Centralization versus separation of regulatory institutions

Dana Foarta
Stanford GSB

Takuo Sugaya
Stanford GSB

October 8, 2016

Abstract

We examine the optimal institutional design in an environment in which an informed agent with reputational concerns can exert costly effort to influence an outcome, and an uninformed decision-maker can choose policy that affects the final outcome. We show that when agent's information is not observable to the decision-maker, the agent's ability to manipulate information leads to higher effort provision; however, information withholding may prevent the decision maker from choosing the optimal policy. In a dynamic environment, the optimal institutional setup varies as a function of the agent's reputation: institutional separation – in which the agent's information is not observable by the decision-maker – is optimal for intermediate reputation levels, while institutional centralization – in which the decision maker and the agent have access to the same information – is optimal for very low and very high reputation.

1 Introduction

A long-standing question for academics and policymakers has been how to optimally design regulatory institutions when it comes to the choice of centralization versus institutional separation. Two consideration that are relevant for this decision are the need for information sharing between different institutions and the importance of clearly separating regulatory tasks. The separation of regulatory tasks can lead to a clearer delimitation of regulatory tasks between different agencies; however, it requires can also lead to worse transmission of information between institutions when the information is collected by one agency and later utilized by another. Separation of regulatory tasks incentivizes regulators to perform well by making it easier for the government to link their reward to informative observations about their performance; however, separation also creates incentives for regulators to withhold or manipulate information transmission between institutions whenever their objectives differ, as shown in the canonical cheap talk model of Crawford and Sobel (1982).

This paper builds a model to shed light on the trade-off between regulatory effort and information transmission. We then ask, given this trade-off, when is it optimal to have centralized rather than separate regulatory institutions, and how does this choice change over time?

The relevance of these questions has become evident in the ongoing policy debates about the architecture of financial regulation. One illustrative debate is that about the relationship between the institutions of bank supervision and the lender of last resort. Bank supervisors monitor risk taking in the banking sector. Their effort in monitoring is meant to reduce excessive risk-taking – banks that are closely monitored are prevented from taking on too much risk – and the probability of a banking crisis – by monitoring, these regulators gather information on the health of the banking system. The lender of last resort uses the information provided by the bank supervisors, and – if there are signs of a crisis – it can intervene by providing funds to reduce the losses from a banking crisis. Therefore, these two institutions are closely linked by their role in preventing financial crises. While this link exists across financial systems, the institutional relationship between bank supervision and the lender of

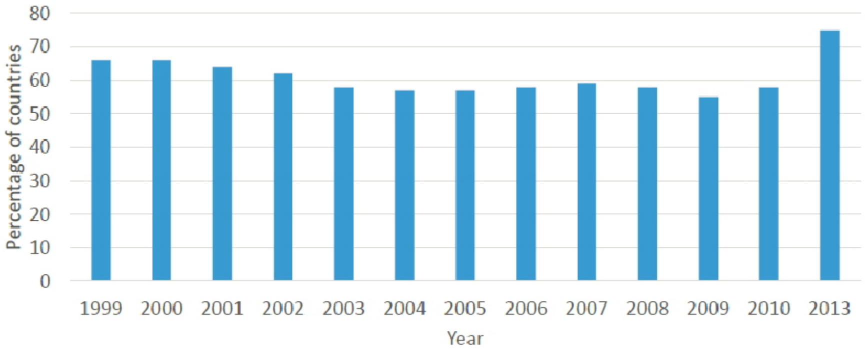

Figure 1: Percentage of countries with institutional centralization of bank supervision inside the central bank

last resort – usually the central bank – varies across countries, as illustrated in Figure 1[1]. In a majority of cases (75% in 2013), the functions of bank supervision and lender of last resort are both performed by the central bank, while in the rest, bank supervision is performed by an institution separate from the lender of last resort. Moreover, several waves of reforms have led countries to switch between institutional structures. In aggregate, over the 1999-2009 period, there has been a shift towards the separated institutional structure, while in the period since 2009 there has been more centralization. The United Kingdom, for instance, moved from a system with institutional centralization to one with separation in 1997/98, with the creation of the Financial Services Authority (FSA) as an institution separated from the Bank of England. This system was changed again in 2010/11, when the responsibility for bank supervision was transferred back to the Bank of England.

We consider the following model that captures these institutional features. An informed agent – called the regulator – can exert costly effort to increase the probability of a good outcome. How costly the effort is depends on the regulator's private type, which can be high or low. After effort is sunk, the regulator observes a noisy signal about whether the final outcome will be good or bad – a banking crisis . Another agent, an uninformed decision maker – called the central bank– can take a costly action – intervene as a lender of last

[1] The data is obtained from "How countries supervise their banking, insurers and securities markets," Central Banking Publications, 2013, and Melecky and Podpiera (2013). It covers 98 countries.

resort– before the final outcome is revealed. The cost of the action is lower than the cost of the bad outcome but higher the cost of the good outcome. Therefore, the central bank wants to intervene if the final outcome is likely to be bad, but she does not want to intervene if the final outcome is good. In a centralized institutional structure, the noisy signal observed by the regulator is also observable to the central bank. In a separated institutional structure, the noisy signal is only observed by the regulator, who can send a message to the central bank about it. Finally, the public in this model only observes the final outcome – the good outcome, the bad outcome or an intervention by the central bank. The regulator has reputational concerns – his payoff depends on the beliefs that the public forms about him, while the central bank only cares about minimizing costs. In a dynamic setting, this sequence of actions is repeated every period and the public updates their belief about the regulator each period. After each period, the public can replace the regulator with a new one.

The model highlights the key trade-off described in the motivation. With separated institutions, reputational concerns lead the agent to withhold information – to not report bad news to the decision maker, in order to 'gamble' for the good outcome. This happens because a good outcome without the decision maker's intervention improves the agent's reputation, while the decision maker's intervention hurts it. Even though a final bad outcome without intervention would be the most detrimental to his reputation, if the intervention leads to a sufficiently bad outcome for him anyway, the agent then prefers to gamble for a possibility that the bad news is a false alarm, and the final outcome will actually be good. In a centralized institutional structure, such a gamble is no longer feasible, since the news about a potential bad outcome is seen directly by the decision maker. This allows the decision maker to intervene in case of bad news, but it reduces the effort put in by the agent.

The model therefore captures the trade-off in choosing an institutional structure. The ability to manipulate information transmission increases effort, and through this it decreases the ex-ante probability of a costly outcome; however, the ability to manipulate information transmission prevents the decision maker from using the agent's information efficiently ex post.

The debate about the reform of the Bank of England reflects this trade-off. When the regulatory agency – the FSA – was separated in 1997/1998, the main argument provided for this change was to "reduc[e] the risk that the central bank would be too accommodative in order to help frail intermediaries."[2] An implication is that too much accommodation might reduce supervisory effort, since an accommodative policy from the central bank makes low supervisory effort less visible. On the other hand, when the supervisory functions were transferred back to the Bank of England in 2011, the main reason cited was the lack of shared information between FSA and the Bank of England during the financial crisis.[3]

We show that this static trade-off is further amplified by the dynamic setting due to the role of reputation building – the update in the public's belief about the agent. When there is institutional separation, the final – good or bad – outcome is reached more often, and the final outcome is more informative as to the effort of the agent. Hence, as in the static model, this 'informativeness' advantage of separated institutions means that, all else equal, the same effort by the agent can be implemented more efficiently when institutions are separated as opposed to centralized. In addition, in the dynamic model, this 'informativeness' means that the public has better information about the type of the agent in equilibrium since different types choose different levels of effort. This improved screening enables the public to make more informed decisions about replacement.

With this 'amplified trade-off', we obtain the following key result. The optimal equilibrium can feature cycling between institutional separation and centralization. Separated institutions are welfare maximizing for the public when the agent's reputation is intermediate. The centralized institution is welfare maximizing when the agent has very low reputation or very high reputation. Separated institutions are shown to dominate the centralized institution setup when observations of the final outcome are most informative for the public.

[2] The quote is from the executive summary of Chapter 14 of "Money, Banking, and Financial Markets" by Stephen G. Cecchetti and Kermit L. Schoenholtz, available at
http://www.mhhe.com/economics/cecchetti/Cecchetti2_Ch14_RegDivofLabor.pdf

[3] For example, see "Responsibility without information: the Bank of England as Lender of Last Resort" in Financial Times, available at
http://blogs.ft.com/maverecon/2007/09/responsibility-html/#axzz4M3r4Tqv9

That is exactly when the public is uncertain about the type of the agent – when reputation is actively being updated; however, once the agent's reputation is sufficiently high or low, the public is fairly convinced about his type and additional updates are not as valuable. We show that this result emerges both in the best Perfect Bayesian Equilibrium and in the optimal dynamic mechanism design problem. The results of the model suggest that transitions between institutional separation and centralization can be welfare improving in environments in which society must balance incentives for regulatory effort and information transmission between institutions.

Related Literature The paper connects to several strands of literature. First, our paper contributes to the literature on strategic information transmission between an informed agent and an uninformed policymaker (Crawford and Sobel, 1982; Benabou and Laroque, 1992; Morris, 2001). We use a version of the canonical cheap talk model of Crawford and Sobel (1982). This model has been used to analyze the interaction between information and policy formation, and has been applied to the question of institutional design, for instance in examining the relationship between the legislative and the executive (Austen-Smith, 1990), or the choice of delegation or restriction on amendments (Bendor and Meirowitz, 2004; Ambrus et al., 2013). Most of the papers examine the relationship between the costly effort to acquire information by the informed party and the commitment of the uninformed party to take actions. Instead, in our paper, the information about the crisis comes for free to the regulator. Hence, for the efficient usage of information, it is optimal for the decision maker to directly observe the information. However, the regulatory agency has reputational concerns and takes costly ex-ante effort to prevent a bad outcome. We focus on the interaction between the messaging strategy, which follows the structure of Crawford and Sobel (1982), and the ex-ante effort choice of the agent, and show that the incentives to put in more effort in order to gamble for a good outcome sometimes makes it optimal for the decision maker not to observe the signal directly. Moreover, we examine this trade-off in a fully dynamic model, deriving both the best sustainable equilibrium for the public as well as the optimal mechanism. In terms of its theoretical contribution, the problem provides the solution to a dynamic mechanism design problem with adverse selection and moral hazard, without monetary transfers.

