracy arising in the analysis under commitment. Notably, the disloyal agent's unique equilibrium undermining policy exhibits non-monotonic undermining intensity—the agent undermines intensively during the probationary phase, scales back at the beginning of the escalation phase, and returns gradually to full undermining by the time he is cleared.

The redundancy of commitment power distinguishes our results from outcomes in models of screening by a durable-good monopolist, for instance Stokey (1981), Gul et al. (1986), Boleslavsky and Said (2013), Daley and Green (2020), and Chaves (2020). In this literature, the adjustment of prices over time serves a screening function similar to the adjustment of stakes in our setting. A key finding is that seller impatience introduces a Coasian force leading the commitment outcome to outperform the best equilibrium without commitment.

In contrast, in our setting commitment is not needed to achieve the principal's preferred outcome.

The remainder of the paper is organized as follows. Section 2 presents the model. We derive the optimal contract under commitment in Section 3. Section 4 applies our results to rationalize stylized facts from espionage investigations. Section 5 discusses robustness to endogenous monitoring, punishments, and relaxations of the zero-sum assumption. Section 6 analyzes the principal-agent relationship without commitment. Section 7 offers concluding remarks and suggests directions for future research. The appendices contain proofs of all results.

2 The model

In this section, we introduce the model. The development here is deliberately informal; we provide some additional formalities in Section 3.

2.1 The environment

A (female) principal employs a (male) agent over a potentially infinite horizon in continuous time. The principal specifies a $[\phi, 1]$-valued *stakes curve* $x = (x_t)_{t\geq 0}$, a reduced form for the agent's information or access to resources, where $\phi \in [0, 1)$ is an exogenous lower bound on feasible stakes.

The principal earns a baseline flow payoff of x_t from employing the agent when period-t stakes are x_t, representing work the agent completes for the principal. However, at each instant, the agent may secretly undermine the principal, modeled by a choice of an undermining intensity $\beta_t \in [0, 1]$. When the stakes are x_t, undermining with intensity β_t inflicts a flow loss of $K\beta_t x_t$ on the principal, where $K > 1$. The principal discounts payoffs at rate $r > 0$, and so given an undermining policy β and a stakes process x, her expected payoffs from employing the agent until time T are

$$\Pi(\beta, x, T) = \mathbb{E} \int_0^T e^{-rt}(1 - K\beta_t)x_t \, dt.$$

The agent's motives and opportunities depend on a preference parameter $\theta \in \{G, B\}$. When $\theta = G$, the agent is *loyal*. A loyal agent has preferences which are perfectly aligned with the principal's. That is, given an undermining policy β and a stakes curve x, the loyal agent's expected payoff from being employed until time T is $\Pi(\beta, x, T)$. Since the loyal agent's motives are perfectly aligned with the principal's, for simplicity we assume that the

loyal agent cannot undermine: $\beta_t = 0$ for all time is the agent's only feasible undermining policy. (His preferences will still play a role when choosing among alternative employment contracts.) On the other hand, when $\theta = B$ the agent is *disloyal*. The disloyal agent has interests diametrically opposed to the principal's, and his expected payoff is $-\Pi(\beta, x, T)$.

Neither working for the principal nor undermining incurs any direct costs to the agent: he cares only about the payoffs the principal receives. If the agent is terminated or does not accept employment, he and the principal both receive an outside option normalized to 0.[3]

2.2 The information structure

Prior to accepting employment with the principal, the agent is privately informed of his type θ. The principal believes that $\theta = G$ with probability $q \in (0, 1)$.

Once the agent is employed, the principal only imperfectly observes whether the agent has undermined. Whenever the agent undermines with intensity β_t over a time interval dt, the principal immediately receives definitive confirmation of this fact with probability $\gamma\beta_t\, dt$, where $\gamma > 0$. Otherwise, the act of undermining goes permanently undetected.

If the agent chooses undermining policy β, the cumulative probability that his undermining has gone undetected by time t is therefore $\exp\left(-\gamma \int_0^t \beta_s\, ds\right)$. Note that the detection rate is *not* cumulative in past undermining — the principal has one chance to detect an act of undermining at the time it occurs. To ensure consistency with this information structure, we assume the principal does not observe her ex post payoffs until the end of the game.

2.3 Discussion of model assumptions

2.3.1 Imperfect monitoring

We have assumed that any steps the agent takes to undermine the principal are detected only stochastically. We view this assumption as a reasonable first approximation to the basic dilemma faced by a disloyal agent such as a spy, who must remain on the job nominally aiding the organization while simultaneously undertaking damaging acts which might be discovered. The inability of the principal to detect all acts of undermining is consistent with our application to espionage investigations, as a government may only become aware that information has been leaked to a foreign power with substantial delay. In other applications, imperfect monitoring can be justified whenever suspicion arises in an organization where the principal employs many agents. In that case, an undetected act of undermining may be

[3] In Section 5, we explore several extensions with richer agent preferences, including reputational benefits from employment and negative termination payments.

interpreted as an instance which the principal is unable to trace back to the agent under suspicion.

We have also assumed that the principal's monitoring technology is known to the agent. In many contexts, however, agents may be uncertain about how, when, and whether they are being monitored. Nonetheless, our Poisson monitoring technology captures some of this uncertainty. Concretely, suppose that the agent varies his methods of undermining over time, and/or the principal rotates her surveillance channels over time. In that case, an act of undermining is detected only if the principal happens to be performing the appropriate forms of surveillance to detect the current act, leading to stochastic detection from the agent's perspective as in our model. Such stochasticity is especially plausible when monitoring requires costly physical surveillance and the principal has limited resources.

2.3.2 Minimal stakes

Our specification of the task structure requires that the project be operated above a minimal stakes level ϕ, which may be strictly greater than zero. This lower bound is designed to capture several different technological limitations on the way in which employees may be treated. First, in some settings organizations may not be able to exclude employees from obtaining some minimum level of access simply by being employed. For instance, employees may require keycard access to the building, and may be able to glean sensitive information from discussions with coworkers, company-wide e-mails, or unguarded recycling bins.

Second, in many contexts employees may not be productive unless they are given minimal levels of access and information commensurate to the needs of their assignment. While an untrusted employee could simply be reassigned to an alternative low-value task, this action would be tantamount to firing the employee and would give up the gains from assigning a loyal employee to a high-value task. This is especially true if the low-value task does not leave opportunities for effective undermining, ruling out screening for loyalty by monitoring the employee's actions. In case none of these limitations is relevant in a particular application, the corresponding solution can be recovered as a special case of our model by setting $\phi = 0$.

3 The commitment outcome

In this section we analyze the outcome of the employment relationship when the principal has commitment power. We characterize the principal's optimal stakes curve and employment policy as well as the disloyal agent's optimal undermining policy. Section 3.1 formally states the contracting problem. Section 3.2 presents the results of the analysis and highlights

important features of the solution, while Section 3.3 discusses the key economic forces at play and provides a heuristic derivation of the optimal stakes curve.

3.1 The contracting problem

The principal offers the agent a menu $\mathcal{M}$ of contracts, with each contract $\mathcal{C}$ committing to a path of stakes and a termination policy specifying at what time and under what conditions the agent will be fired. Both stakes and termination can condition on the history of detected undermining.[4] Each contract additionally makes a recommendation for how the disloyal agent should undermine over time. No transfers are permitted. To ensure that expected payoffs are well-defined, we impose the condition that realized paths of stakes and undermining are càdlàg.

We call a contract *incentive-compatible* if the recommended undermining policy is optimal for the disloyal agent. In our model, it is not necessary to explicitly specify an incentive-compatible undermining recommendation. This is because the principal's payoff depends on the disloyal agent's actions only through that agent's expected utility. As a result, all incentive-compatible undermining policies yield the same payoff to the principal. We will therefore drop an explicit specification of the undermining recommendation from our description of contracts.[5]

Since the principal faces two possible agent types, it is without loss to offer a menu of at most two contracts $\mathcal{M} = (\mathcal{C}^G, \mathcal{C}^B)$, where $\mathcal{C}^\theta$ is the contract intended for the agent of type θ. We call a menu *obedient* if each agent type prefers the contract intended for him over the other agent's contract, and prefers it over declining employment entirely. It is without loss to restrict attention to obedient menus, as any non-obedient menu may be replaced by a payoff-equivalent menu which is obedient.

In our model, the only obedience condition relevant to the principal is the agent's desire to choose the wrong contract. That is, each agent type is always willing to accept employment: under any contract, the loyal agent trivially obtains a non-negative payoff, while the disloyal agent can guarantee himself a non-negative payoff by undermining with full intensity at all times.[6]

[4] For simplicity, we do not analyze randomized contracts.

[5] In several extensions in Section 5, we depart from the assumption of a zero-sum relationship between the principal and disloyal agent. As a result, the principal is no longer indifferent between incentive-compatible undermining policies, and we will explicitly specify an undermining recommendation.

[6] In Section 5.4, we study an extension in which the principal can impose punishments on the agent. In that environment, the obedience constraint that the agent prefers to accept employment becomes relevant.

3.2 The characterization

In Section 3.3, we establish two important facts. First, the principal optimally pools both agents into a single contract. Second, an optimal termination policy either declines to employ the agent (i.e., terminates him immediately), in which case we say the contract is *degenerate*, or else terminates him the first time undermining is detected. We will show later that the principal optimally employs the agent so long as the agent's initial reputation q is sufficiently large. The main content of a non-degenerate contract is therefore the *stakes curve*, which specifies the evolution of stakes in the absence of detected undermining.

We will call a stakes curve *optimal* if it is part of an optimal non-degenerate contract. The following theorem characterizes the optimal stakes curve.

Theorem 1. *There exists a unique optimal stakes curve x^*, which is monotone increasing and proceeds through up to three phases of fixed lengths:*

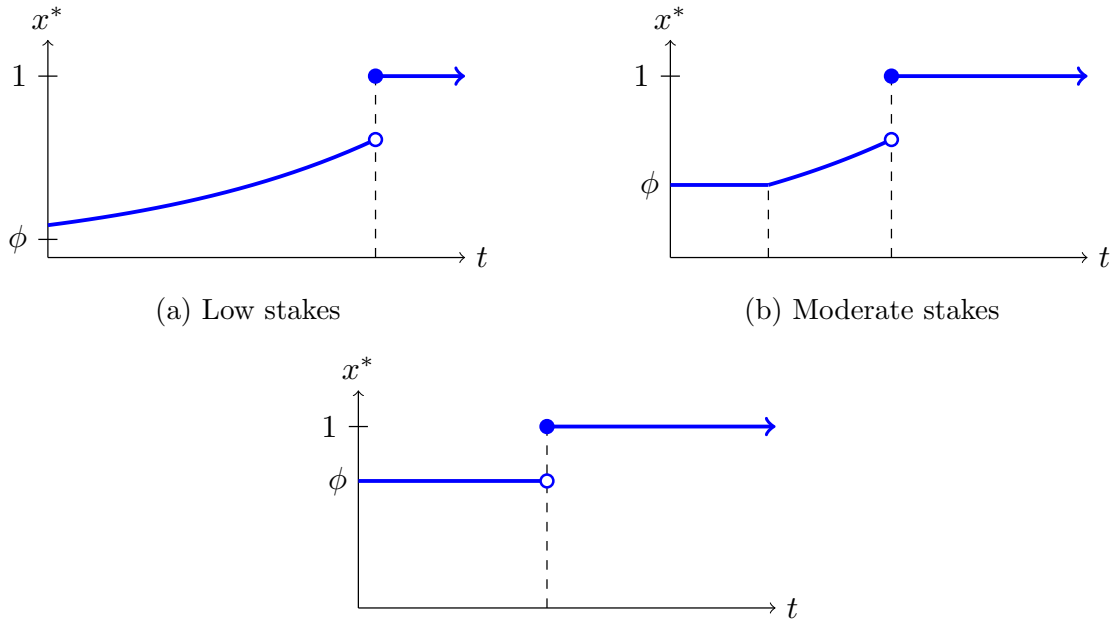
1. *In the* probationary phase, $x_t^* = \phi$,
2. *In the* escalation phase, *stakes grow continuously at rate $r + \gamma/K$,*
3. *In the* trusted phase, $x_t^* = 1$.

If $q < (K-1)/K$, the trusted phase begins at a strictly positive time, and stakes jump discontinuously upward at the beginning of the trusted phase. Further, there exist thresholds $0 < \underline{\phi} < \overline{\phi} < 1$ such that the escalation phase has positive length if $\phi < \overline{\phi}$, while the probationary phase has positive length if $\phi > \underline{\phi}$.

If $q \geq (K-1)/K$, the trusted phase begins at time zero.

Figure 1 illustrates the possible stakes curves that can arise in an optimal contract. The form of the optimal stakes curve depends on both the agent's initial reputation and the lower bound ϕ on stakes. When $q < (K-1)/K$, we refer to the environment as one of low stakes, moderate stakes, or high stakes, depending on which phases are present in the optimal stakes curve. Figure 2 depicts the values of q and ϕ for which the various forms arise. The figure also illustrates the range of parameters for which the optimal contract is non-degenerate; note that all possible forms of the optimal stakes curve may arise for such parameter values.

We will describe the time at which the agent becomes trusted as a *clearance event*. The following lemma reports how the timing of the various phases of the relationship and the size of the jump in stakes when the agent is cleared depend on monitoring quality. The first two comparative statics are illustrated in Figure 3.



(a) Low stakes

(b) Moderate stakes

(c) High stakes

Figure 1: The form of the optimal stakes curve for different values of ϕ.

Lemma 1. *As γ increases:*

- *The size of the stakes jump at clearance decreases,*
- *The agent is cleared earlier,*
- *The length of the escalation phase is zero for small γ, increases for intermediate γ, and decreases for large γ.*

Our characterization yields several predictions about the dynamics of trust in the presence of secret undermining. First, optimal screening of disloyal agents entails discrete jumps in stakes. In particular, it is optimal to eventually clear the agent after a period where stakes are bounded away from their maximal level. Second, the extent to which the principal relies on discrete adjustment of stakes depends on the monitoring quality γ — the more difficult it is to detect undermining, the more stakes are adjusted through a discrete jump at a clearance event rather than through gradual escalation. Third, the agent is cleared later as monitoring quality falls. Fourth, despite the longer period needed to build trust, the length of the escalation phase need not always increase as monitoring becomes more difficult. Instead, the relationship between γ and the length of this phase is nonmonotonic.

In response to the stakes curve characterized in Theorem 1, the disloyal agent's optimal undermining policy is summarized in the following lemma:

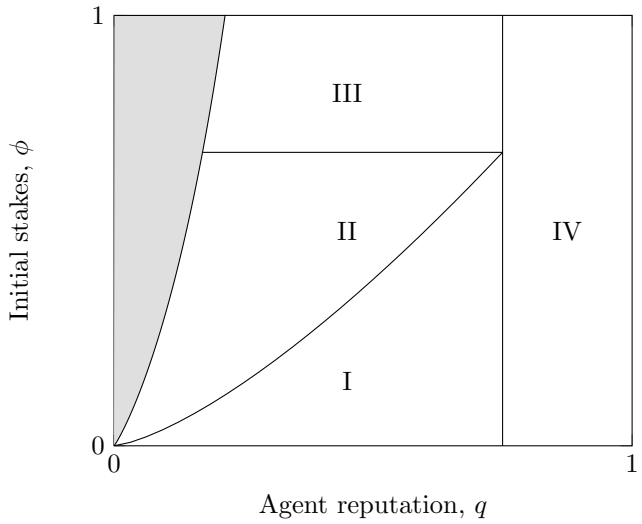
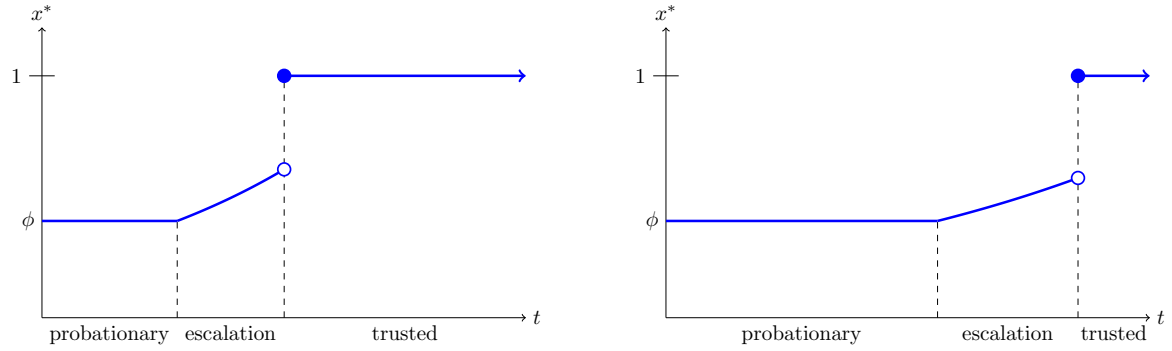


Figure 2: The behavior of optimal stakes in (q, ϕ)-space. Regions I, II, III represent the low-, moderate-, and high-stakes environments, while in region IV stakes are set to 1 immediately. In the shaded region, the optimal contract is degenerate.



(a) Larger γ

(b) Smaller γ

Figure 3: Phases of the optimal stakes curve.

Lemma 2. *Throughout the probationary and trusted phases, the disloyal agent strictly prefers to undermine and sets $\beta_t = 1$. During the escalation phase, the agent is indifferent between undermining or not, and any choice of $\beta_t \in [0, 1]$ is optimal. All undermining policies which are optimal for the agent yield the same profit for the principal.*

Under commitment, the disloyal agent's undermining policy is uniquely determined in the probationary and trusted phases. By contrast, our characterization is agnostic about the pattern of undermining during the escalation phase. In Section 6, we show that when the principal cannot commit to a stakes curve, the agent's undermining behavior is pinned down at all times.

3.3 A heuristic derivation

We now provide a sketch of the construction of the optimal contract. We first establish that without loss of generality, the principal pools all agents together into a single contract rather than separating loyal and disloyal agents into different contracts.

Lemma 3. *Given any obedient menu of contracts* $(\mathcal{C}^G, \mathcal{C}^B)$*, offering the contract* $\mathcal{C}^G$ *to both agent types weakly increases the principal's payoff.*

This result exploits the opposed objectives of the principal and disloyal agent. If the principal offers a menu of contracts which successfully screens the disloyal type into a different contract from the loyal one, she must be increasing the disloyal agent's payoff versus forcing him to accept the same contract as the one taken by the loyal agent. But increasing the disloyal agent's payoff decreases the principal's payoff, so it is always better to simply offer one contract and pool the disloyal agent with the loyal one. Thus screening of disloyal agents must be accomplished by direct detection of undermining rather than via self-reports by the agent.

We next establish the natural result that the principal optimally terminates the agent no later than the first time undermining is detected.

Definition 1. *A contract is* stringent *if it terminates the agent immediately following any detection of undermining.*

Lemma 4. *Given any contract* $\mathcal{C}$*, there exists a stringent contract* $\mathcal{C}'$ *such that the principal's payoff is weakly higher under* $\mathcal{C}'$ *than under* $\mathcal{C}$*, no matter the agent's type.*

Focusing on stringent contracts simplifies the description of a contract, as we need not specify how stakes and termination react to arbitrary histories of detected undermining. Every stringent contract can be described simply by a pair (x, T_F), where $(x_t)_{t\geq 0}$ is a stakes curve and T_F is an exogenous termination date (possibly infinite) at which the contract ends if no undermining is detected.

We now characterize the optimal stringent contract. An important property of this contract is that it makes the disloyal agent willing to undermine with full intensity at all times.

Definition 2. *A contract is a* loyalty test *if undermining with full intensity at all times is an optimal strategy for the disloyal agent.*

Any contract which is not a loyalty test can be modified to produce a superior contract. To understand why, consider an interval of time over which the disloyal agent strictly prefers

not to undermine. Over any such interval, the gains from remaining employed and exploiting higher future stakes must strictly outweigh the losses from failing to undermine the principal early on. The principal therefore has the freedom to raise stakes a bit during this interval without disturbing the disloyal agent's incentives. This modification hurts the disloyal agent, because he continues to refrain from undermining during a period with higher stakes than under the original contract. Therefore, because the principal and disloyal agent have opposed interests, the modification helps the principal when the disloyal agent is present. Further, this modification improves the principal's payoff when the loyal agent is present, as the loyal agent is more productive when stakes are higher. The modified contract therefore dominates the original one.

In light of this observation, designing an optimal stakes curve can be viewed as optimizing the principal's profits subject to the incentive constraint that the resulting curve is a loyalty test. To understand when this constraint binds, it is helpful to first consider a relaxed problem without the incentive constraint, in which the disloyal agent *exogenously* undermines with full intensity at all times. At each time t, supposing that the relationship has not yet been terminated, the principal receives flow payoffs x_t if the agent is loyal and $-x_t(K-1)$ if the agent is disloyal. Given a detection rate γ, the total probability the relationship has survived until time t is $q + (1-q)e^{-\gamma t}$. Let

$$\overline{\pi}_t \equiv \Pr(\theta = G \mid \beta = 1, \text{no detected undermining by time } t)$$

be the agent's time-t reputation, which is rising over time. Then the principal's total expected profits under the contract $\mathcal{C} = (x, T_F)$ are

$$V[\mathcal{C}] = \int_0^{T_F} e^{-rt} \left(q + (1-q)e^{-\gamma t}\right) \left(\overline{\pi}_t - (K-1)(1-\overline{\pi}_t)\right) x_t \, dt.$$

Flow profits are maximized at each time by setting $x_t = \phi$ whenever t is small enough that $\overline{\pi}_t < q^* \equiv (K-1)/K$, and setting $x_t = 1$ afterward. We will let

$$t^* \equiv \inf\{t \ : \ \overline{\pi}_t \geq q^*\}$$

denote the cutoff time at which the optimal stakes curve in the relaxed problem jumps to 1.

If the agent's initial reputation is at least q^*, then profits are maximized by setting stakes to their maximal level immediately. Under this policy, the disloyal agent trivially has no incentive to defer undermining, so the policy solves the unrelaxed problem as well. On the other hand, if the agent's initial reputation is below q^*, then $t^* > 0$ and it must be checked that the disloyal agent at least weakly prefers to undermine at full intensity prior to time

t^*. As might be expected, the disloyal agent is most strongly tempted to defer undermining just prior to time t^*, when his continuation payoff is largest relative to his flow payoff from undermining. Provided that the jump in stakes at time t^* is not too large, the disloyal agent is willing to undermine with full intensity at all times, and the solution to the relaxed problem is also the solution to the original problem.

To see when incentive compatibility holds, consider the disloyal agent's trade-off between undermining just before being cleared or waiting a moment longer. By undermining, the agent obtains a flow benefit of $K\phi\,dt$. On the other hand, the agent risks detection with probability $\gamma\,dt$. If the agent is detected, he is terminated and loses the opportunity for future undermining. Once stakes have reached their maximal level, the total expected value of undermining is $(K-1)/(r+\gamma)$, which reflects time discounting and the probability of detection. Undermining is therefore optimal if and only if

$$K\phi\,dt \geq \gamma\,dt \cdot \frac{K-1}{r+\gamma}.$$

Note that because undermining is only occasionally detected in our setting of imperfect monitoring, both the gains and losses from a particular act of undermining are of the same order, even near a jump in stakes. Discontinuities in the stakes curve can therefore coexist with incentive compatibility, supposing that they are not too large.

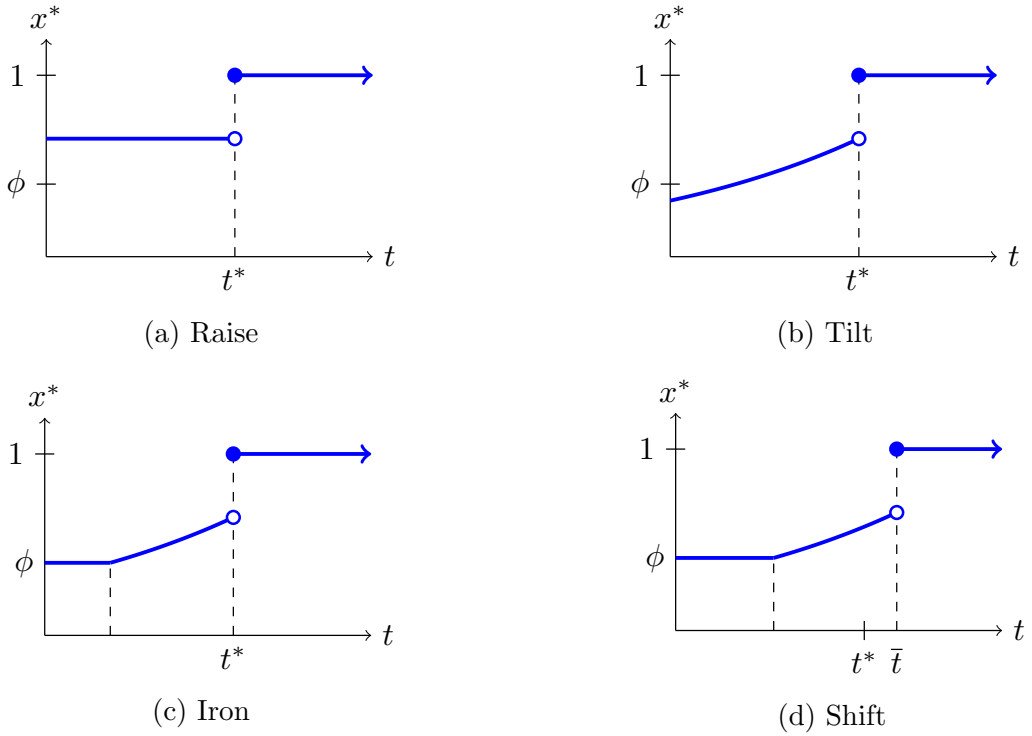
The condition just derived is satisfied for ϕ sufficiently close to 1. For smaller values of ϕ, however, incentive compatibility fails when the agent is close to being cleared. In this case, stakes must rise prior to clearance in order to decrease the size of the jump at clearance and preserve the loyalty test property. Intuitively, two changes must be made to the stakes curve. First, the jump must be made small enough that the agent does not prefer to feign loyalty just prior to the jump. Second, stakes should rise from their initial level as late and as quickly as possible, to ensure that flow losses are minimized when the agent is untrusted.

This modification to the ideal stakes curve can be visualized graphically — the constant stakes level prior to time t^* is shifted upward to ensure undermining is optimal near time t^*, and is then tilted counterclockwise until incentive-compatibility binds. Depending on the value of ϕ, this tilt procedure may produce a stakes curve which violates the lower bound early in the relationship. As a remedy, the principal optimally irons the stakes curve so that it does not drop below ϕ.

This "raise-and-tilt" procedure holds fixed the clearance date at t^*. In fact, the principal will in general also wish to delay clearance until some time $\bar{t} > t^*$. Increasing $\bar{t}$ reduces stakes, and therefore losses, early in the relationship (when the agent's reputation is below q^*), but also reduces stakes and profits late in the relationship (when his reputation is above

q^*). When the loyalty test constraint binds and the clearance date is exactly t^*, shifting delaying clearance entails no first-order losses, as flow profits just to the right of t^* are approximately zero. On the other hand, this shift yields a first-order increase in profits early in the relationship. The unique optimal stakes curve sets $\bar{t} > t^*$ to balance the gains and losses from delaying clearance. Figure 4 illustrates this procedure.

This construction of the optimal stakes curve ensures that undermining at all times is an optimal policy for the disloyal agent. However, it is not necessarily the unique optimal policy. Under the stakes curve x^*, the incentive constraint binds during the escalation phase, when the principal would like to raise stakes as quickly as possible. Therefore, the agent is indifferent over undermining during the escalation phase. Due to the zero-sum nature of the interaction between principal and disloyal agent, none of these policies is privileged from the point of view of maximizing principal profits.



(a) Raise (b) Tilt

(c) Iron (d) Shift

Figure 4: Heuristic derivation of the optimal stakes curve.

The comparative statics presented in Lemma 1 can be understood in light of the preceding discussion. As γ increases, the agent's incentive compatibility constraint just prior to clearance tightens, and so the principal must reduce the jump in stakes at clearance. Additionally, the disloyal agent is weeded out more quickly, and so he is optimally cleared at an earlier date. The reduced jump at clearance suggests that a larger γ should lengthen the escalation phase; however, improved monitoring also raises the growth rate of stakes during

the escalation phase. The net effect of these opposing forces is a nonmonotonicity in the length of the escalation phase with respect to monitoring quality.

4 Application to espionage investigations

The analysis of Section 3 predicts that a principal who can only imperfectly monitor its agents, but who can also adjust the stakes of an agent's assignment, will subject an agent of uncertain loyalty to a dynamic screening process. The agent will be initially placed in a low- but positive-stakes assignment and monitored for evidence of undermining for an extended period, following which he will be discretely promoted to a high-stakes assignment provided that no undermining is detected. These predictions are broadly consistent with observed practices from espionage investigations, providing a rationale for these practices and validating our model in an important application.

Espionage investigations are a good fit for our model because they involve employees in high-stakes assignments whose activities are imperfectly monitored for direct evidence of espionage. According to an analysis of United States government espionage cases by Herbig (2017),[7] 63% of espionage cases involved a spy who was at least 30 when they began spying, while 33% were 40 or older; and in 45% of cases, the spy held a Top Secret (TS) or Top Secret-Sensitive Compartmentalized Information (TS-SCI) clearance at the outset of their espionage career.[8] Many spies are therefore relatively senior with access to high-stakes information. Moreover, 56% of spies operated for at least a year, and 27% operated for 5 or more years.[9] These statistics suggest imperfect monitoring of espionage activities and the opportunity to act repeatedly.

Additionally, once a suspect has been identified through analysis of existing circumstantial evidence, espionage investigations typically focus on obtaining "smoking-gun" proof of an act of espionage, for instance by witnessing an interaction with a handler or a dead drop of information. These investigations often last months to years, and can take time to successfully witness an incriminating act. This investigative process, featuring definitive positive signals obtainable with some probability only when an act occurs, closely matches our model.[10]

[7] This analysis is based on a database of successful espionage investigations since World War 2 maintained by the United Stated government and available at https://www.dhra.mil/perserec/espionage-cases/.

[8] See Herbig (2017), Table 1 ("Age when espionage began") and Table 2 ("Security clearance when espionage began").

[9] See Herbig (2017), Table 6 ("Duration"). These percentages were computed by summing the number of occurrences of each outcome across all three time periods, and then dividing by the total number of cases across time periods.

[10] In some cases, suspected spies have been arrested on the basis of suspicious but inconclusive actions.

As predicted by our model, suspected spies who become the target of an espionage investigation are often subjected to an extended trial period with limited stakes. For instance, Robert Philip Hanssen, a senior FBI agent and Soviet/Russian spy active from 1979 to 2001, was transferred to FBI headquarters for the purpose of restricting his access to sensitive information (Wise, 2003, p. 237).[11] Harold Nicholson, a CIA agent and Russian spy active from 1994 to 1996, was transferred from a post at the CIA's officer training school to agency headquarters to limit his access to information (Lonergan, ????). Paul Garbler, a former CIA station chief in Moscow and suspect during a 1960s CIA mole hunt, was reassigned to a series of increasingly marginal posts within the CIA (Wise, 1992, Ch. 14). And David Murphy, a former head of the CIA Soviet division and another suspect during the mole hunt, was demoted to Paris station chief (Wise, 1992, Ch. 15).

Our model also predicts that suspects should not be completely cut off from sensitive information. Consistent with this prediction, in all of the cases mentioned above, the suspect remained in a position of some responsibility. In particular, both Robert Hanssen and Harold Nicholson retained sufficient access to information that they attempted to pass additional classified documents to their handlers after being reassigned.

Finally, our model predicts that suspects not caught conducting espionage should eventually be cleared and promoted to assignments of significantly greater responsibility. Such clearance events have indeed occurred. Paul Garbler eventually managed to convince his superiors to "partially rehabiliat[e]" him by assigning him to a station chief post in Stockholm (Wise, 1992, p. 208). Similarly, David Murphy was eventually promoted to a senior management role, an action the CIA director at the time described as deliberately " 'taking a chance, putting him into a highly sensitive activity' " (Wise, 1992, p. 226).

5 Robustness

In this section, we analyze a number of extensions of our baseline model. In Section 5.1, we endogenize the intensity of monitoring. In Section 5.2, we introduce a wedge between the agent's benefit from undermining and the harm inflicted on the principal. In Section 5.3, we allow the agent to accrue private benefits from continued employment, for instance due to

For instance, Aldrich Ames, a senior CIA agent and Soviet spy active from 1985 to 1993, was arrested on the basis of incriminating documents found in Ames's home and recorded phone calls between Ames and his wife (United States Senate Select Committee on Intelligence, ????, p. 51-52). Our results are robust to inconclusive evidence of this sort, so long as the evidence is informative enough to make continued screening unprofitable for the principal.

[11] These restrictions were emphasized by Hanssen in a message to his Russian handlers, where he complained that his new posting was a "do-nothing... job outside of regular access to information within the counterintelligence program" (Wise, 2003, p. 238).

career concerns. In Section 5.4, we allow the principal to impose punishments on the agent following detected undermining. A robust finding across all of these settings is that the loyal agent's optimal stakes curve exhibits the same qualitative features as in the baseline model.

5.1 Endogenous monitoring

We have so far assumed that undermining is detected exogenously, without any effort by the principal. In some contexts, the principal may need to devote costly resources toward detecting undermining. For instance, in the espionage investigations highlighted in Section 4, investigations require significant manpower to conduct. We now consider how our results change when monitoring is costly and its intensity can be varied over time. Our main finding is that, so long as monitoring costs are sufficiently small, all qualitative features of the loyal agent's optimal stakes curve remain unchanged under endogenous monitoring. Additionally, the principal eventually stops monitoring the agent once enough time has passed. We present an informal discussion of our findings here, and defer a formal analysis to the online appendix.

We consider a simple linear cost specification: at each moment in time, the principal can commit to a monitoring intensity $\gamma_t \in [0, \overline{\gamma}]$, where $\overline{\gamma}$ is an exogenous upper bound on available resources. To monitor at intensity γ_t, the principal must pay a flow cost of $c \cdot \gamma_t$, where $c > 0$ is the marginal cost of monitoring. We assume that the principal's monitoring costs do not enter the disloyal agent's utility function.[12] For simplicity, we restrict attention to ϕ small enough that the lower bound on stakes is nonbinding. Under endogenous monitoring, a contract consists of a stakes curve, a termination policy, and a monitoring intensity at each point in time.

When monitoring is endogenous, the principal optimally offers the loyal agent a contract which proceeds through escalation and trusted phases qualitatively similar to those in the baseline model. During the escalation phase stakes rise continuously, while in the trusted phase stakes are set to their maximal level. (Because ϕ is small, there is no probationary phase.) Meanwhile, monitoring proceeds through up to three distinct phases: during an initial *vigilant phase*, the principal monitors at the maximal intensity $\overline{\gamma}$; eventually, monitoring enters a *tapering phase*, in which the monitoring intensity declines smoothly; at the end of the tapering phase, the monitoring intensity jumps downward to zero and remains there forever after, during a *complacent phase*.[13] The vigilant monitoring phase lasts at least as long as the escalation phase, so that the agent is monitored at full intensity until stakes

[12] Qualitatively similar features would emerge if the disloyal agent gained utility from the principal's expenditures on monitoring.

[13] The vigilant and complacent phases always occur, while the tapering phase occurs only when $\overline{\gamma}$ is sufficiently large; otherwise, monitoring proceeds directly from the vigilant to the complacent phase.

reach their maximal level.

A key property of this contract is that, when c is sufficiently small (but still positive), stakes jump upward discontinuously at the end of the escalation phase. That is, our prediction of a jump in stakes is robust to small costs of monitoring. This bound on costs need not be very stringent: when the maximal monitoring intensity is sufficiently large, a jump in stakes is part of an optimal contract whenever the principal uses monitoring at all. (When monitoring costs are very high, monitoring becomes an ineffective screening device.)

In addition to the contract outlined above, the principal offers a second *screening contract* intended for the disloyal agent. The key property of this contract is that it never monitors the agent. One optimal screening contract sets a fixed stakes level forever, with the level calibrated to deliver the agent the same level of utility as he would enjoy under the loyal agent's contract.

The basic rationale for screening contracts is that the principal would like to lower the disloyal agent's promised utility by threatening to monitor, but prefers not to actually expend monitoring costs when the disloyal agent is present. Were both agents pooled into one contract, these two goals would be in tension. With two contracts, the principal can discourage the disloyal agent from taking the loyal agent's contract by monitoring, while screening the disloyal agent into a low-stakes contract with no monitoring.

5.2 Negative-sum undermining

Our baseline model assumes that the disloyal agent gains as much from undermining as the principal loses. In some contexts, it may be more realistic to model the principal's (relative) losses as greater than the agent's gains. For instance, the agent's utility from undermining might derive in part from compensation by a third party (e.g., payment for pilfered documents), with such payments being only a fraction of the damage inflicted upon the principal. Alternatively, detected undermining might be exogenously punished by criminal prosecution, with the agent receiving a penalty that reduces the gains from undermining.[14] We now show that our main results are unchanged under such negative-sum undermining.

We model the misalignment between agent gains and principal losses by assuming that the agent gains a flow benefit of $K_A x_t$ from undermining when current stakes are x_t, while the principal incurs flow losses of Kx_t, as in the baseline model. We assume that $K > K_A > 1$, so that the losses imposed on the principal are greater than the gains for the agent (relative to the flow gains/losses from task performance). The following proposition characterizes an

[14]More precisely, the payoff specification in this extension can be microfounded by assuming that detected undermining leads the agent to incur an exogenous punishment, such as from criminal prosecution, whose magnitude is proportional to the crime, i.e., to the level of stakes.

optimal contract in this environment. (Its proof can be found in the online appendix.)