Second, the main motivation of the paper is to explore a set of questions that belong to the broader literature on the allocation of decision-making powers within government (Maskin and Tirole, 2004; Aghion et al., 2004; Alesina and Tabellini, 2007). In line with this literature, we examine the effects of public beliefs about policymakers on incentives and policies. Public beliefs about policymakers' ability create reputational concerns, which in turn affect the transmission of information between policymakers. We therefore focus on the understudied area of how reputational concerns affect the organizational structure of regulation.

The paper is also linked to the literature that studies the structure of financial regulation and its relationship to the central bank (Boot and Thakor, 1993; Peek et al., 1999). We add to this literature the insight that separating bank supervision from the central bank involves an additional trade-off, that of information withholding, and this trade-off may make separation suboptimal. Moreover, we consider a dynamic setting, and show that cycles between the two institutional settings are optimal.

Another related stream of literature on information transmission has employed the Bayesian persuasion approach of Kamenica and Gentzkow (2011). We depart from this literature by focusing on information transmission in a game in which we do not assume commitment. Moreover, even if the agent has commitment power in sending messages, our main result does not change. The agent wants to commit not to reveal bad signals to the decision maker since he wants to gamble for a good outcome even ex ante. This creates a trade-off between the ex-ante incentive to put in more effort and the ex-post efficient use of information, as explained above.

The rest of the paper is organized as follows. Section 2 described the model .Section 3 presents a baseline version of the model in a two period framework with a specific functional form for the supervisor's utility, and section 4 presents the dynamic model. Section 5 concludes, and the Appendix contains the proofs and extensions.

2 Environment

We begin by describing an environment that captures the main forces of the model. There are three players: an informed agent (A), a decision maker (D), and a unitary public (P).

The Informed Agent. The informed agent A has a type $\theta \in \{L, H\}$, which is private information. The type θ has value H with commonly known probability μ. The agent A can put in unobservable effort $e \in [0, 1]$ to influence the probability distribution of a final outcome $y \in Y = \{B, G\}$. Effort is costly for A, and the cost is higher for type L than for type H. Let $c_\theta(e)$ be the cost of implementing e for A of type θ. We assume the following about the cost of effort.

Assumption 1 *The functions $c_H(e)$ and $c_L(e)$ are continuously differentiable, $c'_H(0) = 0$, $\lim_{e \to 1} c'_\theta(e) = \infty$ for each θ, and*

$$c'_\theta(e) > 0,\ c''_\theta(e) > 0,\ c'_H(e) < c'_L(e)\ \ \forall e > 0. \tag{1}$$

The probability of outcome $y = G$ is an increasing and weakly concave function of the effort e: $\Pr(y = G|e) = q(e)$, $q'(e) \geq 0$, and $q''(e) \leq 0$.

After putting in the effort e, the agent A receives a noisy signal $s \in S = \{B, G\}$ about the outcome y. The error in the signal is:[4]

$$\Pr(s = B|y = G, e) = \Pr(s = G|y = B, e) = \varepsilon > 0.$$

The institutional structure determines how the signal s can be used by A: in the separated institutional structure, the agent A can then send a message $m(s) \in M$ to the decision maker

[4] All the results go through if the probability of error depends on y: $\Pr(s = B|y = G, e) = \varepsilon$ and $\Pr(s = G|y = B, e) = \varepsilon$.

D about the signal s and effort e; in the centralized institutional structure, no message is sent, since s is observed directly by D as well. Since we employ the cheap talk message à la Crawford and Sobel (1982), the message space M is arbitrary.

The agent A receives a payoff that depends on the expectation formed by the public about his type and his cost of effort $c_\theta(e)$:

$$u_\theta = u^A(b) - c_\theta(e), \tag{2}$$

where $b \equiv (\mathbb{E}[\Pr(\theta = H|\mathit{public\ outcome})], \mathit{public\ outcome})$, such that $u^A(b)$ is a function of the expectation formed by the public about A's type θ, and we also allow the utility to depend directly on the public outcome.

Given the interpretation of A as a regulator, the utility function captures the benefit to the regulator from having a good reputation with the public. For example, a bad reputation can translate into the regulator being replaced, or his position being restructured in costly ways, as shown in the example of the regulatory reform in the UK.

The Decision Marker. After A puts in effort e and receives signal s, but before the final outcome y is revealed, a decision maker D implements a policy $\iota \in I = \{\iota^G, \iota^B\}$. For each $y \in Y$, there exists a policy $\iota^y \in I$ that minimizes cost to D if the final outcome is y. As explained above, institutional structure determines the information D receives before deciding policy ι: in the separated institutional structure, the decision maker D receives the message $m(s)$ from A; in the centralized institutional structure, she observes s directly.

We assume the following costs associated with the policy ι at outcome y:

$$C = \begin{cases} 0 & \text{if} \quad \iota = \iota^G \ \mathit{and}\ y = G, \\ C_1 & \text{if} \quad \iota = \iota^B, \\ C_2 & \text{if} \quad \iota = \iota^G \ \mathit{and}\ y = B, \end{cases} \tag{3}$$

where $C_1, C_2 \in \mathbb{R}$, $0 < C_1 < C_2$. This specification captures the many situations in which early, preventive intervention is less costly than letting the situation worsen to a crisis point;

however, the actual cost savings happen only if the situation would indeed worsen without intervention, otherwise costs would be saved by not performing an unnecessary preventive intervention.

In the interpretation of D as a lender of last resort that can intervene to reduce the cost of a crisis, C_1 denotes the cost of an intervention in the market to prevent a crisis when there are signs of distress, while $C_2 > C_1$ denotes the cost of intervention when a crisis is unfolding.

We make the following assumption about D's costs:

Assumption 2 *The values C_1 and C_2 satisfy:*

$$1+\frac{q\left(1\right)}{1-q\left(1\right)}\frac{\varepsilon}{1-\varepsilon}<\frac{C_2}{C_1}<1+\frac{q\left(0\right)}{1-q\left(0\right)}\frac{1-\varepsilon}{\varepsilon} \tag{4}$$

and

$$1+\frac{q\left(0\right)}{1-q\left(0\right)}<\frac{C_2}{C_1}. \tag{5}$$

(4) of Assumption 2 implies that it is less costly to choose policy ι^B after a signal $s = B$ rather than ι^G, regardless of the A's effort. In addition, the expected cost of choosing ι^G is lower than the cost of ι^B whenever the signal is $s = G$, regardless of the A's effort. Moreover, (5) implies that, if the decision maker does not obtain any information about s or e, then the expected cost of choosing ι^G is lower than ι^B.

Finally, D's objective is to minimize the costs associated with policy ι:

$$u^D = -C. \tag{6}$$

The Public. The public (P) does not observe the signal s nor the message $m(s)$. It only observes whether the cost paid by D according to (6). It can therefore infer the realization of y if $\iota = \iota^G$, but can only infer that $\iota = \iota^B$ if D's payoff is $-C_1$. Based on these observations, the public updates its beliefs about A's type, following Bayes' rule whenever possible. The public's payoff is assumed to be identical to that of the decision maker:

$$u^P = -C. \tag{7}$$

In the interpretation of A and D as institutions in charge of banking stability, the public only observes whether a banking crisis – the bad outcome – happens if the lender of last resort does not intervene to prevent it. If the lender of last resort intervenes, then the public does not know what would have happened without intervention – just isolated bank failures or a major banking crisis.

Timing. Here we summarize the timing of the move and observability of actions:

1. Nature decides A's privately observed type $\theta \in \{L, H\}$, where the probability of high type is equal to $\Pr(\theta = H) = \mu$.
2. A puts effort $e \in [0, 1]$ by paying the cost $c_\theta(e)$. The effort e is A's private information.
3. Given e, the final outcome $y \in \{G, B\}$ and signal $s \in \{G, B\}$ are drawn from $\Pr(y, s|e)$; nobody observes y or s at this point.
4. A observes s,
   (a) in the separated institutional structure, A sends message $m(s)$ to D,
   (b) in the combined institutional structure, D also observes s.[5]
5. D decides $\iota \in I$.
6. Cost C is observed by P and payoffs are realized.

[5] Here we do not allow A to send message m. If A had an incentive to inform D of his effort, then D could use this new piece of information in addition to the signal s; however, we can show that A would never inform D of his effort, if it changed the distribution of D's actions. See Part 4 of Appendix A.3.2 for the details.

Theoretically, the model has both adverse selection (type) and moral hazard (effort). One may wonder why we need both, not only moral hazard. In fact, in the static model, moral hazard alone will lead us to the qualitatively similar trade-off between separated and centralized institutions;[6] however, in the dynamic model, the public's learning about the agent's type plays an important role in the trade-off. Moreover, it is natural given the motivation of the model to consider the different types of agents – either as regulators with different costs of implementing supervisory policy or as managers with different capabilities.