Proposition 1. *For any $K_A \in (1, K)$, the principal optimally offers a single contract chosen by both agent types. When this contract is non-degenerate, stakes proceed through (up to) three phases as in Theorem 1. During the escalation phase, stakes grow at rate $r + \gamma/K_A$, and the principal optimally recommends undermining with full intensity.*

In most respects, the optimal contract is qualitatively unchanged relative to the baseline model. The one notable difference is that, while the agent continues to be indifferent over undermining during the escalation phase, the principal strictly prefers that the agent undermine as intensively as possible. This preference stems from the fact that the interaction between the principal and disloyal agent is now negative-sum rather than zero-sum, and the wedge between the agent's gains and principal's losses grows larger as stakes grow. As a result, the principal would prefer that the disloyal agent earn his promised utility by undermining when this wedge is small.

5.3 Gains from trade

Our main analysis assumes that the disloyal agent's interests are diametrically opposed to those of the principal. In some contexts, this misalignment may not be complete. In particular, the agent may accrue benefits from continued employment independent of the harm inflicted on the principal, for instance due to reputational benefits of employment or deferred job search costs following termination. Our model can be extended to accommodate such benefits, and our main predictions about stakes dynamics continue to hold in their presence. We present an informal discussion of our findings here, and defer a formal analysis to the online appendix.

We modify our model to accommodate gains from employment by assuming that, as long as the agent is employed, he collects a stream of private gains $g > 0$ regardless of his loyalty.[15] These gains accrue above and beyond any stakes-related payoffs the agent receives. Crucially, they are not incurred at a cost to the principal, and so they represent gains from trade from employment. As a result, the relationship between the principal and a disloyal agent is positive-sum, in contrast to the zero-sum relationship of the baseline model. To streamline our analysis, we assume that ϕ is small enough that the lower bound on stakes is non-binding. We also impose a regularity condition that the discount rate r is not too large.

We find that under gains from trade, the principal offers the loyal agent a contract very similar to the optimal contract of the baseline model. This contract features a gradual

[15] Nothing would change if the size of the flow benefit depended on the agent's loyalty.

escalation phase, followed by graduation to trusted status marked by a jump in stakes to their maximal level. (Because ϕ is small, there is no probationary phase.) Under this contract, detected undermining is punished by immediate termination. In addition to this contract, the principal offers a second *screening contract* intended for the disloyal agent. This screening contract keeps the disloyal agent employed forever (when the agent follows his recommended undermining policy), in order to maximize the gains from trade from employment.

The basic rationale for screening contracts is similar to the endogenous-monitoring results of Section 5.1: The principal would like to lower the disloyal agent's promised utility by threatening termination following undermining, but prefers not to actually destroy gains from trade by acting on this threat. With two contracts, the principal can discourage the disloyal agent from taking the loyal agent's contract by threatening termination, while simultaneously enticing him to take his intended contract by maximizing his private gains from employment.

One optimal screening contract coincides with the loyal agent's contract until the loyal agent's graduation date, at which point the disloyal agent continues through an extended escalation phase. The disloyal agent's reward for delaying graduation is that he eventually reaches a *permissive phase* where undermining is punished not with termination, but with permanent *reassignment* to a low-stakes task forever. The stakes following reassignment are set sufficiently low that the disloyal agent is willing to forego all further undermining, motivated by the prospect of a stream of private gains.

As in the baseline model, the optimal stakes curve keeps the disloyal agent indifferent about undermining whenever stakes are below their maximal level. However, because the model with gains from trade is not zero-sum, the agent's indifference does not translate into indifference by the principal. Instead, the principal optimally recommends that the disloyal agent defer all undermining until the permissive phase. This recommendation maximizes the fraction of the disloyal agent's utility derived from private gains rather than undermining, minimizing the principal's cost of delivering the disloyal agent's promised utility. Only during the permissive phase does the disloyal agent undermine the principal.

5.4 Punishments

Our main analysis supposes that the principal cannot impose any punishments on the agent beyond termination. It may instead be the case that the principal can inflict additional harm on the agent beyond termination, for instance by reporting him to law enforcement for criminal prosecution. Such punishments do not directly benefit the principal, but could

be useful for disciplining the agent and discouraging undermining.[16] We now discuss how the results of our model change if the principal can commit to punishments up to some maximum size. As the analysis is very similar to the gains-from-trade extension of Section 5.3, we omit a formal development.

When punishments are relatively weak, the outcome is very similar to the one arising under gains from trade. Intuitively, the principal would like to threaten to punish the disloyal agent in order to reduce his promised utility, but would prefer not to destroy surplus by actually imposing the punishments. As a result, when the maximal punishment size is small, the principal optimally offers two contracts—one for the loyal agent in which all undermining is punished as harshly as possible, and another for the disloyal agent in which punishments are threatened during the escalation phase but not during the permissive phase. The two contracts follow the same stakes paths until the graduation date for the loyal agent, following which the disloyal agent is subjected to a longer escalation phase and delayed graduation. Notably, the disloyal agent is recommended not to undermine during the escalation phase, and so by following his recommended undermining policy he avoids punishment under the contract intended for him.

On the other hand, if punishments are sufficiently powerful, the principal offers a single contract and the disloyal agent is screened out of employment entirely. This is because when punishments are strong, the disloyal agent is guaranteed a negative utility from employment under an optimal contract, and declines employment voluntarily. In this regime, the optimal contract is qualitatively similar to the baseline model, with detected undermining punished as harshly as possible. This contract screens out disloyal agents through a combination of gradual stakes escalation and the threat of punishment. In the limit of very strong punishments, the disloyal agent can be screened by the threat of punishment even when stakes are set to their maximal level, and screening through stakes is no longer necessary.

6 The no-commitment outcome

While the commitment outcome sets a benchmark for efficient screening, in high-stakes applications the principal may have difficulty committing to future stakes. In this section we study what outcomes are possible without commitment.[17] We show that the stakes curve implemented in the absence of commitment is exactly the optimal commitment stakes curve, demonstrating robustness of our results to relaxation of the commitment assumption.

[16] Our results would be qualitatively unchanged if the principal benefited from punishments imposed on the agent, for instance if the punishment took the form of a financial penalty accruing to the principal.

[17] Throughout this section, we assume that q is sufficiently high that the degenerate contract is not optimal under commitment. Otherwise, the optimal outcome is trivially implementable without commitment.

Further, we show that equilibrium play uniquely determines disloyal agent behavior at all times, resolving the indeterminacy arising during the escalation phase in the commitment benchmark.

The following proposition states the main result of the section. It characterizes the unique sequence of stakes and undermining that can arise along the equilibrium path of any pure strategy Bayes Nash equilibrium.

Proposition 2. *Suppose q is large enough that the optimal commitment contract is non-degenerate. Let x^* be the stakes path characterized in Theorem 1. Then:*

- *x^* is the unique stakes path arising under any (pure-strategy) Bayes Nash equilibrium,*
- *There exists a unique undermining policy β^* arising under any Bayes Nash equilibrium, which satisfies the following properties:*
  - *$\beta_t^* = 1$ during the probationary and trusted phases*
  - *β^* is strictly increasing during the escalation phase*
  - *β^* is continuous except at the beginning of the escalation phase*
- *There exists a perfect Bayesian equilibrium with equilibrium stakes path x^* and undermining policy β^*.*

This proposition has several components. First, it establishes that the optimal commitment stakes path x^* is also the unique equilibrium stakes path arising in any Bayes Nash equilibrium (i.e., without any requirement of sequential rationality). The intuition is as follows. The principal can guarantee herself at least her payoff under commitment, regardless of the (disloyal) agent's strategy, simply by choosing the stakes path x^*. This is because the principal's optimal contractual profits are calculated assuming the agent responds by choosing a strategy which minimizes the principal's profits given x^*. Therefore, no matter what strategy the agent actually follows in equilibrium, the principal must do at least as well as under commitment. On the other hand, the principal's equilibrium profits cannot exceed the commitment payoff, since any equilibrium outcome can be replicated by an appropriate contract. This argument establishes that the principal receives the same payoff with and without commitment. Since Theorem 1 shows that x^* is the unique stakes path achieving this profit, it must be the equilibrium stakes path.

Second, the proposition pins down a unique undermining path which is consistent with equilibrium. Optimality for the agent requires undermining with full intensity during the probationary and trusted phases, but imposes no restriction on undermining behavior during the escalation phase. However, most undermining policies do not induce x^* as a best

response by the principal. Intuitively, the equilibrium rate of undermining controls the proportion of loyal agents remaining in the relationship over time. The more intensively the principal expects the agent to undermine, the more quickly her beliefs about the agent's loyalty rise absent a discovery of undermining. And the level of the principal's beliefs impact her incentives to deviate from x^* either by raising stakes more quickly, to favor an ex post trustworthy agent, or by lowering them, to shield herself from an ex post suspicious agent.

It turns out that under stakes curve x^*, there is exactly one undermining policy which raises the principal's beliefs at a rate such that she continues to prefer following x^* at all times. This undermining policy is pinned down by the defining property that it induces an expected flow payoff of zero for the principal throughout the escalation phase. During this period, the disloyal agent's undermining intensity gradually rises, reaching full intensity at graduation to trusted status. The shape of β^* is plotted in Figure 5. A notable feature is that in a moderate-stakes environment, equilibrium undermining intensity is nonmonotonic, exhibiting a downward jump at the beginning of the escalation phase.

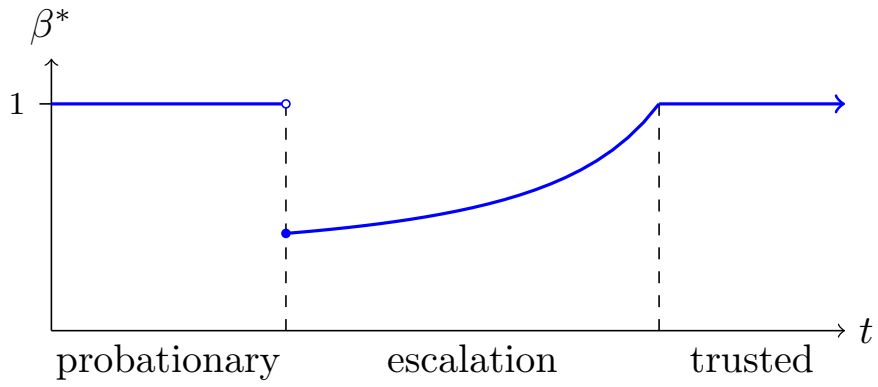


Figure 5: The time-consistent undermining path.

Finally, the proposition establishes that there exists a perfect Bayesian equilibrium supporting (x^*, β^*) as its equilibrium path. This step amounts to constructing a full contingent undermining strategy for the agent, conditioning on the principal's past behavior, which supports the equilibrium path and further is part of an equilibrium in every continuation game. The equilibrium is Markovian in the principal's posterior belief p about the agent's type: Whenever the state is p, on or off the equilibrium path, the continuation strategies yield the same equilibrium path of stakes and undermining in the continuation game.

Under this strategy profile, the agent's undermining is unresponsive to the history of stakes. This does not lead to incentive-compatibility problems for the agent, as he expects that the principal will immediately correct any deviation from equilibrium stakes. The principal, on the other hand, may be tempted to deviate from equilibrium stakes in order to

exploit the stubbornness of the agent's undermining strategy. The principal cannot profit from such deviations due to the defining property of the undermining path β^* mentioned above: Since her expected flow payoff is identically zero in this phase, all stakes levels yield her the same flow payoffs. The principal also receives no long-run benefit from deviating: Since the agent's undermining intensity does not respond to the history of stakes, the principal cannot affect the rate at which she screens out disloyal agents by deviating. As a result, this strategy profile is sequentially rational, inducing a perfect Bayesian equilibrium.

7 Conclusion

This paper studies dynamic screening of disloyal agents in a variable-stakes principal-agent problem, when higher stakes both facilitate efficient performance of the agent's job and enhance the harm of undermining. We show that under commitment, the optimal stakes path features a probationary phase of fixed stakes, an escalation phase with a smooth rise in stakes, and a deterministic date at which the agent is deemed trusted and stakes jump discretely to their maximum level. This jump in stakes is a signature feature of an optimal stakes process in our setting, distinguishing our results from outcomes in existing models of variable-stakes relationships with discrete betrayal actions.

Our main findings are driven by a mechanism which is robust to a number of departures from our baseline assumptions. The optimal stakes curve is the unique equilibrium path of stakes in an environment without commitment. Additionally, the loyal agent's optimal stakes curve is qualitatively unchanged in environments with flexible monitoring, negative-sum undermining, gains from trade, or punishments.

Our analysis leaves open several important questions for future work. We have focused on the relationship between a principal and a single agent who is commonly known to be under suspicion. In some circumstances, the agent may be unsure if he is being investigated, or how many resources the organization has devoted to monitoring him. Additionally, the organization may have several suspects which it needs to investigate either in parallel or in sequence. Enhancing our model to allow for multiple suspects and uncertainty over who the organization is currently investigating, and how intensively, would provide further insights into how organizations root out disloyal elements.

We have also assumed that the agent's loyalty is exogenous. In practice, agents may be willing to support certain goals or strategies, but not others. Combining our model of screening for loyalty with a project-selection problem could shed light on the process by which organizations make and build consensus on major decisions.

References

Albuquerque, R. and H. A. Hopenhayn (2004): "Optimal Lending Contracts and Firm Dynamics," *The Review of Economic Studies*, 71, 285–315.

Atakan, A., L. Koçkesen, and E. Kubilay (2020): "Starting small to communicate," *Games and Economic Behavior*, 121, 265–296.

Boleslavsky, R. and M. Said (2013): "Progressive screening: Long-term contracting with a privately known stochastic process," *Review of Economic Studies*, 80, 1–34.

Chaves, I. N. (2020): "Privacy in Bargaining: The Case of Endogenous Entry," Unpublished.

Daley, B. and B. S. Green (2020): "Bargaining and News," *American Economic Review*, 110, 428–474.

Fudenberg, D., G. Georgiadis, and L. Rayo (2021): "Working to learn," *Journal of Economic Theory*, 197, 105347.

Fudenberg, D. and L. Rayo (2019): "Training and Effort Dynamics in Apprenticeship," *The American Economic Review*, 109, 3780–3812.

Gul, F., H. Sonnenschein, and R. Wilson (1986): "Foundations of dynamic monopoly and the coase conjecture," *Journal of Economic Theory*, 39, 155–190.

Herbig, K. L. (2017): "The expanding spectrum of espionage by Americans, 1947-2015," Tech. rep., Defense Personnel and Security Research Center.

Kreps, D. (2018): "Starting Small to Screen for Mr. Good Bob," Unpublished.

Lonergan, M. (????): "Affidavit in support of complaint, arrest warrant and search warrants: United States v. Harold James Nicholson," .

Rayo, L. and L. Garicano (2017): "Relational Knowledge Transfers," *The American Economic Review*, 107, 2695–2730.

Sobel, J. (1985): "A Theory of Credibility," *The Review of Economic Studies*, 52, 557–573.

Stokey, N. L. (1981): "Rational expectations and durable goods pricing," *The Bell Journal of Economics*, 112–128.

Teschl, G. (2012): *Ordinary differential equations and dynamical systems*, vol. 140, American Mathematical Soc.

Thomas, J. and T. Worrall (1994): "Foreign Direct Investment and the Risk of Expropriation," *Review of Economic Studies*, 61, 81–108.

United States Senate Select Committee on Intelligence (????): "An Assessment of the Aldrich H. Ames Espionage Case and its Implications for U.S. Intelligence," .

Watson, J. (1999): "Starting small and renegotiation," *Journal of Economic Theory*, 85, 52–90.

——— (2002): "Starting small and commitment," *Games and Economic Behavior*, 38, 176–199.

Wise, D. (1992): *Molehunt: the secret search for traitors that shattered the CIA*, Random House Incorporated.

——— (2003): *Spy: The inside story of how the FBI's Robert Hanssen betrayed America*, Random House Incorporated.

A Technical results

In this appendix we prove several technical lemmas aiding in the characterization of incentive-compatible undermining policies. These facts will be used at several points in proofs of results appearing in the main text.

Given a stringent contract $\mathcal{C} = (x, T_F)$, let $U^\beta[\mathcal{C}]$ be the disloyal agent's continuation value process under the undermining policy β, conditional on no undermining having been detected yet. That is,

$$U^\beta[\mathcal{C}]_t = \int_t^{T_F} \exp\left(-r(s-t) - \gamma \int_t^s \beta_u \, du\right) (K\beta_s - 1) x_s \, ds$$

for all $t \leq T_F$, with $U_t^\beta = 0$ for $t > T_F$. This function is absolutely continuous with a.e. derivative

$$\frac{d}{dt} U^\beta[\mathcal{C}]_t = (r + \gamma\beta_t) U_t^\beta - (K\beta_t - 1) x_t, \quad t \leq T_F.$$

Note that the rhs is bounded below by $f(U^\beta[\mathcal{C}]_t, t)$, where

$$f(u,t) \equiv \min\{(r+\gamma)u - (K-1)x_t, ru + x_t\} = \min_{\beta' \in [0,1]} \{(r + \gamma\beta')u - (K\beta' - 1)x_t\}.$$

The following lemma is a comparison theorem tailored to our setting.

Lemma A.1. *Suppose $g(u,t)$ is a function which is strictly increasing in its first argument. Suppose there exist two absolutely continuous functions u_1 and u_2 on $[0,T]$ such that $u_1(0) \geq u_2(0)$ while $u_1'(t) = g(u_1(t),t)$ and $u_2'(t) \leq g(u_2(t),t)$ on $[0,T]$ a.e. Then:*

1. *$u_1 \geq u_2$.*

2. *If in addition $u_1(0) = u_2(0)$ and $u_2'(t) = g(u_2(t),t)$ on $[0,T]$ a.e., then $u_1 = u_2$.*

Proof. Define $\Delta(t) \equiv u_2(t) - u_1(t)$, and suppose by way of contradiction that $\Delta(t_0) > 0$ for some $t_0 \in (0,T]$. Let $t_1 \equiv \sup\{t < t_0 \ : \ \Delta(t) \leq 0\}$. Given continuity of Δ, it must be that $t_1 < t_0$. Further, $t_1 \geq 0$ given $u_1(0) \geq u_2(0)$. And by continuity $\Delta(t_1) = 0$. But also by the fundamental theorem of calculus

$$\Delta(t_0) = \Delta(t_1) + \int_{t_1}^{t_0} \Delta'(t)\, dt.$$

Now, given that $\Delta(t) > 0$ on (t_1, t_0), it must be that

$$\Delta'(t) = u_2'(t) - u_1'(t) \leq g(u_2(t),t) - g(u_1(t),t) < 0$$

a.e. on (t_1, t_0) given that g is strictly increasing in its first argument. Hence from the previous identity, $\Delta(t_1) > \Delta(t_0)$, contradicting $\Delta(t_1) = 0$ and $\Delta(t_0) > 0$. So it must be that $\Delta \leq 0$, i.e. $u_2 \leq u_1$.

Now, suppose further that $u_1(0) = u_2(0)$ and $u_2'(t) = g(u_2(t),t)$ on $[0,T]$ a.e. Trivially $u_2(0) \geq u_1(0)$ and $u_1'(t) \leq g(u_1(t),t)$ on $[0,T]$ a.e., so reversing the roles of u_1 and u_2 in the proof of the previous part establishes that $u_1 \leq u_2$. Hence $u_1 = u_2$. □

The next lemma provides equivalent characterizations of an optimal undermining policy.

Lemma A.2. *Fix a stringent contract $\mathcal{C} = (x, T_F)$, and let β^* be an arbitrary undermining policy. The following are equivalent:*

(i) β^* maximizes the disloyal agent's payoff under $\mathcal{C}$.

(ii) For almost all $t \in [0,T_F]$, $dU^{\beta^*}[\mathcal{C}]_t/dt = f(U^{\beta^*}[\mathcal{C}]_t, t)$.

(iii) For almost all $t \in [0,T_F]$,

$$U^{\beta^*}[\mathcal{C}]_t < Kx_t/\gamma \implies \beta_t^* = 1 \tag{A.1}$$

$$U^{\beta^*}[\mathcal{C}]_t > Kx_t/\gamma \implies \beta_t^* = 0. \tag{A.2}$$

Proof. To economize on notation, we drop explicit conditioning on $\mathcal{C}$ throughout this proof. To establish (ii) $\implies$ (i), let β^* be an undermining policy whose continuation value process U^{β^*} satisfies $dU_t^{\beta^*}/dt = f(U_t^{\beta^*}, t)$ a.e. on $[0, T_F]$. Let $\tilde{\beta}$ be an arbitrary undermining policy, and let τ be the random time with hazard rate $\gamma\tilde{\beta}$. Then for all finite $t \leq T_F$,

$$e^{-rt}U_t^{\beta^*}\mathbf{1}\{\tau > t\} = U_0^{\beta^*} + \int_0^t (-rU_s^{\beta^*} + dU_s^{\beta^*}/ds)e^{-rs}\mathbf{1}\{\tau > s\}ds - \sum_{s\leq t} e^{-rs}U_s^{\beta^*}\mathbf{1}\{\tau = s\}.$$

Define $Q_t^{\tilde{\beta}} \equiv \exp\left(-\gamma \int_0^t \tilde{\beta}_s ds\right) > 0$. Then, taking expectations, we have

$$e^{-rt}U_t^{\beta^*}Q_t^{\tilde{\beta}} = U_0^{\beta^*} + \int_0^t e^{-rs}Q_s^{\tilde{\beta}}\left(-rU_s^{\beta^*} + \frac{dU_s^{\beta^*}}{ds}\right) ds - \int_0^t e^{-rs}U_s^{\beta^*}\gamma\tilde{\beta}_s Q_s^{\tilde{\beta}}\, ds.$$

As U^{β^*} and $Q^{\tilde{\beta}}$ are bounded uniformly in time, taking $t \to T_F$ yields

$$U_0^{\beta^*} = \int_0^{T_F} e^{-rs}Q_s^{\tilde{\beta}}\left[(\gamma\tilde{\beta}_s + r)U_s^{\beta^*} - \frac{dU_s^{\beta^*}}{ds}\right] ds \geq \int_0^{T_F} e^{-rs}Q_s^{\tilde{\beta}}(K\tilde{\beta}_s - 1)x_s\, ds, \tag{A.3}$$

where we have used that $\lim_{t\to T_F} e^{-rt}U_t^{\beta^*} = 0$[18] and that $dU_t^{\beta^*}/dt = f(U_t^{\beta^*}, t) = \min_{\beta'\in[0,1]}\{(r+\gamma\beta')U_t^{\beta^*} - (K\beta' - 1)x_t\}$ on $[0, T_F]$ a.e. The right hand side of (A.3) is the agent's expected payoff from the policy $\tilde{\beta}$, and since $\tilde{\beta}$ is arbitrary, β^* is optimal.

To prove that (i) $\implies$ (ii), recall that $dU_t^{\beta^*}/dt \geq f(U_t^{\beta^*}, t)$ on $[0, T_F]$ a.e., and suppose (in negation of (ii)) that $dU_t^{\beta^*}/dt > f(U_t^{\beta^*}, t)$ for all finite t in a positive-measure subset of $[0, T_F]$. Define $\tilde{\beta}_t = \arg\min_{\beta'\in[0,1]}\{(r + \gamma\beta')U_t^{\beta^*} - (K\beta' - 1)x_t\}$ for all $t \in [0, T_F]$. Then $dU_t^{\beta^*}/dt \geq (r + \gamma\tilde{\beta}_t)U_t^{\beta^*} - (K\tilde{\beta}_t - 1)x_t$ for all $t \in [0, T_F]$, with strict inequality over a positive-measure set. Repeating the calculation performed in the proof that (ii) $\implies$ (i),

$$U_0^{\beta^*} = \int_0^{T_F} e^{-rs}Q_s^{\tilde{\beta}}\left[(\gamma\tilde{\beta}_s + r)U_s^{\beta^*} - \frac{dU_s^{\beta^*}}{ds}\right] ds < \int_0^{T_F} e^{-rs}Q_s^{\tilde{\beta}}(K\tilde{\beta}_s - 1)x_s\, ds,$$

so the agent obtains a strictly higher expected utility from the policy $\tilde{\beta}$ than from β^*; that is, β^* does not maximize the disloyal agent's payoff.

To establish (ii) $\iff$ (iii), recall that $dU_t^{\beta^*}/dt = (r + \gamma\beta_t^*)U_t^{\beta^*} - (K\beta_t^* - 1)x_t$ on $[0, T_F]$ a.e. Thus, (ii) holds iff $(r+\gamma\beta_t^*)U_t^{\beta^*} - (K\beta_t^* - 1)x_t = \min\{(r+\gamma)U_t^{\beta^*} - (K-1)x_t, rU_t^{\beta^*} + x_t\}$ on $[0, T_F]$ a.e. When $U_t^{\beta^*} < Kx_t/\gamma$, this is equivalent to $\beta_t^* = 1$, and when $U_t^{\beta^*} > Kx_t/\gamma$, it is equivalent to $\beta_t^* = 0$, which together are (A.1) and (A.2). □

Given a contract $\mathcal{C}$, let $V^G[\mathcal{C}]$ be the principal's payoff under $\mathcal{C}$ in the presence of the

[18] If $T_F = \infty$, then $e^{-rt} \to 0$ as $t \to T_F$ while $U_t^{\beta^*}$ is bounded. If $T_F < \infty$, then $U_{T_F} = 0$.

loyal agent. Let $(\widetilde{U}[\mathcal{C}]_t)_{t\geq 0}$ be the continuation value process of the disloyal agent under the undermining policy $\beta = 1$. The following lemma will aid us in writing the principal's objective in terms of the disloyal agent's continuation value process, eliminating explicit dependence on the stakes path.

Lemma A.3. *Given any stringent contract $\mathcal{C} = (x, T_F)$,*

$$V^G[\mathcal{C}] = \frac{1}{K-1}\left(\widetilde{U}[\mathcal{C}]_0 + \int_0^{T_F} e^{-rt}\gamma\widetilde{U}[\mathcal{C}]_t\,dt\right).$$

Proof. Since the contract $\mathcal{C}$ is stringent, the loyal agent optimally refrains from undermining at all times. Therefore

$$V^G[\mathcal{C}] = \int_0^{T_F} e^{-rt}x_t\,dt.$$

The process x_t may be eliminated using the ODE

$$\frac{d\widetilde{U}[\mathcal{C}]_t}{dt} = -(K-1)x_t + (r+\gamma)\widetilde{U}[\mathcal{C}]_t$$

satisfied by the continuation utility process $\widetilde{U}[\mathcal{C}]$ for all $t < T_F$. The result is

$$V^G[\mathcal{C}] = \frac{1}{K-1}\int_0^{T_F} e^{-rt}\left((r+\gamma)\widetilde{U}[\mathcal{C}]_t - \frac{d}{dt}\widetilde{U}[\mathcal{C}]_t\right).$$

Using the identity $(r+\gamma)\widetilde{U}[\mathcal{C}]_t - \frac{d}{dt}\widetilde{U}[\mathcal{C}]_t = -e^{(r+\gamma)t}\frac{d}{dt}\left(e^{-(r+\gamma)t}\widetilde{U}[\mathcal{C}]_t\right)$ reduces this expression to

$$V^G[\mathcal{C}] = \frac{1}{K-1}\int_0^{T_F} e^{\gamma t}\frac{d}{dt}\left(e^{-(r+\gamma)t}\widetilde{U}[\mathcal{C}]_t\right)\,dt,$$

which may be integrated by parts to obtain

$$V^G[\mathcal{C}] = \frac{1}{K-1}\left(\widetilde{U}[\mathcal{C}]_0 - e^{-rT_F}\widetilde{U}[\mathcal{C}]_{T_F} + \int_0^{T_F} e^{rt}\gamma\widetilde{U}[\mathcal{C}]_t\,dt\right).$$

Using the terminal utility condition $\widetilde{U}[\mathcal{C}]_{T_F} = 0$ yields the desired expression. □

B Proofs of main results

In this appendix we provide all omitted proofs of results from the main text. We present proofs in order of logical development, beginning with the proofs of Lemmas 3 and 4, which logically precede all other results despite appearing after Theorem 1 in the main text.

Given a contract $\mathcal{C}$, let $V^\theta[\mathcal{C}]$ denote the principal's payoff under this contract in the presence of an agent of type θ, when the agent (if disloyal) follows an optimal undermining policy. Let $U^\theta[\mathcal{C}]$ denote the payoff to the agent of type θ from accepting the contract.

B.1 Proof of Lemma 3

If the menu $\mathcal{M} = (\mathcal{C}^G, \mathcal{C}^B)$ is obedient, then $U^B[\mathcal{C}^B] \geq U^B[\mathcal{C}^G]$. But then $V^B[\mathcal{C}^B] = -U^B[\mathcal{C}^B] \leq -U^B[\mathcal{C}^G] = V^B[\mathcal{C}^G]$, and so the principal would obtain a weakly higher payoff by offering the menu $\mathcal{M}' = (\mathcal{C}^G, \mathcal{C}^G)$.

B.2 Proof of Lemma 4

Fix a contract $\mathcal{C}$, and suppose $\mathcal{C}$ is not stringent. Construct a new contract $\mathcal{C}'$ which terminates the agent following any detected undermining. As this modification lowers the continuation utility achieved by the disloyal agent following detection at any history, it must be that the disloyal agent's ex ante payoff drops as well. So in the presence of the disloyal agent, the principal obtains a higher payoff under $\mathcal{C}'$ than under $\mathcal{C}$. Meanwhile, in the presence of the loyal agent, the principal obtains the same payoff under $\mathcal{C}'$ as under $\mathcal{C}$, since the loyal agent cannot undermine by assumption and is therefore unaffected by this modification.

B.3 Proof of Theorem 1

Define

$$\overline{\phi} \equiv \frac{(K-1)\gamma}{K(\gamma+r)}, \qquad \underline{\phi} \equiv \frac{(K-1)\gamma}{K(\gamma+r)}\left(\frac{Kq}{K-1}\right)^{1+\frac{Kr}{\gamma}}$$
$$t^* \equiv \frac{1}{\gamma}\log\left((K-1)\frac{1-q}{q}\right), \qquad \bar{t}_L \equiv \frac{K}{\gamma}\log\left(\frac{K-1}{Kq}\right),$$
$$\Delta \equiv \left(r+\frac{\gamma}{K}\right)^{-1}\log\left(\frac{\overline{\phi}}{\phi}\right), \qquad \underline{t}_M \equiv t^* - \frac{1}{\gamma}\log\left(Ke^{\frac{\gamma}{K}\Delta} - (K-1)\right), \bar{t}_M \equiv \underline{t}_M + \Delta.$$

We first prove that an optimal contract sets either $T_F^* = 0$ or $T_F^* = \infty$, and that when $T_F = \infty$ the unique optimal stakes path x^* is as follows:

Case 1: $q \geq \frac{K-1}{K}$. Then, $x_t^* = 1$ for all $t \geq 0$.

Case 2: $q < \frac{K-1}{K}$ and $\phi \geq \overline{\phi}$. Then,

$$x_t^* = \begin{cases} \phi & \text{if } t \in [0, t^*) \\ 1 & \text{if } t \geq t^*. \end{cases}$$

Case 3: $q < \frac{K-1}{K}$ and $\phi \in (\underline{\phi}, \overline{\phi})$. Then

$$x_t^* = \begin{cases} \phi & \text{if } t \in [0, \underline{t}_M) \\ \phi e^{(r+\gamma/K)(t-\underline{t}_M)} & \text{if } t \in [\underline{t}_M, \bar{t}_M) \\ 1 & \text{if } t \geq \bar{t}_M. \end{cases}$$

Case 4: $q < \frac{K-1}{K}$ and $\phi \leq \underline{\phi}$. Then

$$x_t^* = \begin{cases} \underline{\phi} e^{(r+\gamma/K)t} & \text{if } t \in [0, \bar{t}_L) \\ 1 & \text{if } t \geq \bar{t}_L. \end{cases}$$

After establishing this result, we conclude the proof by characterizing when $T_F = 0$ and when $T_F = \infty$.

Let $\overline{U} \equiv (K-1)/(r+\gamma)$, $\underline{U} \equiv \phi \overline{U} < \overline{U}$, and $\widehat{U} \equiv K\phi/\gamma$. For any $u \in [0, \overline{U}]$, let $(U^{**}(u)_t)_{t \geq 0}$ be the solution to the ODE

$$\dot{U}_t = \min \left\{ \left(r + \frac{\gamma}{K} \right) U_t, (r+\gamma) U_t - (K-1)\phi \right\}$$

with initial condition $h(0) = u$.[19] Let $\underline{T}^*(u) \equiv \inf\{t : U^{**}(u)_t \leq 0\}$, $\overline{T}^*(u) \equiv \inf\{t : U^{**}(u)_t \geq \overline{U}\}$, and $\widehat{T}^*(u) \equiv \inf\{t : U^{**}(u)_t \geq \widehat{U}\}$.

Note that the rhs of the ODE above equals the first argument of the minimum when $U_t \geq \widehat{U}$, and equals its second argument otherwise. Further, the second argument is positive if and only if $U_t \geq \underline{U}$, where $\underline{U} < \widehat{U}$. As a result, if $u < \underline{U}$, then $U^{**}(u)$ is strictly decreasing,

$$\dot{U}^{**}(u)_t = (r+\gamma) U^{**}(u)_t - (K-1)\phi$$

for all time, and $\underline{T}^*(u) < \infty$ while $\overline{T}^*(u) = \widehat{T}^*(u) = \infty$. Meanwhile if $u \geq \underline{U}$, then $U^{**}(u)$ is weakly increasing and $\underline{T}^*(u) = \infty$. The solution satisfies

$$\dot{U}^{**}(u)_t = \begin{cases} (r+\gamma) U^{**}(u)_t - (K-1)\phi, & t < \widehat{T}^*(u) \\ \left(r + \frac{\gamma}{K} \right) U^{**}(u)_t, & t \geq \widehat{T}^*(u). \end{cases}$$

Further, if $u > \underline{U}$ then $U^{**}(u)$ is strictly increasing and $\lim_{t \to \infty} U^{**}(u)_t = \infty$.

[19] The ODE has a globally unique solution given the fact that the right-hand side has global Lipschitz constant $r + \gamma$ (Teschl, 2012, Corollary 2.6).

Proposition B.1. *Fix a utility level $u \in [0, \overline{U}]$. Among all stringent contracts, there exists a unique*[20] *optimal contract $\mathcal{C}^*(u) = (x^*(u), T_F^*(u))$ delivering utility u to the disloyal agent:*

1. *If $u < \underline{U}$, then $T_F^*(u) = \underline{T}^*(u) < \infty$ and $x^*(u)_t = \phi$ for all $t \le T_F^*(u)$.*
2. *If $u \ge \underline{U}$ and $\widehat{U} < \overline{U}$, then $T_F^*(u) = \infty$ and*

$$x^*(u)_t = \begin{cases} \phi, & t < \widehat{T}^*(u), \\ \frac{\gamma}{K} U^{**}(u)_t, & \widehat{T}^*(u) \le t < \overline{T}^*(u) \\ 1, & t \ge \overline{T}^*(u) \end{cases}$$

3. *If $u \ge \underline{U}$ and $\widehat{U} \ge \overline{U}$, then $T_F^*(u) = \infty$ and*

$$x^*(u)_t = \begin{cases} \phi, & t < \overline{T}^*(u), \\ 1, & t \ge \overline{T}^*(u) \end{cases}$$

In all cases, $\mathcal{C}^*(u)$ is a loyalty test.

Proof. Given a fixed utility promise to the disloyal agent, the principal's value from the contract is determined entirely by her payoff under the loyal agent. Given a contract $\mathcal{C} = (x[\mathcal{C}], T_F[\mathcal{C}])$, let $U[\mathcal{C}]$ denote the disloyal agent's continuation utility path under an optimal undermining policy, and let $\widetilde{U}[\mathcal{C}]$ denote his continuation utility under the undermining policy $\beta = 1$. By Lemma A.3, the principal's payoff $V^G[\mathcal{C}]$ under $\mathcal{C}$ and the loyal agent is

$$V^G[\mathcal{C}] = \frac{1}{K-1}\left(\widetilde{U}[\mathcal{C}]_0 + \gamma \int_0^{T_F[\mathcal{C}]} e^{-rt} \gamma \widetilde{U}[\mathcal{C}]_t \, dt\right).$$

We derive the result by optimizing the alternative objective function

$$\overline{V}^G[\mathcal{C}] = \frac{1}{K-1}\left(U[\mathcal{C}]_0 + \gamma \int_0^{T_F[\mathcal{C}]} e^{-rt} \gamma U[\mathcal{C}]_t \, dt\right). \tag{B.1}$$

Note that $\overline{V}^G[\mathcal{C}] = V^G[\mathcal{C}]$ whenever $\mathcal{C}$ is a loyalty test. As a result, the contract optimizing $\overline{V}^G[\mathcal{C}]$ also optimizes $V^G[\mathcal{C}]$ provided that the optimal contract is a loyalty test.

[20] More precisely, it is unique when $\phi > 0$, or when $\phi = 0$ and $u \ge \underline{U}$. Otherwise, there exist a multiplicity of optimal termination times. This is because the agent's continuation utility under an optimal contract reaches zero in finite time, and once this occurs the principal can deliver this utility by setting stakes to zero and terminating at any future time. All such choices deliver the principal the same continuation value. By convention, we will assume that the principal terminates the agent as soon as possible in such continuations to eliminate any degeneracy.