Definition of Institutional Separation versus Centralization Our focus is on one key difference between institutional centralization and institutional separation, namely the transmission of information. In the centralized institutional structure, both the agent A and the decision maker D observe signal s. In the separated institutional structure, only the agent observes signal s. All the other elements of the model stay the same between the two institutional designs. In particular, we assume that effort is not observable in the centralized institution. One may wonder whether the agent's effort may be easily observable to the decision maker; however, in many centralized institutional structures the decision maker and the agent are still sufficiently independent in terms of their tasks, making it hard for effort to be observed. For example, in the case of banking supervision inside the Bank of England, the Prudential Regulatory Authority (PRA) – in charge of banking supervision – and the Monetary Policy Committee (MPC) – in charge of monetary policy – are run by boards with very little overlap.[7] Hence, the MPC does not directly observe the daily regulatory effort of the PRA.[8]

We also assume that the objectives of the agent and decision maker stay the same regardless of the institutional design. In other words, even if the agent is in the same institution

[6] The details are available from the authors upon request.

[7] They share the Governor and Deputy Governor for Financial Stability as their board members to enhance cooperation. The details of the institution design is explained in the Quarterly Bulletin of the Bank of England 2013 Q1, available at
http://www.bankofengland.co.uk/publications/Documents/quarterlybulletin/2013/qb130102.pdf

[8] Moreover, it is easy to show that, even if the centralized institution allows the decision maker to observe an additional signal about the effort level, unless the signal precision is perfect, the same trade-off between ex ante incentive provision and ex post usage of information shows up.

as the decision maker, they are still different agencies pursuing their own objectives.[9]

3 The Static Ex-ante versus Ex-post Trade-off

We begin with the simplest version of the model, in which we make the following assumption about the payoff function $u^A(b)$:

$$u^A(b) = \begin{cases} 1 & \text{if } \frac{\Pr(\theta=H|C)}{\Pr(\theta=L|C)} \geq \frac{\mu}{1-\mu} \\ 0 & \text{if } \frac{\Pr(\theta=H|C)}{\Pr(\theta=L|C)} < \frac{\mu}{1-\mu} \end{cases} \tag{8}$$

The payoff function $u^A(b)$ reflects the case in which the public punishes A whenever his reputation decreases – whenever the public's belief that A is an H-type is lower than their prior, corresponding to the original distribution of types. Notice that the public outcome in this game is just the cost C to the decision marker. In a full information environment, in which effort e were observable to the public, the public outcome would include the value of e. Then, the payoff to A could be structured so as to only reward the targeted level of effort that maximizes the public's objective. The incentives problem in this model is driven by the fact that e is private information.

We first analyze the static problem in each of the two institutional setups, separated and centralized. Afterwards, we compare these two setups in terms of the effort provision and the expected costs of policy ι.

3.1 Separated Institutions

In the separate institutional structure, the asymmetric information between A and D is the value of the signal s and A's type θ, which maps into a choice of effort e_θ. The agent A alone observes the signal s and transmits message $m(s)$ to D. After receiving the message $m(s)$, the decision maker D makes the policy decision ι.

[9] Moreover, in Appendix A.1.1, we show that the same trade-off between ex ante incentive provision and ex post usage of information shows up, even if they share some of their objectives in the centralized institution, as long as their objectives are not perfectly aligned.

Definition 1 *An equilibrium with separate institutions is a set of effort levels e_θ, messages $m(s)$, and intervention policies ι such that (i) D chooses $\iota(m)$ to maximize (6), (ii) $e(\theta)$ and $m(s)$ maximize (2), and (iii) $\Pr(\theta = H|C)$ is derived by the public following Bayes' Rule whenever possible.*

We establish that an equilibrium with separate regulatory and intervention institutions exists, and we derive its main properties.

Proposition 1 *An equilibrium with separate regulatory and intervention institutions exists, and it has the following properties:*

- ***(Effort strategy)*** *The equilibrium effort choices satisfy $e_H \geq e_L$, with strict inequality if $e_H > 0$.*
- ***(Messaging strategy)*** *In the equilibrium with $e_H > 0$, then there is no meaningful message exchange: D's strategy does not depend on $m(s)$.*
- ***(Intervention Strategy)*** *D implements ι^G in the equilibrium with $e_H > 0$.*
- ***(Beliefs)*** *Equilibrium beliefs satisfy:*

$$\frac{\Pr(\theta = H|C = 0)}{\Pr(\theta = L|C = 0)} \geq \frac{\mu}{1-\mu} \geq \frac{\Pr(\theta = H|C = C_1)}{\Pr(\theta = L|C = C_1)} \geq \frac{\Pr(\theta = H|C = C_2)}{\Pr(\theta = L|C = C_2)}.$$

Proof. In the Appendix Section A.3.1. ■

Intuitively, since the H-type can always pretend to be the L-type with a lower cost, the H-type should obtain the higher value in the equilibrium. If the H-type chose lower effort and enjoyed a higher payoff, then the L-type would deviate to the H-type's effort level in order to save the cost of effort.

Proposition 1 also shows that there exist two equilibria: an uninteresting equilibrium in which $e_H = e_L = 0$,[10] and a more interesting equilibrium in which high type H takes positive

[10] In the equilibrium with $e_H = e_L = 0$, as long as both types take the same message strategy, no matter what C happens, the public belief is constant for each C, and so there is no reward or punishment from the

effort, and $e_H > e_L \geq 0$. In the equilibrium with positive effort, the agent A never informs D when the signal is bad. The intuition for this result is as follows. Suppose that D changed her action based on A's message, so probability of ι^B increased after some message. Once the public observes $C = C_1$, by Assumption 2, they infer that D believes that $y = B$ is more likely. Since $y = B$ happens more often after low effort, they infer that A's type is more likely to be L– the type that puts in less effort. Hence, they update their beliefs negatively, and so $u_A(b) = 0$. Since $u_A(b) = 0$ is the lowest utility that A can obtain, A would like to avoid policy ι^B being implemented. Therefore, A does not send any message to increase the probability of ι^B and gambles for a good outcome, regardless of the signal.

Given this messaging strategy by A, the decision maker D then chooses ι^G since (5) means that D prefers ι^G in the absence of any information about s.

In the application of the model to banking supervision, the central bank does not want to intervene – i.e., choose ι^B– unless it receives bad news from the bank supervisor, so the default policy would be that of no intervention – policy ι^G.

Note that D's policy decision is distorted due to A withholding information about the signal s. The reputational concern leads A to misreport the signal in order to avoid the reputational loss from a policy ι^B. We summarize this result in the following corollary:

Corollary 1 (Information withholding) *In the equilibrium with separated institutions, D's intervention decision is distorted compared to the full information case, in which D observed s: for positive e_H, the decision maker D chooses ι^G when $s = B$.*

Formally, the problem for agent A is to choose e_θ to maximize his expected utility:

$$q(e_\theta)\, u^A(b|C = 0) + (1 - q(e_\theta))\, u^A(b|C = C_2) - c_\theta(e_\theta).$$

outcome. Hence it is in turn optimal to take zero effort. Moreover, since the agent is indifferent between any C (and so messages), there is multiplicity in the message strategy. Depending on the message strategy, the decision maker's strategy also changes.

It is easy to show that reputation goes up after the final good outcome, and it goes down after the final bad outcome. Then, given (8), the above objective is equal to

$$q\left(e_{\theta}\right)-c_{\theta}\left(e_{\theta}\right).$$

The first order condition for effort is then

$$q'\left(e_{\theta}\right)=c'_{\theta}\left(e_{\theta}\right). \tag{9}$$

3.2 One Centralized Institution

We now consider the centralized institutional structure. In this setting, both A and D receive signal s, and there is no information transmission through messages.

Definition 2 *An equilibrium with a centralized institutional structure is a set of effort levels $e(\theta)$ and intervention policies $\iota(s)$ such that (i) D chooses $\iota(s)$ to maximize (6), (ii) $e(\theta)$ maximizes (2), and (iii) $\Pr(\theta=H|C)$ is derived by the public following Bayes' Rule whenever possible.*

We proceed to show to an equilibrium exists and to derive its properties. As in the case with separate institution, there is always an equilibrium with no effort. For the rest of the discussion, we focus on the more interesting equilibrium in which at least the H-type exerts a positive effort.

Proposition 2 *An equilibrium with institutional centralization exists, and it has the following properties:*

- ***(Effort strategy)*** *The equilibrium effort choices satisfy $e_H \geq e_L$, with strict inequality if $e_H > 0$.*
- ***(Intervention Strategy)*** *D implements ι^B after $s=B$ and ι^G after $s=G$.*

- ***(Beliefs)*** *Equilibrium beliefs satisfy:*

$$\frac{\Pr(\theta = H|C = 0)}{\Pr(\theta = L|C = 0)} \geq \frac{\mu}{1-\mu} \geq \frac{\Pr(\theta = H|C = C_1)}{\Pr(\theta = L|C = C_1)} \geq \frac{\Pr(\theta = H|C = C_2)}{\Pr(\theta = L|C = C_2)}.$$

Proof. In the Appendix Section A.3.2. ■

Proposition 2 shows that in the equilibrium with $e_H > 0$, the H-type chooses strictly higher effort than the L-type. The intuition is the same as in the case with separated institutions: the H-type faces a lower marginal cost of effort, so he faces a relatively higher benefit of effort. D now observes the signal directly, and she uses that information to choose the cost-minimizing policy ι.

Although there is no information withholding, compared to the case of separated institutions, another type of distortion emerges, due to A's inability to influence D's policy decision once s is observed.