To optimize the objective $\overline{V}^G$, we first derive two bounds on $dU[\mathcal{C}]_t/dt$ satisfied by any contract $\mathcal{C} = (x[\mathcal{C}], T_F[\mathcal{C}])$. By Lemma A.2, we know that for almost every $t \leq T_F[\mathcal{C}]$,

$$rU[\mathcal{C}]_t = \max_{\beta \in [0,1]} \left\{ (K\beta - 1)x[\mathcal{C}]_t - \gamma\beta U[\mathcal{C}]_t + \frac{d}{dt}U[\mathcal{C}]_t \right\}.$$

If $x[\mathcal{C}]_t \geq \gamma U[\mathcal{C}]_t/K$, then the rhs is maximized by $\beta = 1$ and

$$\frac{d}{dt}U[\mathcal{C}]_t = (r+\gamma)U[\mathcal{C}]_t - (K-1)x[\mathcal{C}]_t.$$

Then $x[\mathcal{C}]_t \geq \gamma U[\mathcal{C}]_t/K$ implies

$$\frac{d}{dt}U[\mathcal{C}]_t \leq \left(r + \frac{\gamma}{K}\right) U[\mathcal{C}]_t.$$

On the other hand, if $x[\mathcal{C}]_t < \gamma U[\mathcal{C}]_t/K$, then the rhs is maximized by $\beta = 0$ and

$$\frac{d}{dt}U[\mathcal{C}]_t = rU[\mathcal{C}]_t + x[\mathcal{C}]_t.$$

Then $x[\mathcal{C}] < \gamma U[\mathcal{C}]_t/K$ implies

$$\frac{d}{dt}U[\mathcal{C}]_t \leq \left(r + \frac{\gamma}{K}\right) U[\mathcal{C}]_t. \tag{B.2}$$

So this bound must hold at all times $t \leq T_F[\mathcal{C}]$. Meanwhile, no matter the optimal undermining policy, at each time

$$rU[\mathcal{C}]_t \geq (K-1)x[\mathcal{C}]_t - \gamma U[\mathcal{C}]_t + \frac{d}{dt}U[\mathcal{C}]_t \geq (K-1)\phi - \gamma U[\mathcal{C}]_t + \frac{d}{dt}U[\mathcal{C}]_t,$$

implying

$$\frac{d}{dt}U[\mathcal{C}]_t \leq (r+\gamma)U[\mathcal{C}]_t - (K-1)\phi. \tag{B.3}$$

Thus the continuation value process under any contract must satisfy

$$\frac{d}{dt}U[\mathcal{C}]_t \leq \min\left\{ \left(r + \frac{\gamma}{K}\right) U[\mathcal{C}]_t, (r+\gamma)U[\mathcal{C}]_t - (K-1)\phi \right\} \tag{B.4}$$

for almost every $t \leq T_F$.

In light of this bound, Lemma A.1 establishes that under any contract $\mathcal{C} = (x[\mathcal{C}], T_F[\mathcal{C}])$ delivering the disloyal agent utility u, $U^{**}(u)_t \geq U[\mathcal{C}]_t$ for all $t \leq T_F[\mathcal{C}]$. Since $\overline{U}$ represents an upper bound on the utility deliverable to the disloyal agent under any contract, we therefore have $U[\mathcal{C}]_t \leq \min\{U^{**}(u)_t, \overline{U}\}$ for all $t \leq T_F[\mathcal{C}]$. One implication of this result

is that $T_F[\mathcal{C}] \leq \underline{T}^*(u)$, since the agent must be terminated when $U[\mathcal{C}]$ reaches zero. Now, observe that $\overline{V}^G$ is pointwise increasing in the utility path, and is additionally increasing in the termination time. The results of the previous paragraph therefore imply that if some contract $\mathcal{C}$ satisfies $T_F[\mathcal{C}] = \underline{T}^*(u)$ and $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\}$ for all $t \leq \underline{T}^*(u)$, then it must maximize $\overline{V}^G$, and further must strictly dominate all contracts which do not satisfy these properties. We next show that $\mathcal{C}^*(u)$ uniquely fulfills these properties.

First consider the case $u < \underline{U}$. In this case, $U^{**}(u)_t < \underline{U} < \widehat{U}, \overline{U}$ for all times. The contract $\mathcal{C} = (x[\mathcal{C}]_t, \underline{T}^*(u))$ therefore satisfies $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\} = U^{**}(u)_t$ for all $t < \underline{T}^*(u)$ if and only if $dU[\mathcal{C}]_t/dt = (r+\gamma)U[\mathcal{C}]_t - (K-1)\phi$ for all such t. But the derivation of the bound (B.3) implies that the bound is saturated at a particular time t if and only if $x[\mathcal{C}]_t = \phi$ and additionally undermining at full intensity is optimal, i.e., $\phi \geq \gamma U[\mathcal{C}]_t/K$. Since $U[\mathcal{C}]_t \leq U^{**}(u)_t < \widehat{U}$, the latter inequality is automatically satisfied. Hence $U[\mathcal{C}]_t = U^{**}(u)_t$ for all $t < \underline{T}^*(u)$ if and only if $x[\mathcal{C}]_t = \phi = x^*(u)_t$ for all $t < \underline{T}^*(u)$. Thus $\mathcal{C}^*(u)$ is uniquely optimal.

Now consider the case $u \geq \underline{U}$, in which case $\underline{T}^*(u) = \infty$. The contract $\mathcal{C} = (x[\mathcal{C}]_t, \infty)$ satisfies $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\}$ for $t \leq \widehat{T}^*(u)$ if and only if

$$\frac{d}{dt}U[\mathcal{C}]_t = (r+\gamma)U[\mathcal{C}]_t - (K-1)\phi$$

for all such times. The reasoning of the previous paragraph implies that the first requirement is satisfied if and only if $x[\mathcal{C}]_t = \phi = x^*(u)_t$ for $t < \widehat{T}^*(u)$. Meanwhile, if $\widehat{U} < \overline{U}$, $\mathcal{C}$ satisfies $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\}$ for $t \in [\widehat{T}^*(u), \overline{T}^*(u))$ if and only if

$$\frac{d}{dt}U[\mathcal{C}]_t = \left(r + \frac{\gamma}{K}\right) U[\mathcal{C}]_t$$

for all such times. But the derivation of the bound (B.2) implies that it is saturated if and only if $x[\mathcal{C}]_t = \gamma U[\mathcal{C}]_t/K$. Thus this requirement is satisfied if and only if $x[\mathcal{C}]_t = x^*(u)_t$ for $t \in [\widehat{T}(u), T^*(u))$. Finally, $\mathcal{C}$ satisfies $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\}$ for $t \geq \overline{T}^*(u)$ if and only $U[\mathcal{C}]_t = \overline{U}$ for all such times. But $U[\mathcal{C}]_t = \overline{U}$ at some time t if and only if $x[\mathcal{C}]_t = 1$ for all future t. Hence this requirement is satisfied if and only if $x[\mathcal{C}]_t = 1 = x^*(u)_t$ for all $t \geq T^*(u)$. Putting these findings together yields the result that $U[\mathcal{C}]_t = \min\{U^{**}(u)_t, \overline{U}\}$ for all t if and only if $x[\mathcal{C}]_t = x^*(u)_t$ for all t. Thus $\mathcal{C}^*(u)$ is uniquely optimal.

It remains only to establish that $\mathcal{C}^*(u)$ is a loyalty test, which implies that it maximizes V^G as well as $\overline{V}^G$. Consider first $u < \underline{U}$. In that case $x[\mathcal{C}^*(u)]_t = \phi > \gamma U_t^{\mathcal{C}^*(u)}/K$ for all $t < \underline{T}^*(u)$, which by Lemma A.2 implies that $\beta = 1$ is optimal. Now consider $u \geq \underline{U}$. In that case $x[\mathcal{C}^*(u)]_t = \phi > \gamma U_t^{\mathcal{C}^*(u)}/K$ for all $t < \widehat{T}^*(u)$, which implies that $\beta_t = 1$ is optimal

for such times. Meanwhile, if $\widehat{U} < \overline{U}$ then $x[\mathcal{C}^*(u)]_t = \gamma U_t^{\mathcal{C}^*(u)}/K$ for all $t \in [\widehat{T}^*(u), \overline{T}^*(u))$, implying that $\beta_t = 1$ is optimal for all such times. Finally, for times $t \geq \overline{T}^*(u)$, $x[\mathcal{C}^*(u)]_t = 1$, in which case $\beta_t = 1$ is trivially optimal for all such times. So $\mathcal{C}^*(u)$ is a loyalty test, as claimed. □

In light of Proposition B.1, all that remains is to statically maximize the principal's expected payoff, taking into account uncertainty over the agent's type, with respect to the initial utility U_0 promised to the disloyal agent. Let $V^*(u)$ denote the principal's optimized expected payoff across all contracts delivering $U_0 = u$ to the disloyal agent. Let $u^* = \arg\max_{u \in [0, \overline{U}]} V^*(u)$.

We first show that either $u^* = 0$ or else $u^* \geq \underline{U}$. To see this, note that when $u \in (0, \underline{U})$, Proposition B.1 implies that

$$V^*(u) = q(1 - e^{-r\underline{T}^*(u)})\frac{\phi}{r} - (1-q)u.$$

The utility path $U^{**}(u)$ and termination time $\underline{T}^*(u)$ in this regime have the explicit solutions

$$U^{**}(u) = \underline{U} + (u - \underline{U})e^{(r+\gamma)t}, \qquad \underline{T}^*(u) = (r+\gamma)^{-1} \log \frac{\underline{U}}{\underline{U} - u}.$$

Hence

$$V^*(u) = q\left(1 - (1 - u/\underline{U})^{r/(r+\gamma)}\right)\frac{\phi}{r} - (1-q)u.$$

This function is strictly convex in u over $[0, \underline{U}]$, and so within this interval is uniquely maximized at one of its endpoints. Hence $u^* \notin (0, \underline{U})$, as claimed.

An immediate implication of this result is that either $T_F^*(u^*) = 0$ or $T_F^*(u^*) = \infty$, with the former corresponding to $u^* = 0$ and the latter corresponding to $u^* \geq \underline{U}$. We next compute $u^{**} = \arg\max_{u \in [\underline{U}, \overline{U}]} V^*(u)$, and in a final step compare $V^*(u^{**})$ to $V^*(0) = 0$ to obtain u^*.

Suppose $u \in [\underline{U}, \overline{U}]$. In light of Lemma A.3 and Proposition B.1, $V^*(u)$ may be written

$$V^*(u) \equiv -\left(1 - \frac{K}{K-1}q\right)u + \frac{q\gamma}{K-1}\int_0^\infty e^{-rt} \min\{U^{**}(u)_t, \overline{U}\}\, dt. \tag{B.5}$$

Case 1. If $q \geq (K-1)/K$, then this objective is strictly increasing in u for all u, meaning that $u^{**} = \overline{U}$. It follows from Proposition B.1 that $x_t^*(u^{**}) = 1$ for all $t \geq 0$. This solution corresponds to case 1 above.

Case 2. Assume going forward that $q < (K-1)/K$. First consider the case $\widehat{U} \geq \overline{U}$, which is equivalent to $\phi \geq \overline{\phi}$. In this case,

$$V^*(u) = -(1-q)u + q\left(\frac{\phi}{r} + e^{-r\overline{T}^*(u)}\frac{1-\phi}{r}\right).$$

Note that $\widehat{T}^*(u) \geq \overline{T}^*(u)$ for all $u \leq \overline{U}$ when $\widehat{U} \geq \overline{U}$, and so the utility path $U^{**}(u)$ has the same explicit solution as in the $u < \underline{U}$ case for times $t \leq \overline{T}^*(u)$, with

$$\overline{T}^*(u) = (r+\gamma)^{-1}\log\frac{\overline{U}-\underline{U}}{u-\underline{U}}.$$

Therefore

$$V^*(u) = -(1-q)u + q\left(\frac{\phi}{r} + \left(\frac{u-\underline{U}}{\overline{U}-\underline{U}}\right)^{r/(r+\gamma)}\frac{1-\phi}{r}\right).$$

This value function is strictly concave in u, with derivative

$$\frac{dV^*}{du} = -(1-q) + \frac{q}{K-1}\left(\frac{u-\underline{U}}{\overline{U}-\underline{U}}\right)^{-\gamma/(r+\gamma)}.$$

This expression vanishes at

$$u_H \equiv \underline{U} + (\overline{U}-\underline{U})\left(\frac{q}{1-q}\frac{1}{K-1}\right)^{1+r/\gamma},$$

which lies in $(\underline{U}, \overline{U})$ given that $0 < q < (K-1)/K$ and therefore $0 < q/(1-q) < K-1$. Hence $u^{**} = u_H$. Note that $\overline{T}^*(u^{**}) = t^*$, corresponding to case 2 above.

Cases 3 and 4. Now consider the case $\widehat{U} < \overline{U}$, i.e., $\phi < \overline{\phi}$. For $u \geq \widehat{U}$, the utility path $U^{**}(u)$ has the explicit solution

$$U^{**}(u) = ue^{(r+\gamma/K)t},$$

and $\widehat{T}^*(u) = 0$ while

$$\overline{T}^*(u) = (r+\gamma/K)^{-1}\log\frac{\overline{U}}{u}.$$

Then $V^*(u)$ may be explicitly computed to be

$$V^*(u) = -u + \frac{qK}{r+\gamma}\frac{r+\gamma/K}{r}\left(\frac{u}{\overline{U}}\right)^{r/(r+\gamma/K)}.$$

This is a strictly concave function of u, with derivative

$$\frac{dV^*}{du} = -1 + \frac{qK}{K-1}\left(\frac{u}{\overline{U}}\right)^{-\frac{\gamma/K}{r+\gamma/K}}.$$

When $q < (K-1)/K$, this expression is negative for u sufficiently close to $\overline{U}$, and it vanishes at

$$u_L \equiv \overline{U}\left(\frac{qK}{K-1}\right)^{1+rK/\gamma}.$$

Note that $u_L \geq \widehat{U}$ if and only if $\phi \leq \underline{\phi}$. If this inequality is violated, then V^* is strictly decreasing for all $u \geq \widehat{U}$.

Next, we turn to $u < \widehat{U}$. (This case is relevant only if $\phi = 0$, in which case all expressions that follow are well-defined.)

$$V^*(u) = -(1-q)u + q\left(\int_0^{\widehat{T}^*(u)} x^*(u)_t\, dt + e^{-r\widehat{T}^*(u)}\int_0^{\infty} e^{-rt} x^*(u)_{t+\widehat{T}^*(u)}\, dt\right).$$

Now, observe that $x^*(u)_t = \phi$ for $t < \widehat{T}^*(u)$, while $x^*(u)_t = x^*(\widehat{U})_{t-\widehat{T}^*(u)}$ for $t \geq \widehat{T}^*(u)$. Hence this expression is equivalently

$$V^*(u) = -(1-q)u + q\left(\frac{\phi}{r}\left(1 - e^{-r\widehat{T}^*(u)}\right) + e^{-r\widehat{T}^*(u)} q^{-1}(V^*(\widehat{U}) + (1-q)\widehat{U})\right).$$

For $t < \widehat{T}^*(u)$ the utility path $U^{**}(u)$ has the explicit solution

$$U^{**}(u) = \underline{U} + (u - \underline{U})e^{(r+\gamma)t},$$

and so

$$\widehat{T}^*(u) = (r+\gamma)^{-1}\log\frac{\widehat{U} - \underline{U}}{u - \underline{U}}.$$

So $V^*(u)$ may be written

$$V^*(u) = q\frac{\phi}{r} - (1-q)u + \left(\frac{u-\underline{U}}{\widehat{U}-\underline{U}}\right)^{r/(r+\gamma)}\left(V^*(\widehat{U}) + (1-q)\widehat{U} - q\frac{\phi}{r}\right).$$

Recalling that $V^*(\underline{U}) = q\phi/r - (1-q)\widehat{U}$, this may therefore be written

$$V^*(u) = q\frac{\phi}{r} - (1-q)u + \left(\frac{u-\underline{U}}{\widehat{U}-\underline{U}}\right)^{r/(r+\gamma)}\left(V^*(\widehat{U}) - V^*(\underline{U})\right).$$

The derivative of this expression is

$$\frac{dV^*}{du} = -(1-q) + \frac{r}{r+\gamma}\frac{V^*(\widehat{U}) - V^*(\underline{U})}{\widehat{U}-\underline{U}}\left(\frac{u-\underline{U}}{\widehat{U}-\underline{U}}\right)^{-\gamma/(r+\gamma)},$$

or after simplification

$$\frac{dV^*}{du} = -(1-q) + q\left(\frac{K}{K-1}\left(\frac{\overline{\phi}}{\phi}\right)^{\frac{\gamma/K}{r+\gamma/K}} - 1\right)\left(\frac{u-\underline{U}}{\widehat{U}-\underline{U}}\right)^{-\gamma/(r+\gamma)}.$$

Since $\phi < \overline{\phi}$, this derivative is strictly decreasing in u, i.e., V^* is a strictly concave function on $[\underline{U}, \widehat{U}]$. Further, $dV^*/du > 0$ for u sufficiently close to $\underline{U}$. The previous expression for the derivative vanishes at

$$u_M \equiv \underline{U} + (\widehat{U}-\underline{U})\left(\frac{q}{1-q}\left(\frac{K}{K-1}\left(\frac{\overline{\phi}}{\phi}\right)^{\frac{\gamma/K}{r+\gamma/K}} - 1\right)\right)^{1+r/\gamma}.$$

The bound $u_M < \widehat{U}$ holds if and only if $\phi > \underline{\phi}$. Otherwise, V^* is strictly increasing on $[\underline{U}, \widehat{U}]$.

Note that regardless of the comparison between ϕ and $\underline{\phi}$, V^* is single-peaked on $[\underline{U}, \overline{U}]$. If $\phi \leq \underline{\phi}$, then $u^{**} = u_L$, while if $\phi > \underline{\phi}$ then $u^{**} = u_M$. In the first case, $\overline{T}^*(u^{**}) = \bar{t}_L$, corresponding to case 3 above. In the second case, $\widehat{T}^*(u^{**}) = \underline{t}_M$ and $\overline{T}^*(u^{**}) = \underline{t}_M + \overline{T}^*(\widehat{U}) = \underline{t}_M + \Delta$, corresponding to case 4 above.

To complete the proof, we need only compare $V^*(u^{**})$ to $V^*(0) = 0$ to determine the optimal contract. If $V^*(u^{**}) \geq 0$, then the contract outlined at the beginning of the proof is optimal, while otherwise the degenerate contract which terminates the agent immediately is optimal. The following lemma performs the comparison.

For $x \in \{H, M, L\}$, let $\mathcal{S}^x$ denote the set of (q, ϕ) pairs in the high-, moderate-, and low-stakes regimes, respectively, and let $\mathcal{S}^+ = \{(q,\phi) : q \geq \frac{K-1}{K}\}$.

Lemma B.1. *There exists a nonempty set $\mathcal{S}^\emptyset$ of (q,ϕ) pairs for which $0 > V^*(u^{**})$, and this set satisfies $\mathcal{S}^\emptyset = \{(q,\phi) : q < \underline{q}(\phi)\}$ for an increasing, continuous function $\underline{q} : [0,1] \to [0, q^*)$ with $\underline{q}(0) = 0$. Whenever $q > \underline{q}(\phi)$, $V^*(u^{**}) > 0$. If $q = \underline{q}(\phi)$, then $V^*(u^{**}) = 0$.*

The set $\mathcal{S}^\emptyset$ intersects $\mathcal{S}^H$ and $\mathcal{S}^M$ but it does not intersect $\mathcal{S}^L$ or $\mathcal{S}^+$, and neither $\mathcal{S}^H$ nor $\mathcal{S}^M$ is a subset of $\mathcal{S}^\emptyset$.