Corollary 2 (Effort withholding) *Suppose also that we focus on the equilibrium with $e_H > 0$ in both institutions. In the equilibrium with institutional centralization, A's effort choice is lower than in the case with institutional separation.*

The result emerges if we consider A's problem in the centralized institutional structure:

$$\max_{e_\theta} q(e_\theta)(1-\varepsilon) u^A (b|C = 0) + (1 - q(e_\theta)) \varepsilon u^A (b|C = C_2)$$
$$+ [q(e_\theta)\varepsilon + (1 - q(e_\theta))(1-\varepsilon)] u^A (b|C = C_1) - c_\theta(e_\theta).$$

Again, it is easy to show that

$$\frac{\Pr(\theta = H|C = 0)}{\Pr(\theta = L|C = 0)} > \frac{\mu}{1-\mu},$$

and

$$\frac{\Pr(\theta = H|C = C_1)}{\Pr(\theta = L|C = C_1)}, \frac{\Pr(\theta = H|C = C_2)}{\Pr(\theta = L|C = C_2)} < \frac{\mu}{1-\mu}.$$

Hence, given (8), the above objective becomes

$$q\left(e_\theta\right)(1-\varepsilon)\,q\left(e_\theta\right)-c_\theta\left(e_\theta\right).$$

The first order condition for effort is then

$$(1-\varepsilon)\,q'\left(e_\theta\right)=c'_\theta\left(e_\theta\right). \tag{10}$$

Note that the effort in the separate institution characterized in (9) is as if ε were zero with the centralized institution. Let e^ε_θ be the solution to (10). Applying the implicit function theorem to the first order condition, we obtain

$$\frac{de^\varepsilon_\theta}{d\varepsilon}=-\frac{q'\left(e^\varepsilon_\theta\right)}{c''_\theta\left(e^\varepsilon_\theta\right)-(1-\varepsilon)\,q''\left(e^\varepsilon_\theta\right)}<0,$$

which implies that e^ε_θ decreases in the centralized institutional structure compared to the separated institutions.

The centralized institutional structure removes the ex-ante incentive to supply effort, because A no longer has the ability to withhold information from D. Since D intervenes after $s=B$, this reduces the probability that the good outcome will be reached, reducing the expected reputational payoff to A from supplying effort.

3.3 Discussion of the Static Trade-off

Corollaries 1 and 2 illustrate the main trade-off embedded in the choice of institutional structure: with institutional separation, A has the ex-ante incentive to put in more effort, but D's policy decision is costlier due to information withholding; in the centralized structure, the policy decision is optimal, but the incentive for higher effort is removed.

Let e^s_θ denote A's effort choice in the separated structure, and let e^ε_θ denote A's effort choice in the centralized structure. The expected increase in costs when moving from the separated structure to the centralized structure are given by

$$\begin{aligned} \Delta W\left(\varepsilon, C_2, C_1\right) \equiv\; & -\sum_{\theta} \Pr(\theta)\left(1-q\left(e_{\theta}^{s}\right)\right) C_2 \\ & +\sum_{\theta}\left\{\Pr(\theta)\left(\left(1-\varepsilon\right)\left(1-q\left(e_{\theta}^{\varepsilon}\right)\right)+\varepsilon q\left(e_{\theta}^{\varepsilon}\right)\right) C_1\right. \\ & \left.+\Pr(\theta)\varepsilon\left(1-q\left(e_{\theta}^{\varepsilon}\right)\right) C_2\right\}. \end{aligned} \tag{11}$$

The first term of the expression subtracts the expected cost of the policy under institutional separation. The second term of the sum add the expected cost of the policy under institutional centralization, when the signal is $s = B$. Finally, the third term add the expected cost of the policy under institutional centralization when the signal is $s = G$, but the outcome is $y = B$.

The analysis of ΔW allows us to derive two key comparative statics. Consider an increase in the relative cost C_1/C_2. From the above expression, it follows that such an increase has the effect of increasing ΔW. Centralization increases D's use of policy ι^B as opposed to ι^G, and this change in policy decreases costs if C_1/C_2 increases. By contrast, an increase in the signal noise ε has the effect of increasing ΔW. The centralized institution allows D to use the information provided by the signal s. The benefit of this information is higher when it is more accurate – when ε is small – since (i) we do not need to pay the unnecessary cost C_1 if $y = G$ and (ii) the distortion of effort is smaller, as seen in (10).

Proposition 3 *The expected cost of policy ι is larger with institutional centralization compared to institutional separation if ε and $\frac{C_1}{C_2}$ are sufficiently high.*

The above result also highlights the fundamental reason why centralization does not always dominate separation: the inability of D to commit not to choose ι^B after $s = B$. With commitment, centralization would always achieve an outcome at least as good as separation. The choice of institutional structure therefore trades off commitment (of not observing s first of all in the separated institution) – which increases effort – and flexibility – which allows D to use the information from the signal when there are likely to be savings of the cost from it.

Going back to the motivating example of the structure of banking supervision, the static trade-off implies that separation of bank supervision from the lender of last resort should provide incentives for higher supervisory effort. This effort should translate into banks taking less risk. Some suggestive evidence to this end is provided in Eichengreen and Dincer (2011), who look at on bank performance and supervisory structure in 140 from 1998 to 2010. They find that countries with separated institutions have fewer nonperforming loans as a share of GDP, controlling for inflation, per capita income, and country and/or year fixed effects. The other side of the trade-off, however, is that the final bad outcome is reached more often with institutional separation. The link between institutional structure and banking crises has been examined by Goodhart and Schoenmaker (1995), using data from 24 countries over the 1980s. Their data shows that bank failures happened with more than double the frequency in countries with separated rather than centralized institutions and these failures required a more frequent use of public funds for bank rescues (as opposed to commercial bank funds in rescues coordinated through the central bank).

4 Dynamic Institutional Choice

After establishing the static trade-off between effort and information transmission, a natural next step is to ask what it implies for institutional design, a process that is inherently dynamic, since belief of the agent's type keeps being updated and different belief leads to different choice of institution and replacement strategy by the public. To answer this question, we extend the static model to a dynamic one. We show that the static trade-off survives in a dynamic framework, but reputational dynamics add a novel element to the trade-off between institutional structures: the centralized institution uses the signal information efficiently for policy decisions, but this slows down learning about the agent's type. When policy ι^B is used more frequently, outcomes become less informative about the agent's effort. This effect is strongest when the reputation of the agent is intermediate, so exactly when there is sufficient prior information to update beliefs strongly, but the current observed outcome is not informative enough to make use of this prior reputational buildup.

We show these dynamics with two sets of results. First, we consider the best sustainable equilibrium (as introduced in Chari and Kehoe (1990)) and show that it corresponds to the best Perfect Bayesian Equilibrium (PBE) for the public without communication with the agent. We argue this equilibrium concept captures the key dynamic interactions that motivate our model. Second, we also consider the best PBE for the principal with communication between the principal and the agent. This problem corresponds to the dynamic mechanism design problem. We show that the dynamic trade-off described above is also present if we allow communication.

4.1 Dynamic setup

The setup of the dynamic game is as follows. In each period t, the incumbent agent is endowed with the prior belief μ_t given the public history. At the beginning of each period, the public P decides whether to keep or to replace him. Upon replacement, a new agent is selected from the pool of possible agents, and the belief about the agent's type reverts to the initial prior given the distribution of types in the pool – the prior μ_H that the agent is an H-type.

Notice that the payoff from reputation $u^A(b)$ described in (8) has an immediate translation as the agent being kept for the next period or replaced. Therefore, the agent's payoff each period is given by

$$u^A(b) = \begin{cases} 1 & \text{if } A \text{ is kept} \\ 0 & \text{if } A \text{ is replaced} \end{cases}.$$

The per-period payoff for D and P is still equal to $u^D = u^P = -C$. The discount factor is common and equal to $\delta \in [0, 1)$.

In the motivating example about regulatory institutions, replacement happens through the government – the leadership of a regulatory agency is removed and a new leadership is appointed.

Then, the public decides the institutional structure, $\psi \in \{0, 1\}$, where $\psi = 0$ denotes a separated structure and $\psi = 1$ a centralized one. For simplicity, we assume that there is no switching cost to changing the institutional structure.[11] After the institutional structure is chosen, the intra-period play is the same as described in section 2, steps 2–6, except that there is no communication in step 4.[12]

Finally, we assume that at the end of the period, the value of the signal s_t becomes public. So, the public history at the beginning of period $t+1$ is given by $\{C_j, s_j\}_{j=0}^{t}$. Intuitively, after a policy is implemented and an outcome is observed, audits take place and the information available to the decision maker at the time of the decision is revealed to the public. Since the audits only reveal tangible information, we continue to assume that effort e is not revealed, as in the standard moral hazard models.

Throughout this section, we make the following additional assumption about the cost of effort:

Assumption 3 *The cost function $c_L'(e)$ satisfies $c_L'(0) \to \infty$.*

Under Assumption 3, the marginal cost of effort of the L-type agent is sufficiently high such that he chooses not work. If the L-type puts positive effort, then only difference is that we additionally keep track of L's incentives. See Appendix A.2 for the details.

4.2 Best Sustainable Equilibrium

We first consider the case in which there is no communication between P and A about A's type.[13] As we show below, the best PBE without communication is equivalent in terms of on-path outcome to the best equilibrium for P in the class of sustainable equilibrium introduced by Chari and Kehoe (1990). To construct an equilibrium in this class, we consider

[11] If there were a cost of institutional re-organization, then the institutional change would happen only if the benefit from the change were sufficiently large.

[12] This prohibition of communication is without loss of generality: if there is a benefit of incentivizing the agent to tell the truth about s in step 4, it would be optimal for P to pick $\psi = 1$. Moreover, if there is a benefit of declaring the type, it would be more efficient for P to incentivze A to do so before P picks ψ.