Proof. We construct the function $\underline{q}$ piecewise over $[0, \overline{\phi})$ and $[\overline{\phi}, 1]$. Inspecting the value function (B.5) reveals that the principal obtains a strictly positive payoff from setting $T_F = \infty$ when $q \geq q^*$, i.e., when $(q,\phi) \in \mathcal{S}^+$. It therefore suffices to consider the high-, moderate-,

and low-stakes cases, with $q < q^*$. We will write $V^*(u; \phi, q)$ to make explicit the dependence on (ϕ, q) in (B.5).

First consider the high-stakes region $\phi \geq \overline{\phi}$. Since $\phi \geq \overline{\phi}$, we have $\underline{U} > 0$, so $V^*(u; \phi, 0) = -u \leq -\underline{U} < 0$ for all $u \in [\underline{U}, \overline{U}]$, and thus $V^*(u^{**}; \phi, 0) < 0$. On the other hand, we just saw that the principal can attain a strictly positive payoff whenever $q \geq q^*$. Hence, the equation $V^*(u^{**}; \phi, q) = 0$ implicitly defines a continuous function $\underline{q}^H : [\overline{\phi}, 1] \to (0, q^*)$ such that for $\phi \in [\overline{\phi}, 1]$, $V^*(u^{**}; \phi, q) < 0$ (and the unique optimal contract is degenerate) iff $q < \underline{q}^H(\phi)$. Since $V^*(u^{**}; \phi, q)$ is strictly decreasing in ϕ and strictly increasing in q in the high-stakes region, $\underline{q}^H$ is strictly increasing.

Now consider $\phi < \overline{\phi}$. Define

$$\Phi(p) \equiv \frac{(K-1)\gamma}{K(\gamma + r)} \left(\frac{Kp}{K-1} \right)^{1 + \frac{Kr}{\gamma}}$$

to be the optimal initial stakes when initial beliefs are p and the lower bound constraint does not bind. (Note that $\Phi(q) = \underline{\phi}$.) For each $\phi \in (0, \overline{\phi})$, we are in the moderate-stakes region whenever $q \in [0, (\Phi)^{-1}(\phi))$. For such q, $V^*(u^{**}; \phi, q)$ is strictly decreasing in ϕ as the lower bound constraint binds and is strictly increasing in q. Moreover, as in the high-stakes region, $V^*(u; \phi, 0) = -u < 0$ for $\phi > 0$ and $u \in [\underline{U}, \overline{U}]$, so $V^*(u^{**}; \phi, 0) < 0$. Thus, the equation $V^*(u^{**}; \phi, q) = 0$ implicitly defines a continuous, increasing function $\underline{q}^M : [0, \overline{\phi}) \to (0, q^*)$ such that for $(q, \phi) \in \mathcal{S}^M$, $V^*(u^{**}; \phi, q) < 0$ if and only if $q < \underline{q}^M(\phi)$, where we specify $\underline{q}^M(0) = 0$.

Meanwhile, we are in the low-stakes region whenever $q \in [(\Phi)^{-1}(\phi), q^*)$, and the explicit forms for V^* and u^{**} can be used to calculate that $V^*(u^{**}; \phi, q) = q(Kq/(K-1))^{Kr/\gamma}\gamma/(r(r+\gamma)) \geq 0$, with equality for $(q, \phi) \in \mathcal{S}^L$ iff $(q, \phi) = (0, 0)$. Hence, $\underline{q}^M(\phi) \leq (\Phi)^{-1}(\phi) < q^*$, with the first inequality strict except when $\phi = 0$.

Now define the function $\underline{q} : [0, 1] \to [0, q^*)$ by

$$\underline{q}(\phi) = \begin{cases} \underline{q}^M(\phi), & \phi \in [0, \overline{\phi}) \\ \underline{q}^H(\phi), & \phi \in [\overline{\phi}, 1], \end{cases}$$

which is continuous since $V^*(u^{**}; \phi, q)$ is continuous, in particular at the boundary between $\mathcal{S}^H$ and $\mathcal{S}^M$. Define $\mathcal{S}^\emptyset$ as in the lemma statement. By the construction above, the principal strictly prefers not to hire the agent if and only if $(q, \phi) \in \mathcal{S}^\emptyset$, and $\mathcal{S}^\emptyset$ intersects $\mathcal{S}^H$ and $\mathcal{S}^M$ but not $\mathcal{S}^L$ or $\mathcal{S}^+$. Moreover, since $\underline{q} < q^*$, $\mathcal{S}^H$ is not a subset of $\mathcal{S}^\emptyset$; and since and $\underline{q}^M(\phi) < (\Phi)^{-1}(\phi)$ for $\phi \in (0, \overline{\phi})$, $\mathcal{S}^M$ is not a subset of $\mathcal{S}^\emptyset$. □

B.4 Proof of Lemma 1

We make use of various closed-form expressions derived in the proof of Theorem 1. The size of the stakes jump at graduation is $1 - \max\{\phi, \overline{\phi}\}$, which is weakly decreasing in γ since $\overline{\phi} \equiv \frac{K-1}{K}\frac{\gamma}{\gamma+r}$ is strictly increasing in γ.

Next, consider the graduation time. By continuity of the optimal stakes curve wrt the input parameters, it suffices to show that this time is decreasing in γ within each stakes environment. For high stakes, the graduation time is t^*, and for low stakes, it is $\bar{t}_L$; both of these are decreasing in γ by inspection. For moderate stakes, the graduation time is $\bar{t}_M = \underline{t}_M + \Delta$. As Δ is independent of q, we have $d\bar{t}_M/dq = d\underline{t}_M/dq = -1/(q(1-q)\gamma)$, and thus $d^2\bar{t}_M/(dq\,d\gamma) = [\gamma^2 q(1-q)]^{-1} > 0$. Hence, to obtain an upper bound on $d\bar{t}_M/d\gamma$, we can increase q until $\underline{\phi} = \phi$, at the boundary of the low-stakes environment:

$$\begin{aligned} d\bar{t}_M/d\gamma|_{\phi=\underline{\phi}} &= \frac{(K-1)\left[-\left(\frac{Kq}{K-1}-1\right)r\gamma + \frac{K}{K-1}(r+\gamma)[r+\gamma(1-q)]\log\left(\frac{Kq}{K-1}\right)\right]}{(1-q)\gamma^2(r+\gamma)(Kr+\gamma)} \\ &< \frac{(K-1)\left(\frac{Kq}{K-1}-1\right)}{(1-q)\gamma^2(r+\gamma)(Kr+\gamma)}\left[-r\gamma + \frac{K}{K-1}(r+\gamma)[r+\gamma(1-q)]\right] < 0, \end{aligned}$$

where we have used $q < (K-1)/K$ and the inequality $\log x < x - 1$ for $x > 0$. Hence, $d\bar{t}_M/d\gamma < 0$ in the moderate-stakes environment.

Toward proving the remaining result in the lemma, define $\underline{\gamma} > 0$ as the unique value of γ solving the equation $\overline{\phi} = \phi$ and likewise define $\overline{\gamma} > \underline{\gamma}$ as the unique solution to $\underline{\phi} = \phi$. Hence, the environment is high-stakes for $\gamma \leq \underline{\gamma}$, moderate-stakes for $\gamma \in (\underline{\gamma}, \overline{\gamma})$, and low-stakes for $\gamma \geq \overline{\gamma}$.

For high stakes ($\gamma \leq \underline{\gamma}$), the escalation phase is degenerate. For low stakes ($\gamma \geq \overline{\gamma}$), the escalation phase length is $\bar{t}_L$, which is decreasing in γ as noted above. For moderate stakes ($\gamma \in (\underline{\gamma}, \overline{\gamma})$), the length of the escalation phase is Δ. By continuity of the optimal stakes curve in the model parameters, $\Delta \to 0$ as $\gamma \downarrow \underline{\gamma}$ and $\Delta \to \bar{t}_L|_{\gamma=\overline{\gamma}} > 0$ as $\gamma \uparrow \overline{\gamma}$, and thus Δ must be increasing in γ at some point in $(\underline{\gamma}, \overline{\gamma})$. Hence, to establish the result, it suffices to establish that Δ is quasiconcave in γ over this interval. Differentiation wrt γ yields

$$d\Delta/d\gamma = \frac{K}{(Kr+\gamma)^2}\left[\frac{r(Kr+\gamma)}{\gamma(r+\gamma)} - \log\left(\frac{\overline{\phi}}{\phi}\right)\right].$$

The term in brackets has derivative $-r(Kr+\gamma)(r+2\gamma)/(\gamma^2(r+\gamma)^2) < 0$ wrt γ. Given that Δ is increasing in γ at some point in $(\underline{\gamma}, \overline{\gamma})$, either $d\Delta/d\gamma$ is positive on this interval, or it changes sign from positive to negative exactly once. In either case, Δ is quasiconcave in γ, as desired.

B.5 Proof of Lemma 2

Let U^* be the disloyal agent's value function given the stakes curve x^*. By Lemma A.2, an undermining policy β is optimal if and only if for almost all t, $U_t^* < Kx_t^*/\gamma$ implies $\beta_t = 1$ and $U_t^* > Kx_t^*/\gamma$ implies $\beta_t = 0$. The construction of an optimal contract in Theorem 1 demonstrates that for almost all $t \in [0, \infty)$, U^* satisfies the identity $x_t^* = \left((r+\gamma)U_t^* - \dot{U}_t^*\right)/(K-1)$ and the constraint $\dot{U}_t^* \leq (r + \gamma/K)U_t^*$, with equality in the latter for $t \in [\underline{t}, \overline{t})$, where $\underline{t} = \widehat{T}^*(u^{**})$ and $\overline{t} = \overline{T}^*(u^{**})$. Combining these yields $U_t^* \leq \frac{K}{\gamma}x_t^*$, with equality for $t \in [\underline{t}, \overline{t})$. It follows from (A.1) and (A.2) that β is optimal if and only if for almost every $t \in [0, \underline{t}) \cup [\overline{t}, \infty)$, $\beta_t = 1$. In other words, the agent strictly prefers to undermine during the probationary and trusted phases and is indifferent during the escalation phase. By the zero-sum property, the principal is indifferent over all undermining policies which are optimal for the agent.

C No-commitment results

C.1 Formal definitions

Preliminary to proving the results appearing in Section 6, we develop a notion of strategies in the no-commitment game between the principal and disloyal agent tailored to our setting, and use them to define Bayes Nash equilbrium and perfect Bayesian equilibrium.[21] We leverage the fact that only one player (the principal) acts observably to sidestep technical issues arising in more general settings.

The game takes place on a state space $\Omega = \mathbb{X} \times \mathbb{B} \times \mathbb{T}$, where:

- $\mathbb{X}$ is the set of càdlàg, $[\phi, 1]$-valued, increasing functions,
- $\mathbb{B}$ is the set of càdlàg, $[0, 1]$-valued functions,
- $\mathbb{T}$ is set of elements $\tau \in (\mathbb{R}_+ \cup \{\infty\})^{\mathbb{N}}$ such that:
  1. $\tau_k \leq \tau_{k+1}$ for all k, with the inequality strict whenever $\tau_k < \infty$,
  2. $\lim_{k\to\infty} \tau_k = \infty$.

Each element $\omega \in \Omega$ is a vector $\omega = (x, \beta, \tau)$, where x is the path of stakes; β is the path of undermining; and τ is the vector of times at which undermining is observed. Note that τ may have entries of ∞, which correspond to histories with only a finite number of observations of

[21] For simplicity, we do not explicitly model the type of the agent, and simply define payoffs appropriately to reflect the probabilistic presence of the disloyal agent.

undermining. Also, at most a finite number of observations of undermining are allowed by any finite time. $\mathbb{X}$, $\mathbb{B}$, and $\mathbb{T}$ are endowed with the Borel Σ-algebras generated by sup norm.

Let X and B be the coordinate processes $X(\omega) = \omega(x)$ and $B(\omega) = \omega(\beta)$, and define the counting process N by

$$N_t(\omega) = \sum_{k=1}^{\infty} \mathbf{1}\{t \geq \omega(\tau)_k\}.$$

Define $\mathbb{F}^P$ to be the filtration induced by N^- and $\mathbb{F}^A$ to be the filtration induced by (X^-, N^-), where $N^- \equiv N_{t-}$ and $X_t^- \equiv X_{t-}$.

Definition C.1. *A strategy profile is a triple (χ, Λ, ζ), where:*

- *χ is an $\mathbb{F}^P$-adapted, càdlàg, $[\phi, 1]$-valued stochastic process,*
- *Λ is an $\mathbb{F}^P$-stopping time,*
- *ζ is an $\mathbb{F}^A$-adapted, càdlàg, $[0, 1]$-valued stochastic process.*

(χ, Λ) are chosen by the principal, and represent her control of stakes and termination, respectively. ζ is chosen by the agent and represents his undermining policy, depending on the history of stakes and observed undermining.[22] Intuitively, (χ, ζ) jointly determine a probability measure over stakes paths and undermining. However, as both condition on the path of observed undermining, whose distribution in turn depends on ζ, the mapping must be carefully defined and shown to be coherent. We next formally a mapping from (χ, ζ) to probability measures on Ω.

Fix a stakes-undermining pair (χ, ζ). $\mathbb{F}^P$-adaptedness of χ implies existence of a unique measurable function $\widetilde{X}^\chi : \mathbb{T} \to \mathbb{X}$ such that $\widetilde{X}^\chi(\omega(\tau)) = \chi(\omega)$ for every ω. Similarly, $\mathbb{F}^A$-adaptedness of ζ implies existence of a unique measurable function $\widetilde{B}^\zeta : \mathbb{X} \times \mathbb{T} \to \mathbb{B}$ such that $\widetilde{B}^\zeta(\omega(x), \omega(\tau)) = \zeta(\omega)$ for every ω. Also, for each $\tau \in \mathbb{T}$ and $k = 0, 1, ...$, define τ^k to be the vector such that $\tau_j^k = \tau_j$ for $j \leq k$, and $\tau_j^k = \infty$ for $j > k$.

Then given a pair (χ, ζ), a probability distribution $\mu^\tau(\chi, \zeta)$ over $\mathbb{T}$ can be defined via the conditional distributions

$$\begin{aligned} &\Pr(\tau_k \leq \Delta t + \tau_{k-1} \mid \tau_1, ..., \tau_{k-1}) \\ &= \begin{cases} 0, & \Delta t < 0, \\ 1 - \exp\left(-\gamma \int_0^{\Delta t} \widetilde{B}^\zeta_{s+\tau_{k-1}}(\widetilde{X}^\chi(\tau^{k-1}), \tau^{k-1})\, ds\right), & \Delta t \geq 0. \end{cases} \end{aligned}$$

[22] To avoid technical issues involving non-right-continuous stakes paths following a deviation by the principal, we assume that the agent can condition only on stakes prior to time t, but that time-t stakes are observed only after the agent makes his time-t undermining decision.

whenever $\tau_{k-1} < \infty$, and $\Pr(\tau_k = \infty) = 1$ whenever $\tau_{k-1} = \infty$. These conditional distributions define the probability of the kth jump using the expected cumulative hazard rate under ζ, assuming that no further jumps arrive after time τ_{k-1} and stakes evolve according to χ.

For this construction to induce a well-defined probability distribution $\mu^\tau(\chi, \zeta)$ over $\mathbb{T}$, its support must be a subset of $\mathbb{T}$. Clearly $\mu^\tau(\chi, \zeta)$ puts probability 1 on vectors τ which are strictly increasing so long as elements are finite. Further, for each $t < \infty$,

$$\Pr(\tau_k \leq t \mid \tau_1 \leq t, ..., \tau_{k-1} \leq t) \leq 1 - \exp(-\gamma t),$$

and so

$$\mu^\tau(\chi, \zeta)(\{\lim_{k\to\infty} \tau_k \leq t\}) \leq \prod_{k=1}^{\infty} (1 - \exp(-\gamma t)) = 0.$$

In other words, $\mu^\tau(\chi, \zeta)\left(\{\lim_{k\to\infty} \tau_k = \infty\}\right) = 1$, so $\mu^\tau(\chi, \zeta)$ places probability 1 on $\mathbb{T}$, as required.

The measure $\mu^\tau(\chi, \zeta)$ induces a natural measure $\mu(\chi, \zeta)$ over Ω as follows. For every measurable subset $E \subset \Omega$, let

$$\overline{E}^{(\chi,\zeta)} = \{\tau \in \mathbb{T} \ : \ (\widetilde{X}^\chi(\tau), \widetilde{B}^\zeta(\widetilde{X}^\chi(\tau,), \tau), \tau) \in E\}.$$

(Note that $\overline{E}^{(\chi,\zeta)}$ is a measurable subset of $\mathbb{T}$ whenever E is a measurable subset of Ω, given that compositions and vectors of measurable functions are measurable.) Then define $\mu(\chi, \zeta)(E) = \mu^\tau(\chi, \zeta)(\overline{E}^{(\chi,\zeta)})$. Note in particular that under $\mu(\chi, \zeta)$, $X(\omega) = \chi(\omega)$ and $B(\omega) = \zeta(\omega)$ a.s.

Using the measure $\mu(\chi, \zeta)$, payoffs to the principal and disloyal agent from a given strategy profile may be defined as

$$\Pi(\chi, \Lambda, \zeta) = q\, \mathbb{E}^{(\chi,0)} \left[\int_0^\Lambda e^{-rt} X_t (1 - K B_t)\, dt\right] + (1 - q)\, \mathbb{E}^{(\chi,\zeta)} \left[\int_0^\Lambda e^{-rt} X_t (1 - K B_t)\, dt\right]$$

and

$$U^B(\chi, \Lambda, \zeta) = \mathbb{E}^{(\chi,\zeta)} \left[\int_0^\Lambda e^{-rt} X_t (K B_t - 1)\, dt\right]$$

and where $\mathbb{E}^{(\chi,\zeta)}$ represents expectations under the measure $\mu(\chi, \zeta)$. Note that the principal's payoff is a weighted average of two expected payoffs, one assuming the agent plays the strategy $\chi' = 0$, in which case $B = 0$ with probability 1, and one assuming the agent plays the strategy $\chi' = \chi$. These terms capture the payoff contributions from the loyal and disloyal agent, respectively.

Definition C.2. *A* Bayes Nash equilibrium *is a strategy profile* (χ, Λ, ζ) *such that* $\Pi(\chi, \Lambda, \zeta) \geq \Pi(\chi', \Lambda', \zeta)$ *for all* (χ', Λ') *and* $U^B(\chi, \Lambda, \zeta) \geq U^B(\chi, \Lambda, \zeta')$ *for all* ζ'.

We now turn to the definition of a perfect Bayesian equilibrium. Given any $\mathbb{F}^A$-adapted agent strategy ζ, define a belief process π^ζ by

$$\pi_t^\zeta = \begin{cases} \left(1 + \frac{1-q}{q} \exp\left(-\gamma \int_0^t \zeta_s \, ds\right)\right)^{-1}, & N_t = 0, \\ 0, & N_t > 0. \end{cases}$$

Note that π^ζ is also $\mathbb{F}^A$-adapted, and accords with Bayes' rule whenever it applies. (Bayes' rule does not apply if N increments at a time when $\zeta_t = 0$. We resolve the resulting belief indeterminacy after such histories in favor of sure beliefs that the agent is disloyal.) The strategy ζ induces a natural family of continuation strategies at each time t for the continuation games beginning at that time. This family of strategies is indexed by the history $h_t = (x_s, n_s)_{s<t}$, where

$$n_t = \sum_{k=1}^{\infty} \mathbf{1}\{t \geq \tau_k\},$$

and the continuation game at time t with history h_t is isomorphic to the original game with initial beliefs $\Pr(\theta = G) = \pi_t$. We will denote the induced strategy in each continuation game by $\zeta^{\geq t}(h_t)$.