[13] This assumption is made partially because one may think that the agent declaring he is low-type is not realistic, but mainly because it captures the main dynamic trade-off in a transparent way.

a public randomization device – or observable mixed strategy– for P. This randomization device is used to determine the replacement decision at the beginning of each period and the subsequent decisions within the period. That is, P draws a random variable $z_t \sim$ Uniform $[0,1]$, which is observed by everybody. The probabiity of replacement given z_t is then p_{z_t}.

Definition 3 *Given the history $\{h_t, \mu_t\}$ of public outcomes up to period t and the public's beliefs about the agent's type, a sustainable equilibrium consists of continuation strategies $\{\psi_t(h_t, \mu_t, z_t), p_{z_t}\}_t^\infty$ for the public, $\{e_t(h_t, \mu_t, \psi_t, z_t)\}_t^\infty$ for the agent, and*

$\left\{\iota_t(h_t, \mu_t, z_t, \psi_t, s_{t|\psi_t=1})\right\}_t^\infty$ for the decision maker such that each of the players maximizes their expected continuation utilities.[14]

Notice that for the dynamic analysis we can ignore the messaging in the separated institution. This is due to the intra-period results of Proposition 1 – that D's optimal policy decision is independent of message – and s_t being publicly revealed at the end of period t.

We now proceed to analyze the problem recursively, focusing on the best sustainable equilibrium for P. For this, we study the problem of maximizing $P's$ welfare given a promised payoff V to the H-type agent. Hence the state variable each period is a pair consisting of V and the belief μ.[15]. Given our motivation, this structure resembles a feasible reward structure for the agent. It allows P to write a feasible contract for A that promises a certain level of total benefits, which cannot be made state continent.

Let $J(\mu, V)$ be P's welfare given the two state variables, belief μ and promised utility V. Upon replacement, the belief goes back to μ_H, but P can choose the promised utility to maximize its welfare. Hence, P's utility after A is replaced, denoted by $\bar{J}$, should satisfy

$$\bar{J} = \max_V J(\mu_H, V).$$

The problem for P is to select a vector $\alpha_z = (p_z, \psi_z, e_z, \iota_z, V'_z)$ for each possible realization of z, where p_z is the probability of replacement, ψ_z is the institutional structure that period,

[14] Where $s_{t|\psi_t=1}$ denotes the fact the policy decision depends in s_t only when $\psi_t = 1$.

[15] Since the L-type does not work, we do not keep track of his utility.

e_z is the proposed effort for the H-type,[16] ι_z is D's policy choice, and V'_z determines the continuation promised utility. Here, the public outcome, denoted o_z, consists of ψ_z, s, ι_z, and C.[17] Let $V'_z(o_z)$ denote the next-period promised utility for A of type H after the realization of the public outcome o_z. Since the on-path outcome has full support, the principal always believes that the recommended effort e_z is taken. Hence the next belief μ' is determined by the Bayes rule, which depends on the prior μ, the recommended effort e_z, and the public outcome o_z. Let $\mu'(\mu, e_z, o_z)$ denote this updated belief.

In particular, P chooses α_z to solve the following dynamic program:

$$\begin{aligned} J(\mu, V) &= \max_{\alpha_z} \int_z \Big[p_z \bar{J} + (1-p_z) \Big\{ (1-\delta) u^P(\mu, \psi_z, e_z, \iota_z) \\ &\quad + \delta \sum_{o_z} \Pr(o_z | \mu, \psi_z, e_z, \iota_z) J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V'_z(o_z)) \Big\} \Big] dz, \end{aligned} \tag{12}$$

where

$$u^P(\mu, \psi_z, e_z, \iota_z) = -\mu \mathbb{E}[C | \theta = H, \psi_z, e_z, \iota_z] - (1-\mu) \mathbb{E}[C | \theta = B, \psi_z, e_z, \iota_z]$$

is the principal's instantaneous welfare.

The principal is subject to the following constraints:

$$V = \int_z (1-p_z) \left\{ (1-\delta)[1 - c_H(e_z)] + \delta \sum_{o_z} \Pr(o_z | \psi_z, e_z, \iota_z) V'_z(o_z) \right\} dz; \tag{13}$$

$$e_z \in \arg\max \left\{ (1-\delta)[1 - c_H(e_z)] + \delta \sum_{o_z} \Pr(o_z | \psi_z, e_z, \iota_z) V'_z(o_z) \right\}. \tag{14}$$

Constraint (13) is the promise keeping that P is bound to in equilibrium, and (14) is the incentive compatibility constraint for A.[18]

[16] The L-type L always chooses $\sigma = 0$.

[17] Note that we assume s is observable because of the audit.

[18] Notice that, from P's view point, the distribution of the outcome $\Pr(o_z | \mu, e_z)$ depends on μ, while for the promise keeping and incentive compatibility, we have $\Pr(o | e_z)$ since A knows his own type.

Finally, notice that the dynamic objectives of P and D are identical. This happens because they have the same instantaneous payoff, and the signal information is revealed to the public after each period's play. Hence we do not have any incentive compatibility constraint for ι_z. From now on, we assume that the principal decides the policy ι_z directly.

Note that there is always a PBE in which A exerts no effort, expecting replacement every period; given this, P pick the centralized institution and D chooses ι^B every period. Let $\underline{SW}$ be P's utility under this "no effort equilibrium," which does not depend on beliefs.[19] The lowest value P could obtain would be $J(\mu, 1)$ since the promised value 1 allows A to stay without putting in effort. But this value corresponds to the "no effort equilibrium." We show next that, by switching to this worst PBE after any deviation from policy α_z given z, we can sustain a PBE that maximizes (12) subject to the above constraints.

Definition 4 *The best sustainable equilibrium for P is characterized by the mapping from $(\mu, V) \in [0,1] \times [0,1]$ to a vector $(p_z, \psi_z, e_z, \iota_z, V_z')_{z\in[0,1]}$ which maximizes (12) subject to (13) and (14).*

We can restrict the domain of the states (μ, V) to $[0,1] \times [0,1]$. The belief μ has to be in $[0,1]$. The promised utility has to be in $[0,1]$ since (i) 0 is the minimum payoff that the agent receives; and (ii) 1 is the highest feasible utility for the agent, which is implemented by keeping the agent and allowing him put in effort $e = 0$ forever.

4.2.1 Equilibrium Properties

We first show that $P's$ value function $J(\mu, V)$ is concave is V, convex in μ, and increasing in μ:

Lemma 1 *$J(\mu, V)$ is concave is V, convex in μ, and increasing in μ (strictly increasing if $V \in (0,1)$).*

[19] In the equilibria with effort, P's utility, denoted by $u^P(\mu, \iota_z, \sigma_z)$, depends on the prior and current strategy, which is weakly increasing in μ since the low type does not work for sure.

Proof. See Appendix Section A.3.3. ■

Since we allow the public randomization, the concavity of $J(\mu, V)$ with respect to V follows from the standard arguments. To see why $J(\mu, V)$ is convex in μ, assume that P receives additional information to update its belief about A's type. The belief is martingale, so the updated belief is a mean-preserving spread of the original belief μ. Since P can always ignore this new information, this mean-preserving spread is always (at least weakly) welfare-improving. Finally, $J(\mu, V)$ is increasing in μ since P could always choose the same continuation strategy as μ when the belief is $\mu' > \mu$. Since the L-type chooses $e_L = 0$, this allows P to obtain at least weakly higher welfare with μ' compared to μ.

We can then derive the following implications about the shape of $J(\mu, V)$.

Lemma 2 *The value function $J(\mu, V)$ satisfies the following:*

1. *$J(\mu, 0) = \bar{J}$ for each μ;*

2. *For each μ, there exists $V(\mu)$ such that $J(\mu, V)$ is linear for $V \in [0, V(\mu)]$, where $V(\mu) \geq 1 - \delta$ with strict inequality for $\mu \geq \mu_H$. Moreover, the slope for the linear part is negative for $\mu < \mu_H$; zero for $\mu = \mu_H$; and positive for $\mu > \mu_H$.*

3. *There exists $V^*(\mu) \in \arg\max_V J(\mu, V)$. For each μ, at $V = V^*(\mu)$, the following property holds:*

   (a) $V_z'(o_z) \leq \arg\max_V J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V)$ for each event o_z with negative belief update $\mu'(\mu, \psi_z, e_z, \iota_z, o_z) \leq \mu$;

   (b) $V_z'(o_z) \geq \arg\max_V J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V)$ for each event o_z with positive belief update $\mu'(\mu, \psi_z, e_z, \iota_z, o_z) \geq \mu$.

Proof. See Appendix Section A.3.5. ■

The above properties have the following intuition. First, if the value promised to A is 0, only immediate replacement can fulfill this promise. Hence, we have $J(\mu, 0) = \bar{J}$ for each μ.

Second, for a sufficiently small promised value V, the replacement must happen with positive probability. Since this positive probability is generated by randomization between keeping and replacing A, the principal P's indifference means that $J(\mu, V)$ must be linear for V below some promised utility $V(\mu)$.

The third property follows from concavity: the function $J(\mu, V)$ has a maximum, denoted $V^*(\mu)$. At $V = V^*(\mu)$, assume outcome o_z is observed such that the belief is updated negatively – $\mu' < \mu$. The negative belief update implies that o_z is the outcome which happens less often with higher effort. Hence reducing $V_z'(o_z)$ incentivizes A to work hard. This increased effort has two benefits: the instantaneous welfare u^P is improved; and since the higher effort allows for a higher belief update (the H-type is more distinguishable from the L-type), by the convexity of J with respect to μ, the principal P's continuation welfare improves. As long as

$$V_z'(o_z) > \arg\max_V J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V),$$

reducing $V_z'(o_z)$ directly improves P's continuation welfare as well. Once

$$V_z'(o_z) < \arg\max_V J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V),$$

however, decreasing $V_z'(o_z)$ directly reduces P's continuation welfare. Intuitively, promising too little continuation payoff of A forces P to replace the agent even if P believes that A is an $H-$type.[20] The analysis when outcome o_z occurs such that the belief is updated positively is analogous.

Finally, we establish that the sustainable equilibrium in our setup corresponds to the best perfect Bayesian equilibrium for P when there cannot be communication between P and A:

Lemma 3 *The sustainable equilibrium corresponds to the best perfect Bayesian equilibrium*

[20] Note that, by the second property, we have $\arg\max_V J(\mu'(\mu, \psi_z, e_z, \iota_z, o_z), V) = 0$ if $\mu'(\mu, \psi_z, e_z, \iota_z, o_z)$ is lower than the prior in the pool, μ_H.

for P without communication between principal and agent.

Proof. In Appendix Section A.3.4. ■

Note that there is always a PBE in which A exerts no effort, expecting replacement every period; given this, P pick the centralized institution and D chooses ι^B every period. Let $\underline{SW}$ be P's utility under this "no effort equilibrium," which does not depend on beliefs.[21] In the sustainable equilibrium, the lowest value P could obtain would be $J(\mu, 1)$ since the promised value 1 allows A to stay without putting in effort. But this value corresponds to the "no effort equilibrium." Then, switching to this worst PBE after any deviation from policy α_z given z, we can sustain a PBE with the same outcome as the sustainable equilibrium.

4.2.2 The Optimal Dynamic Policy

We now solve for the optimal dynamic policy. Given Lemma 2, we can show that, without observing s, it is optimal for D to take ι^G. In addition, without loss, we can assume that the centralized institution will utilize the information about the signal s.

Proposition 4 *The optimal strategy ι_z satisfies the following:*

- *If $\psi_z = 0$ (separated institutional structure), then we have $\iota_z = i^G$;*
- *If $\psi_z = 1$ (centralized institutional structure), then we have $\iota_z = i^B$ after $s = B$ and $\iota_z = i^G$ after $s = G$.*

The optimal policy for P is to choose i^G whenever it does not have additional information about the signal. By (5), this is myopically optimal given A's effort. In addition, this policy choice increases A's effort, which increases P's instantaneous utility. Also, policy ι^G allows for a larger belief update, since the outcome is a better indication of effort. Due to the convexity of $J(\mu, V)$, the larger belief update is also improving P's dynamic payoff. The second result is obtained because, when D can observe the signal, it is optimal to use that

[21] In the equilibria with effort, P's utility, denoted by $u^P(\mu, \iota_z, \sigma_z)$, depends on the prior and current strategy, which is weakly increasing in μ since the low type does not work for sure.

information and to choose i^B after $s = B$ and i^G after $s = G$. Otherwise, choosing $\psi = 0$ would be one of the optimal strategies. In other words, we can assume that the centralized institution is picked only if the gains from reducing the expected cost of the outcome outweigh the disincentive for A to put in effort and the lower update in beliefs.

For the following set of results in the analysis, we make a simplifying assumption regarding the error in the signal s_t.

Assumption 4 $\Pr(s = B|y = G, e) = \varepsilon > 0$ *and* $\Pr(s = G|y = B, e) = 0$.

Under this assumption, we consider the case in which the error happens in the signal only when the final outcome is good. All the results so far hold without this assumption, so all the qualitative trade-offs between the two institutional designs are robust to $\Pr(s = G|y = B, e) > 0$.

Lemma 4 *Under Assumption 4, the public outcome is boiled down to one of three cases:* $o_z \in \{0, C_1, C_2\}$*, where* $o_z = C_1$ *corresponds to* $\psi = 1$ *and* ι^B *and the other two outcomes correspond to* $\psi = 0$ *and* ι^G.

Proof. In the Appendix Section A.3.6. ■

Given this simplification of the public outcomes, we can pin down the continuation utility for A after each event.

Lemma 5 *The continuation utility for* A *is higher after outcome* $o_z = 0$ *than after outcome* $o_z \in \{C_1, C_2\}$*. Moreover, the difference between* A*'s expected continuation payoff after* $o_z = 0$ *and the continuation payoff after* $o_z \in \{C_1, C_2\}$ *increases as a function of effort* e_H.

Proof. In the Appendix Section A.3.7. ■

The intuition for this result is that A is rewarded with a higher promised payoff after the outcomes that indicate higher effort. Conversely, A is punished with lower promised effort after the outcome that indicates low effort.

To derive this result, consider first the separated institutional structure. From (14), A's choice of effort satisfies

$$c'_H(e_z) = \frac{\delta}{1-\delta} q'(e_z) \left[V'_z(0|\psi=0) - V'_z(C_2|\psi=0)\right]. \tag{15}$$

It then follows that

$$V'_z(0|\psi=0) - V'_z(C_2|\psi=0) = \frac{1-\delta}{\delta} \frac{c'_H(e_z)}{q'(e_z)} > 0, \tag{16}$$

so there is a positive gap between the continuation utility for A after outcome $o_z = 0$ and outcome $o_z = C_2$. Moreover, this gap is increasing in the targeted effort e_z.

To show that $V'_z(0|\psi=0)$ itself is increasing in e_z, notice that Proposition 4 implies that (13) becomes:

$$V_z = (1-\delta)\left[1 - c_H(e_z)\right] + \delta q(e_z)\left[V'_z(0|\psi=0) - V'_z(C_2|\psi=0)\right] + \delta V'_z(C_2|\psi=0). \tag{17}$$

Then, given (16), $V'_z(C_2|\psi=0)$ is determined given e_z and V_z,

$$\delta V'_z(C_2|\psi=0) = V_z - (1-\delta)\left[1 - c_H(e_z) + q(e_z)\frac{c'_H(e_z)}{q'(e_z)}\right].$$

It then follows from this expression that $V'_z(C_2|\psi=0)$ is a decreasing function of e_z. Given (16), this implies that $V'_z(0|\psi=0)$ is increasing in e_z.

Under the centralized institution, the same steps as above yield

$$V'_z(0|\psi=1) - V'_z(C_1|\psi=1) = \frac{1-\delta}{\delta} \frac{c'_H(e_z)}{(1-\varepsilon)q'(e_z)}, \tag{18}$$

showing that the there is a positive difference in continuation values for A. Similarly, Proposition 4 implies that $V'_z(C_1|\psi=1)$ can be derived from (13) as

$$\delta V'_z(C_1|\psi=1) = V_z - (1-\delta)\left[1 - c_H(e_z) + q(e_z)\frac{c'_H(e_z)}{q'(e_z)}\right], \tag{19}$$

showing that it is a decreasing function of effort e_z. Finally, (18) then implies that $V_z'(0|\psi = 1)$ increases in e_z.

Comparing (16) and (18), we obtain the following Lemma.

Lemma 6 *Fixing any targeted effort level e_z, the continuation utility for A satisfies*

$$V_z'(0|\psi = 1) - V_z'(C_1|\psi = 1) > V_z'(0|\psi = 0) - V_z'(C_2|\psi = 0).$$

The result shows that P must incentivize A with a higher relative reward under the centralized institution in order to obtain the same level of effort as with the separated institutions. The intuition for this result is the following. Under the centralized institution, the observed outcome is less informative about A's effort, because D can act on the signal information. Since effort counts less for A's reputation, A is less willing to work hard unless he is promised a higher reward.

The results so far show how the static trade-off translates to the dynamic framework. For any effort level e, P's instantaneous utility is higher in the centralized institutional structure, since D's use of the signal information lowers the expected cost of outcome. But to attain the same level of effort in the centralized institution as in the separated one, P has to promise a higher reward to A after the good outcome. But promising a higher continuation utility to A reduces P's future payoff, since P essentially accepts less effort in the future. To see it formally, note that 3b) of Lemma 2 shows that, if we start from the optimal promised value given the current belief, then increasing the continuation payoff for A after the good outcome reduces the continuation welfare for P. This corresponds to static trade-off of ex ante high-powered incentives versus ex post efficient use of the information.

The following result highlights the new trade-off that emerges when dynamics are considered.

Proposition 5 *Suppose Assumptions 1, 2, and 4 hold. Fixing any targeted effort level e_z, the following dynamic trade-off emerges: under institutional separation ($\psi = 0$) learning* *about A's type happens at a faster pace – the update in beliefs each period is larger –, but the instantaneous utility is lower compared to institutional centralization ($\psi = 1$).*

Consider the update in beliefs after the worst outcome in each institutional structure given the equilibrium play, i.e. $o_z = C_1$ in the centralized institution and $o_z = C_2$ in the separated institution. It then follows that

$$\underbrace{\left|\mu - \frac{\mu\left[1-(1-\varepsilon)\,q\left(e_z\right)\right]}{1-(1-\varepsilon)\left[\mu q\left(e_z\right)+(1-\mu)\,q\left(0\right)\right]}\right|}_{\substack{\mu'(C_1|\psi=1):\ \text{belief update} \\ \text{with the centralized institution}}} < \underbrace{\left|\mu - \frac{\mu\left[1-q\left(e_z\right)\right]}{1-\left[\mu q\left(e_z\right)+(1-\mu)\,q\left(0\right)\right]}\right|}_{\substack{\mu'(C_2|\psi=0):\ \text{belief update} \\ \text{with separated institutions}}}, \tag{20}$$

so the beliefs are updated more strongly under the separated institutional structure. The reason is that the final bad outcome $y = B$ is more informative of A's effort than the signal $s = B$. This difference in the speed of learning translates into P's utility. Since P's belief is a martingale, this means that the updated belief distribution with $\psi = 0$ – the separate institutions – is a mean-preserving spread of the distribution with $\psi = 1$ – the centralized institution. By Lemma 1, since $V(\mu, V)$ is convex in μ, this mean-preserving spread increases P's utility. The faster learning under the separated institutions increases the payoff from this institutional choice.