Definition C.3. *A* perfect Bayesian equilibrium *is a strategy profile* (χ, Λ, ζ) *such that:*

- χ *is* $\mathbb{F}^P$*-adapted,* Λ *is an* $\mathbb{F}^P$*-stopping time, and* ζ *is* $\mathbb{F}^A$*-adapted,*
- (χ, Λ, ζ) *is a Bayes Nash equilibrium,*
- *For every time* $t \geq 0$ *and history* $h_t = (x_s, n_s)_{s<t}$*, there exists a principal strategy* (χ', Λ') *such that* $(\chi', \Lambda', \zeta^{\geq t}(h_t))$ *is a pure-strategy Bayes Nash equilibrium of the game with initial stakes* X_t *and initial beliefs* π_t^ζ.

Our notion of perfect Bayesian equilibrium economizes on notation by not explicitly recording the principal's off-path continuation play as part of the equilibrium strategy profile. Nevertheless, the equilibrium definition requires that, both on and off path, the agent's continuation play be part of a Nash equilibrium of the continuation game. Note that the requirement of Bayes Nash equilibrium already implies this condition for all on-path histories, and thus the on-path continuation play constitutes an equilibrium of the continuation game. The extra requirement is that some other equilibrium strategy profile justify the agent's play following any off-path history.[23]

[23]This definition also does not formally require that the *same* continuation equilibrium is played at each

C.2 Proof of Proposition 2

We first establish that any Bayes Nash equilibrium must immediately terminate the disloyal agent following undermining on the equilibrium path. (We defer a proof of this result to Section C.2.1.)

Lemma C.1. *Suppose $\phi > 0$. Fix a strategy profile σ. Let τ_D be the first time the agent is detected undermining, and τ_F be the time at which the agent is terminated. If $\Pr_\sigma(\tau_F > \tau_D) > 0$, then σ is not a Bayes Nash equilibrium.*

In light of this lemma, we will restrict attention to strategy profiles in which the principal immediately terminates the agent following detected undermining. We will call such strategy profiles *stringent*. In principle this restriction could rule out equilibria in non-stringent strategies under which the principal does not immediately terminate the agent following detection, but only when the agent does not actually undermine. However, we will see that the unique stringent equilibrium features a binding IC constraint whenever it is optimal for the agent not to undermine. As a result, no non-stringent equilibria exist.[24]

A stringent strategy profile can be summarized as a triple $\sigma = (x, T_F, \beta)$, where x is the stakes curve prior to detected undermining; T_F is the time at which the principal terminates the agent supposing no undermining is detected; and β is the disloyal agent's undermining strategy prior to being detected. We will call a stringent strategy profile *incentive-compatible* if β maximizes the disloyal agent's payoff under the stringent contract (x, T_F).

Define an indexed set of optimal stakes curves $x^{**}(p)$ for $p \in [0, 1]$, where $x^{**}(p)$ is the (unique) optimal commitment stakes curve when the agent is loyal with probability p. This family of curves is characterized by Theorem 1 as q is allowed to vary. In particular, $x^{**}(q) = x^*$.

Definition C.4. *A stringent strategy profile (x, T_F, β) is* time-consistent *if $T_F = \infty$ and at each time t, $(x_s)_{s \geq t} = x^{**}(\pi_t)$, where π_t are the principal's time-t posterior beliefs that $\theta = G$ conditional on no detected undermining and undermining policy β.*

Note in particular that any time-consistent strategy must implement the optimal stakes curve x^*.

history along the revised equilibrium path following a deviation. However, it is easy to see that along the revised equilibrium path, continuation play of the revised equilibrium strategy continues to satisfy the equilibrium requirement for successive subgames.

[24] If $\phi = 0$, then additional equilibria exist in which the principal instead sets stakes to zero rather than terminating immediately. The selected continuation in this case makes no difference to pre-detection equilibrium outcomes. We resolve this indeterminacy by selecting the immediate-termination equilibrium.

We now establish that there exists a unique β^* under which the stringent strategy profile (x^*, ∞, β^*) is incentive-compatible and time-consistent. (We defer a proof of this result to Section C.2.2.) Let $\underline{t}^* \equiv \inf\{t \ : \ x_t^* > \phi\}$ and $\overline{t}^* \equiv \inf\{t \ : \ x_t^* = 1\}$. Clearly $0 \leq \underline{t}^* \leq \overline{t}^* < \infty$.

Lemma C.2. *There exists a unique undermining policy β^* such that the stringent strategy profile (x^*, ∞, β^*) is incentive-compatible and time-consistent. Letting π^* be the sequence of posterior beliefs induced by β^* conditional on no detected undermining, β^* and π^* satisfy:*

- $\beta_t^* = 1$ *for* $t \in [0, \underline{t}^*) \cup [\overline{t}^*, \infty)$,
- *If* $\underline{t}^* < \overline{t}^*$, *then* β^* *is a strictly increasing, continuous function on* $[\underline{t}^*, \overline{t}^*)$ *with* $\beta_t^* = 1/(K(1-\pi_t^*))$ *for all* $t \in [\underline{t}^*, \overline{t}^*)$ *and* $\beta^*_{\overline{t}^*\_} = 1$.

By varying q in the proof of Lemma C.2, a direct corollary is that for each $p \in [0,1]$, there exists a unique time-consistent, incentive-compatible undermining policy $\beta^{**}(p)$ corresponding to the optimal stakes curve $x^{**}(p)$ when beliefs are p. Another immediate consequence is that for all t, $(\beta^*)_{s \geq t} = \beta^{**}(\pi_t^*)$. For otherwise, the undermining policy $\widetilde{\beta} = ((\beta^*)_{s \in [0,t)}, \beta^{**}(\pi_t^*)_{s \geq t})$ would constitute another incentive-compatible, time-consistent undermining policy under x^* and initial belief q, contradicting the uniqueness guaranteed by Lemma C.2.

Our next result shows that time-consistency and incentive-compatibility are necessary conditions for equilibrium (We defer a proof of this result to Section C.2.3.)

Proposition C.1. *Every Bayes Nash equilibrium in stringent strategies is time-consistent and incentive-compatible.*

This result implies that there are no Bayes Nash equilibria in stringent strategies other than (x^*, ∞, β^*). We complete the analysis by proving that the stringent strategy profile (x^*, ∞, β^*) is a perfect Bayesian equilibrium. (We defer a proof of this result to Section C.2.4.)

Proposition C.2. *The stringent strategy profile (x^*, ∞, β^*) is a perfect Bayesian equilibrium.*

C.2.1 Proof of Lemma C.1

Let $\sigma = (\sigma_P, \sigma_A)$ be a Bayes Nash equilibrium strategy profile, where σ_P is the principal's strategy and σ_A is the disloyal agent's strategy. Suppose that $\Pr_\sigma(\tau_D < \infty) > 0$. We will derive a series of restrictions on continuation behavior after detected undermining.

Suppose first that, following detected undermining, the principal's continuation payoff is non-negative with positive probability, but on a positive-probability subset of these histories she does not immediately terminate the agent. The disloyal agent's continuation payoff following detection is nonpositive after such histories. But then as $\phi > 0$, the agent could strictly improve his continuation payoff following all such histories by undermining with maximum intensity going forward. Since the agent is detected undermining with strictly positive probability, this modification must also improve the agent's ex ante payoff over σ_A, holding fixed the principal's strategy, contradicting equilibrium. So under σ, the principal must immediately terminate the agent following detected undermining whenever her continuation payoff is non-negative.

Suppose now that, following detected undermining, the principal's payoff is strictly negative with positive probability. Then by terminating the agent immediately, the principal strictly improves her continuation payoff following such histories. Since the agent is detected undermining with strictly positive probability, this modification must strictly improve the principal's ex ante payoff, holding fixed the agent's strategy. This again contradicts the assumption that σ is an equilibrium. So the principal must receive a non-negative continuation payoff following detected undermining.

Collecting these results, if $\Pr_\sigma(\tau_D < \infty) > 0$, then necessary conditions for σ to be an equilibrium are that 1) the principal's continuation payoff following detection is zero, and 2) this continuation payoff must be delivered by immediate termination. An immediate implication is that $\Pr_\sigma(\tau_F \leq \tau_D) = 1$. To complete the proof, note that $\Pr_\sigma(\tau_F > \tau_D) > 0$ implies $\Pr_\sigma(\tau_D < \infty) > 0$. Thus if $\Pr_\sigma(\tau_F > \tau_D) > 0$, σ cannot be an equilibrium.

C.2.2 Proof of Lemma C.2

Let β be an incentive-compatible undermining policy, and let $(\pi)_{t\geq 0}$ be the associated reputation process conditional on no detection of undermining. Note that π must be weakly increasing in time given that lack of detection is (weakly) good news about the agent's loyalty. Lemma 2 implies that incentive compatibility is equivalent to $\beta_t = 1$ on $[0, \underline{t}^*) \cup [\overline{t}^*, \infty)$. We will freely use this feature of β going forward.

For each $u \in [0, \overline{U}]$ and $p \in [q, 1]$, let $V^*(u, p)$ denote the principal's optimized commitment value function, as a function of the promised utility u to the disloyal agent and the agent's initial reputation is p. This value function is as defined in the proof of Theorem 1, with the dependence of this value on the agent's reputation made explicit. Define $u^*(p) \equiv \arg\max_{u\in[0,\overline{U}]} V^*(u, p)$. We assume that q is large enough that $u^*(q) > \underline{U}$, in which case $u^*(p)$ is single-valued for all $p \in [q, 1]$ and satisfies $u^*(p) > \underline{U}$. Let U^* denote the disloyal agent's continuation value process under the stringent contract (x^*, ∞) and an optimal

undermining policy. Note that $U_0^* = u^*(q)$, and $u^*(q) < \overline{U}$ given that $q < (K-1)/K$. In an abuse of notation, for each $u \in [0, \overline{U}]$, let $x^*(u)$ be as characterized in Proposition B.1. The definition of x^{**} implies the identity $x^{**}(p) = x^*(u^*(p))$ for all p.

Time-consistency is equivalent to the requirement that $(x_s^*)_{s \geq t} = x^{**}(\pi_t)$ for all t. Suppose first that $q \geq (K-1)/K$. Then $\overline{t}^* = 0$, so that $x_t^* = 1$ and $\beta_t = 1$ for all time, and for all $t > 0$, $\pi_t > (K-1)/K$. This latter inequality implies that $x^{**}(\pi_t) = 1$. Thus (x^*, β) is time-consistent. For the remainder of the proof, assume that $q < (K-1)/K$. This implies in particular that $\overline{t}^* > 0$.

As a preliminary result, we establish that (x^*, β) is time-consistent if and only if $U_t^* = u^*(\pi_t)$ for all t. To see that time-consistency implies $U_t^* = u^*(\pi_t)$, note that time-consistency requires that for all t, $(x_s^*)_{s \geq t} = x^{**}(\pi_t)$. But $x^{**}(\pi_t) = x^*(u^*(\pi_t))$, and by definition of $x^*(u)$ the contract $(x^*(u^*(\pi_t)), \infty)$ delivers the disloyal agent an expected utility of $u^*(\pi_t)$. Hence $U_t^* = u^*(\pi_t)$ for all t. In the other direction, suppose that $U_t^* = u^*(\pi_t)$ holds for all t. By construction, the stakes curve x^* satisfies $(x_s^*)_{s \geq t} = x^*(U_t^*)$ for all t. Then if $U_t^* = u^*(\pi_t)$, it follows that $(x_s^*)_{s \geq t} = x^*(u^*(\pi_t)) = x^{**}(\pi_t)$ for all t, establishing time-consistency. Going forward, we will make free use of this equivalent characterization of time-consistency.

We now establish that time-consistency implies $\overline{t}^* = \inf\{t \ : \ \pi_t \geq (K-1)/K\}$. Note that U^* is increasing and $\overline{t}^* = \inf\{t \ : \ U_t^* = \overline{U}\}$, while $u^*(p) = \overline{U}$ if and only if $p \geq (K-1)/K$. Then for $t > \overline{t}^*$, $U_t^* = u^*(\pi_t)$ iff $\pi_t \geq (K-1)/K$. Meanwhile for $t < \overline{t}^*$ we have $U_t^* < \overline{U}$, in which case $U_t^* = u^*(\pi_t)$ implies that $\pi_t < (K-1)/K$.

Now, suppose that $\overline{t}^* = \inf\{t \ : \ \pi_t \geq (K-1)/K\}$. We will prove that under this condition, (x^*, ∞, β) is time-consistent if and only if $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \overline{t}^*)$. For $t \geq \overline{t}^*$, the hypothesized condition implies that $u^*(\pi_t) = \overline{U}$. Then as $U_t^* = \overline{U}$ for $t \geq \overline{t}^*$, the condition $U_t^* = u^*(\pi_t)$ is satisfied on this range of times. Meanwhile for $t \in (0, \overline{t}^*)$, $\pi_t < (K-1)/K$ and $u^*(\pi_t) \in (\underline{U}, \overline{U})$, so $U_t^* = u^*(\pi_t)$ if and only if U_t^* satisfies the first-order condition $\frac{\partial V^*}{\partial u}(U_t^*, \pi_t) = 0$.[25] Hence, time-consistency is satisfied if and only if $\frac{\partial V^*}{\partial u}(U_t^*, \pi_t) = 0$ for all $t < \overline{t}^*$.

Given any t and $u \in [\underline{U}, \overline{U}]$, we will define an undermining policy $\beta^t(u)$ under which both $(x^*(U_t^*), \infty, \beta^t(u))$ and $(x^*(u), \infty, \beta^t(u))$ are incentive-compatible contracts. Let $\widehat{T}^*(u)$ and $\overline{T}^*(u)$ be as defined in the proof of Proposition B.1, and note that the proof of that proposition implies that $\underline{t}^* = \min\{\widehat{T}^*(U_0^*), \overline{T}^*(U_0^*)\}$ while $\overline{t}^* = \overline{T}^*(U_0^*)$. Set

$$\beta^t(u)_s = \begin{cases} \beta_{t+s}, & s \in [\widehat{T}^*(U_t^*), \overline{T}^*(U_t^*)) \cap [\widehat{T}^*(u), \overline{T}^*(u)), \\ 1, & \text{otherwise.} \end{cases}$$

[25] As we establish in the proof of Theorem 1, the optimizer $u^*(p)$ is uniquely defined by this first-order condition whenever $u^*(p) \in (\underline{U}, \overline{U})$.

Because this undermining policy chooses an intensity less than 1 only when the agent is indifferent under both $x^*(U_t^*)$ and $x^*(u)$, it is incentive-compatible under both policies. Note that $\beta^t(U_t^*) = (\beta_s)_{s \geq t}$ for every t.

For all t and $u, u' \in [\underline{U}, \overline{U}]$, define the auxiliary value function

$$\widehat{V}(u, u', t) = \int_0^\infty x^*(u)_s e^{-rs} [\pi_t + (1 - \pi_t)(1 - K\beta^t(u')_s) e^{-\gamma \int_0^s \beta^t(u')_y \, dy}] \, ds. \tag{C.1}$$

$\widehat{V}(u, u', t)$ is the principal's payoff when initial beliefs are π_t, the principal chooses contract $(x^*(u), \infty)$, and the disloyal agent chooses undermining policy $\beta^t(u')$. Note that $\widehat{V}(u, u, t) = V^*(u, \pi_t)$ for any t and u given that $\beta^t(u)$ is, by construction, incentive-compatible under $x^*(u)$. Further, $\widehat{V}(U_t^*, u', t) = V^*(U_t^*, \pi_t)$ for any t and u' given that $\beta^t(u')$ is, by construction, incentive-compatible under $x^*(U_t^*)$.

The derivative $\frac{\partial V^*}{\partial u}(U_t^*, \pi_t)$ can be written in terms of $\widehat{V}$ as

$$\frac{\partial V^*}{\partial u}(U_t^*, \pi_t) = \frac{\partial \widehat{V}}{\partial u}(U_t^*, U_t^*, t) + \frac{\partial \widehat{V}}{\partial u'}(U_t^*, U_t^*, t).$$

As noted earlier, $\widehat{V}(U_t^*, u', t)$ is independent of u', and so the second derivative in the previous expansion vanishes. It is straightforward to calculate the remaining derivative, yielding

$$\begin{aligned} \frac{\partial V^*}{\partial u}(U_t^*, \pi_t) = & \int_{\min\{\widehat{T}^*(U_t^*), \overline{T}^*(U_t^*)\}}^{\overline{T}^*(U_t^*)} \left. \frac{\partial x^*(u)_s}{\partial u} \right|_{u = U_t^*} e^{-rs} [\pi_t + (1 - \pi_t)(1 - K\beta_{t+s}) e^{-\gamma \int_0^s \beta_{t+y} \, dy}] \, ds \\ & - \frac{\partial \overline{T}^*}{\partial u}(U_t^*) \Delta x^*(U_t^*)_{\overline{T}^*(U_t^*)} e^{-r\overline{T}^*(U_t^*)} [\pi_t + (1 - \pi_t)(1 - K) e^{-\gamma \int_0^{\overline{T}^*(U_t^*)} \beta_{t+y} \, dy}], \end{aligned} \tag{C.2}$$

where $\Delta x^*(u)_t \equiv x^*(u)_t - x^*(u)_{t-}$. This calculation relies on several basic properties of $x^*(u)$: $x^*(u)_t = \phi$ for $t < \min\{\widehat{T}^*(U_t^*), \overline{T}^*(U_t^*)\}$; $x^*(u)_t = 1$ for $t \geq \overline{T}^*(U_t^*)$; $x^*(u)_t$ is differentiable wrt u for all $t \in (\widehat{T}^*(u), \overline{T}^*(u))$; and $x^*(u)_t$ is continuous except at $\overline{T}^*(U_t^*)$. It further relies on the fact that $\overline{T}^*$ is differentiable at U_t^* for all $t \in (0, \bar{t})$.