Finally, we can show that the advantage of the separated institutions over the speed of learning is highest for intermediate values of the belief μ.

Corollary 3 *The difference in learning pace between institutional structures is largest when beliefs about A being an H-type are intermediate and smallest when beliefs are either* 0 *or* 1.

Proof. In the Appendix Section A.3.8. ■

The result follows from examining the dependence of the updated belief μ' on the prior μ. The intuition is that the advantage of leaning fast is largest when there is both enough uncertainly about A's type – a high value of learning – and enough prior information about A – enough information so that any new observation can be valued.

Proposition 5 has a direct implication for the possibility of institutional change in the equilibrium path.

Corollary 4 *Cycles between institutional centralization and separation can emerge on the equilibrium path.*

This conclusion follows because the advantage of institutional separation is highest when the dynamic trade-off is strongest in favor fast learning – when the beliefs about A are intermediate. When P does not have much information about A, the gains from choosing institutional separation are lowest. The gains increase as A has a higher reputation, and then decrease again as P has learned enough about A's type so that the marginal contribution of new observations to learning is low. This dynamic leads to cycles between centralization when A's reputation is low, separation when A's reputation is intermediate, and then centralization again when A's reputation is sufficiently high.

In our motivation for this model, this dynamic can explain the fluctuations we see in the data between centralization and separation, as well as the different institutional choices made by different countries during the same time period.

4.3 Best PBE with Communication

The analysis of the dynamic model so far has focused on the resulting equilibria of the game in which the public does not have any communication channels with the regulator. A natural next step is to ask what is the best PBE if we allow the most general communication channel between P and A. The answer to this question is interesting for two main reasons. First, we are providing the natural complement to the previous section in which we characterized the best PBE without communication. Second, one may wonder what is the optimal mechanism in the problem we are presenting in this paper. We show that the best PBE with communication corresponds to the solution for the mechanism design problem with both adverse selection and moral hazard, and without transferable utility.

We begin by characterizing the optimal dynamic mechanism, and then we prove that the solution corresponds to the on-path outcome of the best PBE for the principal. We extend the dynamic setup to allow the agent to send a cheap talk message from an arbitrary message space M_t to the principal P. Intuitively, if P can incentivize the agent to declare his type, then the principal may want to change the institutional design according to this revelation.[22]

To characterize the optimal mechanism, by the revelation principle, we can focus on the following type of "contracts:" The principal prepares two contracts $\mathcal{C}_H$ and $\mathcal{C}_L$, where each contract specifies the replacement probability after each history, and it is designed so that type θ self-selects contract $\mathcal{C}_\theta$ upon arrival, using the communication: $\mathcal{C}_H \succcurlyeq_H \mathcal{C}_L$ and $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$. Here $\succcurlyeq_\theta$ is type θ's preference. As in the analysis of the best PBE, we focus on the case in which the L-type does not work: $c_L'(0) = \infty$.[23]

The next lemma shows that $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$ is binding.

Lemma 7 *The constraint $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$ is binding.*

Proof. See Appendix A.3.9. ■

Intuitively, if the constraint were not binding, the principal would like to replace the agent as soon as he declares that he is a low type.

Given this lemma, we start with the relaxed problem in which the planner maximizes the ex ante social welfare subject to the constraint that L is indifferent between the two contracts, and then verify that $\mathcal{C}_H \succcurlyeq_H \mathcal{C}_L$ is also satisfied.

4.3.1 Optimal Contract for the L-type

Consider a contract for the high type $\mathcal{C}_H$. Suppose that the low type declares H and enters the contract $\mathcal{C}_H$. Since $c_L'(0) = \infty$, he does not work. Let p_t be the probability that the

[22] In the static setup, the agent does not have incentive to declare he is an $L-$type, since this would cause replacement. In the dynamic setup, however, he can "punish" P by deviating from the promised continuation play. Therefore, P may be able to incentivize A to tell the truth about his type.

[23] In Appendix A.2, we explain that the same method to solve for the optimal mechanism works even if L-type works.

contract $\mathcal{C}_H$ terminates when no effort is exerted. The value that type L obtains from $\mathcal{C}_H$ is

$$V_{LH}\left(\mathcal{C}_H\right)=\sum_{t=1}^{\infty}p_t\left(1-\delta^t\right),$$

where p_t is the probability of replacement.

Given $\mathcal{C}_H$, the optimal contract for the L-type, denoted by $\mathcal{C}_L^*\left(\mathcal{C}_H\right)$, maximizes P's utility given $\theta=L$ subject to the constraint that the promised utility is no less than $V_{LH}\left(\mathcal{C}_H\right)$. To solve for $\mathcal{C}_L^*\left(\mathcal{C}_H\right)$ at each $\mathcal{C}_H$, consider the principal's problem given $\theta=L$:

$$\begin{aligned} J_L\left(V\right) &= \max_{\left(p_z,\iota_z,V_z'\right)_z}\int_z[p_z\bar{J}+\left(1-p_z\right)\{-\left(1-\delta\right)\mathbb{E}\left[C|\psi_z,e_z=0,\iota_z\right] \\ & +\delta\sum_{o_z}\Pr\left(o_z|\psi_z,e_z=0,\iota_z\right)J_L\left(V_z'\left(o_z\right)\right)\}]dz. \end{aligned}$$

Here we use the same notation as in Section 4.2. Since the $L-$type agent does not work, there are no learning or incentive considerations for this agent, so it is optimal to always pick $\psi_z=1$. Hence we omit the choice of ψ_z. Moreover, given Assumption 1, it is optimal to implement policy ι^B after $s=B$. Hence the public outcome o_z consists of $\psi_z=1$ and C.[24]

The principal is subject to the promise keeping constraint:

$$V=\int_z\left(1-p_z\right)\left\{\left(1-\delta\right)+\delta\sum_{o_z}\Pr\left(o_z|\psi_z,e_z=0,\iota_z\right)V_z'\left(o_z\right)\right\}dz.$$

We can solve this problem explicitly:

Lemma 8 *The value function for P in the optimal contract with the $L-$type agent is given by*

$$J_L\left(V\right)=\bar{J}-\left(\bar{J}-\underline{SW}\right)V. \tag{21}$$

Proof. In the Appendix Section A.3.10. ■

[24] Note that, given $\psi_z=1$ and the optimal ι, the cost $C=C_1$ uniquely pins down $s=B$; $C=C_2$ means $s=G$ and $y=B$; and $C=0$ means $s=y=G$.

Lemma 8 shows that $J_L(V)$ is equivalent to a convex combination of $\bar{J}$ and $\underline{SW}$. That is, P flips a coin and, with some positive probability, it replaces A, and with some probability, it gives A the highest promised utility 1. In other words, P tries to replace A with the highest probability, by promising the highest continuation payoff conditional on not being replaced. Intuitively, since the L-type does not work and there is discounting, it is optimal to replace A as soon as possible with maximum probability.

Given the optimal contract $\mathcal{C}_L^*(\mathcal{C}_H)$, the H-type always wants to tell the truth:

Lemma 9 *For each $\mathcal{C}_H$, with $\mathcal{C}_L^*(\mathcal{C}_H)$ implicitly given in Lemma 8, we have $\mathcal{C}_H \succsim_L \mathcal{C}_L^*(\mathcal{C}_H)$.*

Proof. In the Appendix Section A.3.11. ■

The result follows from the following three facts: first, since the H-type has the lower cost of effort, he must obtain a weakly higher payoff than the L-type from choosing $\mathcal{C}_H$; second, the payoff from contract $\mathcal{C}_L^*(\mathcal{C}_H)$ is the same for all types, since it is optimal not to put any effort regardless of type; and third, we construct $\mathcal{C}_L^*(\mathcal{C}_H)$ so that the L-type is indifferent between $\mathcal{C}_H$ and $\mathcal{C}_L^*(\mathcal{C}_H)$. Hence, H-type weakly prefers $\mathcal{C}_H$ to $\mathcal{C}_L^*(\mathcal{C}_H)$.

4.3.2 Optimal Contract for the H-type

When P designs the contract $\mathcal{C}_H^*$ for the H-type, she needs to take into account the effect of $\mathcal{C}_H^*$ on the value that the L-type obtains from $\mathcal{C}_H^*$, $V_{LH}(\mathcal{C}_H^*)$, since it affects P's utility through the L-type's incentive compatibility constraint $\mathcal{C}_L^* \succsim_L \mathcal{C}_H^*$. In particular, suppose now that P is at history h^t after A declares type H. Let $\Pr(h^t|\theta)$ be the probability that the contract reaches history h^t if A is type θ; let $J_H(h^t)$ be P's continuation welfare given an agent of type H; and let $V_{LH}(h^t)$ be L-type's continuation payoff. From the ex ante point of view, when we change $J_H(h^t)$ by Δ_H, then P's ex ante welfare changes by $\mu_H \Pr(h^t|H)\Delta_H$. At the same time, when we change $V_{LH}(h^t)$ by Δ_{LH}, then the ex ante value $V_{LH}(\mathcal{C}_H^*)$ changes by $\Pr(h^t|L)\Delta_{LH}$. Since $J_L(V_{LH})$ is linear, this changes $J_L(V_{LH})$ by $-\left(\bar{J} - \underline{SW}\right)\Pr(h^t|L)\Delta_{LH}$. Hence the effect of the ex ante welfare of P is $-\left(\bar{J} - \underline{SW}\right)(1-\mu_H)\Pr(h^t|L)\Delta_{LH}$. In total, the optimal strategy for P is to maximize

$$\mu_H \Pr\left(h^t|H\right) J_H\left(h^t\right) - \left(\bar{J} - \underline{SW}\right)(1-\mu_H)\Pr\left(h^t|L\right) V_{LH}\left(h^t\right). \tag{22}$$

Note that here we use the linearity of $J_L\left(V_{LH}\right)$ to separate the decision problem in history h^t from one in history $\tilde{h}^t \neq h^t$. The marginal effect of changing $V_{LH}\left(h^t\right)$ on the ex-ante welfare is independent of the value $V_{LH}\left(\tilde{h}^t\right)$, since $J_L\left(V_{LH}\right)$ is linear. Otherwise, the marginal effect would depend on the ex-ante value V_{LH} that the L-type receives from telling a lie, which depends not only on $V_{LH}\left(h^t\right)$ but also on $V_{LH}\left(\tilde{h}^t\right)$.