Bayes' rule and the identity $\bar{t}^* = t + \overline{T}^*(U_t^*)$ for $t < \bar{t}$ imply that

$$\pi_{\bar{t}^*} = \pi_{t + \overline{T}^*(U_t^*)} = \pi_t \Big/ \left(\pi_t + (1 - \pi_t) \exp\left(-\gamma \int_0^{\overline{T}^*(U_t^*)} \beta_{t+y} \, dy \right) \right).$$

Meanwhile, by assumption $\pi_{\bar{t}^*} = (K - 1)/K$. Hence the final term in (C.2) vanishes. Moreover, for $s \in (\widehat{T}^*(u), \overline{T}^*(u))$, $x^*(u)_s = \frac{\gamma \widehat{U}}{K} e^{(r + \gamma/K)s}$. Hence $\frac{\partial V^*}{\partial u}(U_t^*, \pi_t) = 0$ for all $t < \bar{t}^*$ is

equivalent to $F(t) = 0$ for all $t < \overline{t}^*$, where

$$F(t) \equiv \int_{\min\{\widehat{T}^*(U_t^*), \overline{T}^*(U_t^*)\}}^{\overline{T}^*(U_t^*)} e^{\gamma s/K} \left[\pi_t + (1-\pi_t)(1-K\beta_{t+s}) \exp\left(-\gamma \int_0^s \beta_{t+y}\, dy\right)\right] ds.$$

If $\widehat{U} \geq \overline{U}$, with $\widehat{U}$ as defined in the proof of Theorem 1, then $\widehat{T}^*(U_t^*) \geq \overline{T}^*(U_t^*)$ for all t and trivially $F(t) = 0$ for all $t < \overline{t}^*$. On the other hand, when $\widehat{U} < \overline{U}$ then $\widehat{T}^*(U_t^*) \leq \overline{T}^*(U_t^*)$ for all t. Then using the identities $\overline{T}^*(U_t^*) = \overline{T}^*(U_0^*) - t$ and $\widehat{T}^*(U_t^*) = \max\{\widehat{T}^*(U_0^*) - t, 0\} = \max\{\underline{t}^* - t, 0\}$ holding for all $t < \overline{t}^*$, $F(t)$ may be equivalently written using a change of variables as

$$F(t) = \int_{\max\{\underline{t}^*, t\}}^{\overline{t}^*} e^{\gamma(s'-t)/K} \left[\pi_t + (1-\pi_t)(1-K\beta_{s'}) \exp\left(-\gamma \int_t^{s'} \beta_{y'}\, dy'\right)\right] ds'.$$

Differentiating this expression wrt t and using the Bayes' rule identity $\dot{\pi}_t = \gamma\beta_t\pi_t(1-\pi_t)$ yields derivative

$$F'(t) = -(\pi_t + (1-\pi_t)(1-K\beta_t))\mathbf{1}\{t \geq \underline{t}^*\} + \gamma\left(\beta_t(1-\pi_t) - \frac{1}{K}\right) F(t)$$

for all $t \neq \underline{t}^*$. Now, $F(t) = 0$ for $t < \overline{t}^*$ implies that $F'(t) = 0$ for $t \in [0, \overline{t}^*) \setminus \{\underline{t}^*\}$. Hence the previous expression reduces to $\pi_t + (1-\pi_t)(1-K\beta_t) = 0$ for all $t \in (\underline{t}^*, \overline{t}^*)$, which after rearrangement yields $\beta_t = 1/(K(1-\pi_t))$.

In the other direction, suppose that $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \overline{t}^*)$. In that case, the previous expression for $F'(t)$ implies that $F'(t) = 0$ for all $t \in [0, \overline{t}^*) \setminus \{\underline{t}^*\}$. Since the boundary condition $F(\overline{t}^*) = 0$ holds trivially, it must be that $F(t) = 0$ for all $t < \overline{t}^*$. Thus $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \overline{t}^*)$ is necessary and sufficient for time-consistency.

We next show that if $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \overline{t}^*)$, then $\overline{t}^* = \inf\{t \,:\, \pi \geq (K-1)/K\}$, so that the second condition is redundant for characterizing time-consistency. Given any incentive-compatible undermining policy, (C.2) holds at $t = 0$ regardless of the value of $\pi_{\overline{t}^*}$. Further, if $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \overline{t}^*)$, then the logic above implies that the first term in this identity vanishes. Thus

$$\frac{\partial V^*}{\partial u}(U_0^*, q) = -\frac{\partial \overline{T}^*}{\partial u}(U_0^*)\Delta x_{\overline{t}^*}^* e^{-r\overline{t}^*}\left(q + (1-q)(1-K)e^{-\gamma\int_0^{\overline{t}^*}\beta_t\, dt}\right).$$

Using Bayes' rule, this is equivalently

$$\frac{\partial V^*}{\partial u}(U_0^*, q) = -\frac{\partial \overline{T}^*}{\partial u}(U_0^*)\Delta x_{\overline{t}^*}^* e^{-r\overline{t}^*} q\left(1 + (1-K)\frac{1-\pi_{\overline{t}^*}}{\pi_{\overline{t}^*}}\right).$$

It is easy to verify that $\partial\overline{T}^*/\partial u < 0$ for all u. Further, $\Delta x^*_{\bar{t}^*} > 0$, and $\frac{\partial V^*}{\partial u}(U^*_0, q) = 0$ given that $q < (K-1)/K$. Then this identity implies that $(K-1)(1-\pi_{\bar{t}^*})/\pi_{\bar{t}^*} = 1$, or equivalently $\pi_{\bar{t}^*} = (K-1)/K$, as desired.

We complete the proof by showing that there exists a unique incentive-compatible undermining policy β^*, with associated belief process π^*, satisfying the necessary and sufficient condition for time-consistency $\beta^*_t = 1/(K(1-\pi^*_t))$ for almost all $t \in (\underline{t}^*, \bar{t}^*)$. Incentive-compatibility requires that $\beta^*_t = 1$ for $t \in [0, \underline{t}^*) \cup [\bar{t}^*, \infty)$. So consider times $t \in (\underline{t}^*, \bar{t}^*)$. Substituting $\beta^*_t = 1/(K(1-\pi^*_t))$ into the Bayes' rule ODE $\dot{\pi}^*_t = \gamma\beta^*_t\pi^*_t(1-\pi^*_t)$ yields $\dot{\pi}^*_t = \gamma\pi^*_t/K$. Let $(q'_t)_{t\geq 0}$ be the unique Carathéodory solution to the initial value problem

$$\dot{\pi}_t = \begin{cases} \gamma\pi_t(1-\pi_t), & 0 \leq t < \underline{t}^*, \\ \gamma\pi_t/K, & \underline{t}^* \leq t < \bar{t}^* \\ \gamma\pi_t(1-\pi_t), & \bar{t}^* \leq t \end{cases}$$

with initial condition $\pi_0 = q$. Define the undermining process β^* by $\beta^*_t = 1$ on $[0, \underline{t}^*) \cup [\bar{t}^*, \infty)$ and $\beta^*_t = 1/(K(1-q'_t))$ on $[\underline{t}^*, \bar{t}^*)$. Then provided that $\beta^*_t \in [0,1]$ for all t, as we verify below, this construction ensures that β^* is the essentially unique undermining process inducing posterior belief process $\pi^* = q'$. Uniqueness follows from the requirement that realized paths of undermining be càdlàg. Therefore if β^* is feasible, it is the unique undermining process which is incentive-compatible and satisfies $\beta_t = 1/(K(1-\pi_t))$ for almost all $t \in (\underline{t}^*, \bar{t}^*)$.

Feasibility is automatic if $\underline{t}^* = \bar{t}^*$, so suppose that $\underline{t}^* < \bar{t}^*$. Let $t^\dagger \equiv \inf\{t \ : \ q'_t \geq (K-1)/K\}$. If $t^\dagger \geq \bar{t}^*$, then feasibility is ensured, so suppose that $t^\dagger < \bar{t}^*$. We derive a contradiction by showing that $\frac{\partial V^*(U^*_0, q)}{\partial u} > 0$. Construct an undermining process $\beta^\dagger$ by $\beta^\dagger_t = \beta^*_t$ for $t < t^\dagger$ and $\beta^\dagger_t = 1$ for $t \geq t^\dagger$. Let $\pi^\dagger$ be the associated posterior belief process. As $\beta^\dagger$ is an incentive-compatible undermining process, (C.2) evaluated at $t = 0$ implies that

$$\frac{\partial V^*}{\partial u}(U^*_0, q) = F^\dagger(0) - \frac{\partial \overline{T}^*}{\partial u}(U^*_0)\Delta x^*_{\bar{t}^*} e^{-r\bar{t}^*} q \left(1 + (1-K)\frac{1-\pi^\dagger_{\bar{t}^*}}{\pi^\dagger_{\bar{t}^*}}\right),$$

where $F^\dagger(t)$ satisfies the ODE

$$\frac{dF^\dagger}{dt} = -(\pi^\dagger_t + (1-\pi^\dagger_t)(1-K\beta^\dagger_t))\mathbf{1}\{t \geq \underline{t}^*\} + \gamma\left(\beta^\dagger_t(1-\pi^\dagger_t) - \frac{1}{K}\right)F^\dagger(t)$$

and boundary condition $F^\dagger(\bar{t}^*) = 0$. Note that $F^\ddagger(t) = 0$ is a supersolution to this ODE for all $t < \bar{t}^*$, a strict supersolution for $t \geq t^\dagger$, and satisfies the same terminal condition as $F^\dagger$. Hence $F^\dagger(0) > F^\ddagger(0) = 0$. As additionally $\pi^\dagger_{\bar{t}^*} > (K-1)/K$ while $\partial\overline{T}^*/\partial u < 0$, it follows

that $\frac{\partial V^*}{\partial u}(U_0^*, q) > 0$, a contradiction. So it must be that $t^\dagger \geq \overline{t}^*$, as desired.

C.2.3 Proof of Proposition C.1

Fix any Bayes Nash equilibrium in stringent strategies $\sigma = (x, T_F, \beta)$. Let π be the path of the principal's beliefs, conditional on no observed undermining, induced by the undermining policy β. Also let $V^*(p)$ be the principal's optimal commitment payoff when the agent is loyal with prior probability $p \in [0, 1]$.

Suppose first that σ is not incentive-compatible. Then there exists another undermining policy $\widetilde{\beta}$ which strictly increases the disloyal agent's ex ante payoff, supposing that the principal commits to the stringent contract (x, T_F). But as the equilibrium response to detected undermining is identical to the commitment response, $\widetilde{\beta}$ must also strictly increase the disloyal agent's ex ante payoff given the principal's equilibrium strategy. This contradicts the assumption that σ is a Bayes Nash equilibrium, so (x, β) must be incentive-compatible.

Now suppose that σ is incentive-compatible. We argue that $T_F = \infty$. Suppose by way of contradiction that $T_F < \infty$. Note that $\pi_{T_F} \geq \pi_0 = q$, and by assumption $V^*(q) > 0$, so that also $V^*(\pi_{T_F}) > 0$ given that $V^*(p)$ is weakly increasing in p. Then by deviating to $T_F = \infty$ and the continuation stakes curve $x^{**}(\pi_{T_F})$, the principal can strictly improve her continuation payoff at time T_F. Since the principal must reach time T_F without detecting undermining with strictly positive probability, this deviation strictly increases her ex ante payoff. This result contradicts the assumption that σ is an equilibrium, so $T_F = \infty$.

Next, we argue that for all $t \geq 0$, $(x_s)_{s \geq t} = x^{**}(\pi_t)$, with establishes time-consistency. Fix any $t \geq 0$. We first claim that the principal's continuation payoff must be at least $V^*(\pi_t)$. Note that a feasible continuation strategy is to play x^{**} forever after and terminate the agent immediately if undermining is detected. As $V^*(\pi_t)$ is the principal's payoff from such a policy when the agent responds by *minimizing* her payoff, it is a lower bound on the principal's payoff from this continuation strategy, and thus a lower bound on her continuation payoff. Next, we claim that $V^*(\pi_t)$ is also an upper bound on her continuation value. By assumption, $(x_s, \beta_s)_{s \geq t}$ is incentive-compatible, so under commitment the principal could offer $(x_s)_{s \geq t}$, recommend $(\beta_s)_{s \geq t}$, and terminate the agent immediately upon detection (and no sooner). Thus, the agent's continuation payoff must be *at most* $V^*(\pi_t)$, so we have shown that it must equal $V^*(\pi_t)$. Since $x^{**}(\pi_t)$ is the uniquely optimal stakes curve in the commitment problem, $(x_s)_{s \geq t} = x^{**}(\pi_t)$.

C.2.4 Proof of Proposition C.2

Throughout this proof, we will make use of the following reference undermining and belief paths. Define a function $\beta^\dagger : [0,1] \to [0,1]$ by

$$\beta^\dagger(p) = \beta^{**}(p)_0.$$

$\beta^\dagger$ will be used to define a Markovian undermining policy for the agent for any posterior belief p. Note that under $\beta^\dagger$, the agent plays a best response assuming the principal plays the optimal commitment stakes curve $x^{**}(p)$ going forward. For each belief p, define an associated posterior belief process $\pi(p)$ as the unique solution to the ODE

$$\dot{\pi}(p)_t = \gamma \beta^\dagger(\pi(p)_t)\pi(p)_t(1 - \pi(p)_t)$$

with initial condition $\pi(p)_0 = p$. Note in particular that $\pi(p)$ satisfies Bayes' rule: $\pi(p)_t = p/Q(p)_t$ for all times, where

$$Q(p)_t \equiv p + (1-p)\exp\left(-\gamma \int_0^t \beta^\dagger(\pi(p)_s)\, ds\right).$$

Additionally, by construction $\pi(q)_t = \pi^*_t$ for all time, where π^* is the belief process constructed in the proof of Lemma C.2. Thus, as established in that proof, $\pi(q)_{\bar{t}^*} = (K-1)/K$. An analogous property holds for $\pi(p)$ evaluated at the first time $x^{**}(p)$ reaches 1.

Define a principal strategy (χ, Λ) by $\chi_t = x^*_t$ and $\Lambda = \inf\{t \ : \ N_t > 0\}$. Define an $\mathbb{F}^A$-adapted reference belief process $\widehat{\pi}$ as $\widehat{\pi}_t = 0$ whenever $N_t > 0$, with $\widehat{\pi}_t = \pi(q)_t$ otherwise. Define an agent strategy ζ by $\zeta_t = \beta^\dagger(\widehat{\pi}_t)$ if $N_t = 0$, and $\zeta_t = 1$ otherwise. Several properties of this strategy will be important in what follows. First, π^ζ satisfies the same ODE as $\widehat{\pi}$ when $N_t = 0$, and both are zero whenever $N_t > 0$, and so $\pi^\zeta = \widehat{\pi}$. Second, for any time t, if $(N_s)_{s \geq t} = 0$, then $(\zeta_s)_{s \geq t} = \beta^{**}(\widehat{\pi}_t)$.

We prove that (χ, Λ, ζ) constitute a perfect Bayesian equilibrium. Consider first the agent's strategy. Fix a time t and any history following which $N_t > 0$. Then the agent's strategy is trivially a best response to immediate firing. So suppose $N_t = 0$ and $\widehat{\pi}_t = p \in [0,1]$. The final property of the previous paragraph implies that if $(N_s)_{s \geq t} = 0$, then $(\zeta_s)_{s \geq t} = \beta^{**}(p)$. Thus ζ induces an undermining policy which, by definition of β^{**}, is a best response to a strategy which induces $(X_s)_{s \geq t} = x^{**}(p)$ and fires the first time that $N_t > 0$. In particular, ζ is a best response to (χ, Λ) in the time-zero game, as $\chi = x^* = x^{**}(q)$ and $\Lambda = \inf\{t \ : \ N_t > 0\}$.

Now consider the principal's strategy. Fix a time t and a history following which $N_t > 0$.

In this case $\widehat{\pi}_t = 0$ and the disloyal agent undermines unconditionally at all future times under ζ, and so the principal's unique best response is immediate termination. So consider instead histories following which $N_t = 0$ and current beliefs are $p \in [0, 1]$. We will establish that 1) if $p \geq \underline{q}(\phi)$, then $x^{**}(p)$ with firing after observing undermining is a best response to $(\zeta_s)_{s \geq t}$, and 2) if $p < \underline{q}(\phi)$, then immediate firing is a best response. This result will complete the proof that (χ, Λ, ζ) is a PBE, as in particular it implies that the strategy profile is a Bayes Nash equilibrium, and that the agent's continuation strategy in every continuation game is part of a Bayes Nash equilibrium of that game.

The optimality of firing following detected undermining is immediate, because the principal forms a posterior belief that the agent is certainly disloyal and expects the agent to undermine with full intensity forever afterward. So we need only characterize optimal behavior prior to observing undermining. Since the agent's reputation rises over time, achievable continuation payoffs rise over time in the absence of detected undermining. Thus an optimal firing policy either fires immediately, or else never fires prior to uncovering undermining.

We characterize the principal's optimal stakes curve assuming the principal does not fire the agent prior to detecting undermining, and show that it is $x^{**}(p)$. Given that this stakes curve induces the undermining policy $\beta^{**}(p)$, the principal's payoff from firing only after uncovering undermining is therefore exactly as in the commitment game. It follows that immediate firing is an optimal policy iff $p < \underline{q}(\phi)$. For the remainder of the proof we ignore the possibility of immediate firing, and characterize the optimal stakes path assuming firing only after observing undermining.

The agent's strategy is Markovian in the state $\widehat{\pi}_t$, with $\beta_t = \beta^\dagger(\widehat{\pi}_t)$, and the evolution of $\widehat{\pi}$ similarly depends only on $\widehat{\pi}_t$. So $(x_s)_{s \geq t}$ is a best response to ζ in a continuation game with $\widehat{\pi}_t = p$ if and only if it solves

$$\sup_{x \in \mathbb{X}} \int_0^\infty v(t; p) x_t \, dt,$$

where

$$\begin{aligned} v(t; p) &\equiv p e^{-rt} + (1 - p) \exp\left(-rt - \int_0^t \beta^\dagger(\pi(p)_s) \, ds\right) (1 - K\beta^\dagger(\pi(p)_t)) \\ &= e^{-rt} Q(p)_t [\pi(p)_t + (1 - \pi(p)_t)(1 - K\beta^\dagger(\pi(p)_t))]. \end{aligned}$$

Define $\underline{t}(p) \equiv \min\{\widehat{T}^*(u^*(p)), \overline{T}^*(u^*(p))\}$ and $\bar{t}(p) \equiv \overline{T}^*(u^*(p))$. We complete the proof by showing that $x^{**}(p)$ maximizes this objective function pointwise. Note that $(\pi(p)_t)_{t \geq 0}$ is strictly increasing in time, and Lemma C.2 established that $\pi(p)_{\bar{t}(p)} = (K-1)/K$. Therefore:

- For $t \in [0, \underline{t}(p))$, we have $\beta^\dagger(\pi(p)_t) = \beta^{**}(p)_t = 1$ and $\pi(p)_t < (K - 1)/K$. Thus,

$v(t;p)_t < 0$ on this interval, and $v(t;p)_t x_t$ is maximized by $x_t = \phi = x^{**}(p)_t$.

- For $t \in [\underline{t}(p), \bar{t}(p))$, we have $\beta^\dagger(\pi(p)_t) = \beta^{**}(p)_t = 1/(K(1-\pi(p)_t))$, so $v(t;p)_t = 0$, and all $x_t \in [\phi, 1]$ are optimal, including $x^{**}(p)_t$.

- For $t \in [\bar{t}(p), \infty)$, we have $\beta^\dagger(\pi(p)_t) = \beta^{**}(p)_t = 1$ and $\pi(p)_t \geq (K-1)/K$, so $v(t;p)_t \geq 0$ and $v(t;p)_t x_t$ is maximized by $x_t = 1 = x^{**}(p)_t$.