By affine transformation, expression (22) becomes

$$\mu\left(h^t\right) J_H\left(h^t\right) + \left(1-\mu\left(h^t\right)\right)\left[\bar{J} - \left(\bar{J} - \underline{SW}\right)V\right], \tag{23}$$

where

$$\mu\left(h^t\right) = \frac{\mu_H \Pr\left(h^t|H\right)}{\mu_H \Pr\left(h^t|H\right) + (1-\mu_H)\Pr\left(h^t|L\right)}.$$

The expression for P's expected continuation utility given in (23) highlights the importance of the L-type's existence. When P constructs contract $\mathcal{C}_H^*$, it must account for what would happen if the L-type reached period t. The value of contract $\mathcal{C}_H^*$ for the L-type affects the incentive compatibility constraint $\mathcal{C}_L^* \succcurlyeq_L \mathcal{C}_H^*$, and through this it affects P's expected welfare. The importance of this consideration depends on how likely the L-type is to reach each history compared to the H-type. This relative importance is equivalent to P's belief in the best PBE. Hence, the problem for P becomes similar to the one analyzed in the best PBE,and we derive the following main result:

Proposition 6 *The maximization problem for P under the optimal dynamic mechanism is the same as in the best PBE without communication, except that the instantaneous welfare is replaced with*

$$-\mu(1-\delta)\mathbb{E}\left[C|\theta = H, \psi_z, \iota_z, e_z\right] + (1-\mu)\underline{SW}.$$

Proof. In the Appendix Section A.3.12. ■

The instantaneous welfare differs for the following reason. Once A declares his type as L in $\mathcal{C}_L$, then P can tailor the policy response for the L-type versus the H-type. Specifically, P's optimal response after A declares type L is to choose $\psi_z = 1$, $i_z = \iota^B$ if $s = B$ and $i_z = \iota^G$ if $s = G$. This achieves welfare $\underline{SW}$.

Since this difference in the instantaneous welfare does not affect any of the statements of Lemma 1, exactly the same trade off between the separate and centralized institutions shows up in the mechanism design solution.

Corollary 5 *The optimal mechanism can feature cycles between institutional centralization and separation.*

The analysis of the optimal mechanism reveals that the same dynamic trade-off emerges even when A is an H-type and can communicate his type to P. Even if only H-type declares H on equilibrium path, the institutions must be designed to account for the possibility that A may be an L-type who tells a lie. Some histories are easier for $L-$type to reach, while others are more difficult. This implies that both replacement and changes in institutional structure over time are part of an optimal contract.

To conclude the analysis, we now verify that the solution for the mechanism design problem corresponds to the best PBE for P with communication:

Lemma 10 *The solution for the mechanism design problem corresponds to the best Perfect Bayesian Equilibrium for P with communication.*

Proof. See Appendix Section A.3.13. ■

Intuitively, since the mechanism design allows commitment for P, it should achieve weakly higher welfare. Yet, as in Lemma 3, we can construct a punishment to support the same outcome in a PBE with communication. Therefore, the solution to the mechanism design problem achieves the same welfare as that of the best PBE with communication.

Finally, we can discuss the robustness and limitations of this result. First, the method is applicable even if the $L-$type works, as long as J_L is linear. Otherwise, the effect of changing $V_{LH}(h^t)$ after a specific history h^t on the ex ante welfare would depend on the values of $V_{LH}\left(\tilde{h}^t\right)$ after other histories $\tilde{h}^t$. The global linearity of $J_L(V_{LH})$ depends on the assumption that L-type does not work. However, we have linearity near $V_{LH} = 0$ even if the L-type works. Intuitively, if the promised value V_{LH} is very low, then replacement happens with a positive probability. We guess and verify that, as long as the cost function satisfies Assumption 1, $J_L(V_{LH})$ is locally linear in V_{LH}. This is then enough to derive the recursive formula. We provide the details in the Appendix Section A.2.

Second, if we have more than two types of agents, then we do not a priori know which constraint is binding. With two types, we know from Lemma 7 that $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$ is binding. Otherwise, the principal could lower $\mathcal{C}_L$ by immediately replacing the L-type, which is not incentive compatible. Suppose we also have third type, M, whose marginal cost of effort is in the middle of that of types H and L. For the contract designed for M, $\mathcal{C}_M$, and his preference $\succcurlyeq_M$, we do not necessarily know whether $\mathcal{C}_M \succcurlyeq_M \mathcal{C}_H$ or $\mathcal{C}_M \succcurlyeq_M \mathcal{C}_L$. Even if neither of them is binding, we do not know whether the principal wants to replace M. This decision depends on whether keeping M is welfare-enhancing as opposed to only having the two other types. Hence there is a large degree of freedom around which constraint binds, as often is the case in models with both adverse selection and moral hazard, without transferable utility.

5 Conclusion

This paper considered the institutional design problem in which costly effort provision is possible in a model of information transmission between an informed agent and an uninformed decision maker. We show that introducing effort in this environment leads to a trade-off between information manipulation and effort provision. The ability to withhold information from the decision maker encourages more effort, through the incentives of regulators to improve their reputation. Yet, information withholding has the drawback of possibly leading to costlier crises, as the decision maker is not able to make an informed decision that may improve welfare.

We argue this trade-off is relevant for the design of regulatory institutions. In the case of banking supervision institutions, the model helps explain the variation in regulatory structures across countries. It shows that the choice of institutional centralization versus separation reflects the trade-off of effort provision versus information transmission. Moreover, the model helps explain why cycles between these two institutional structures may occur over time, as in the example of the Bank of England. The model shows that such changes are optimal, as means of providing incentives for higher regulatory effort.

References

Aghion, P., Alesina, A., Trebbi, F., 2004. Endogenous political institutions*. The Quarterly journal of economics 119 (2), 565–611.

Alesina, A., Tabellini, G., 2007. Bureaucrats or politicians? part i: a single policy task. The American Economic Review 97 (1), 169–179.

Ambrus, A., Azevedo, E. M., Kamada, Y., Takagi, Y., 2013. Legislative committees as information intermediaries: a unified theory of committee selection and amendment rules. Journal of Economic Behavior & Organization 94, 103–115.

Austen-Smith, D., 1990. Information transmission in debate. American Journal of political science, 124–152.

Benabou, R., Laroque, G., 1992. Using privileged information to manipulate markets: Insiders, gurus, and credibility. The Quarterly Journal of Economics 107 (3), 921–958.

Bendor, J., Meirowitz, A., 2004. Spatial models of delegation. American Political Science Review 98 (02), 293–310.

Biais, B., Mariotti, T., Plantin, G., Rochet, J.-C., 2007. Dynamic security design: Convergence to continuous time and asset pricing implications. The Review of Economic Studies 74 (2), 345–390.

Boot, A. W., Thakor, A. V., 1993. Self-interested bank regulation. The American Economic Review, 206–212.

Chari, V. V., Kehoe, P. J., 1990. Sustainable plans. Journal of political economy, 783–802.

Crawford, V. P., Sobel, J., 1982. Strategic information transmission. Econometrica: Journal of the Econometric Society, 1431–1451.

Eichengreen, B., Dincer, N., 2011. Who should supervise? the structure of bank supervision and the performance of the financial system. NBER Working Paper (w17401).

Fong, K., 2009. Evaluating skilled experts: Optimal scoring rules for surgeons.

Goodhart, C., Schoenmaker, D., 1995. Should the functions of monetary policy and banking supervision be separated? Oxford Economic Papers, 539–560.

Kamenica, E., Gentzkow, M., 2011. Bayesian persuasion. American Economic Review 101, 2590–2615.

Mailath, G. J., Samuelson, L., 2001. Who wants a good reputation? The Review of Economic Studies 68 (2), 415–441.

Martimort, D., Semenov, A., 2008. The informational effects of competition and collusion in legislative politics. Journal of Public Economics 92 (7), 1541–1563.

Maskin, E., Tirole, J., 2004. The politician and the judge: Accountability in government. American Economic Review, 1034–1054.

Morris, S., 2001. Political correctness. Journal of Political Economy 109 (2), 231–265.

Peek, J., Rosengren, E. S., Tootell, G. M. B., 1999. Is bank supervision central to central banking? The Quarterly Journal of Economics 114 (2), 629–653.

Stokey, N. L., 1989. Recursive methods in economic dynamics. Harvard University Press.

Thomas, J., Worrall, T., 1990. Income fluctuation and asymmetric information: An example of a repeated principal-agent problem. Journal of Economic Theory 51 (2), 367–390.

A Appendix

A.1 Additional Results

A.1.1 Generalization of the Agent's Utility

One may wonder whether the static trade-off obtained in Section 3 is due to the form of the agent's utility function, which limits the benefits of reputation to only one outcome. In this section, we show that the model is in fact robust to more general utility forms, in which A's reputation changes continuously as a function of the possible outcomes.

Consider the general form for A's utility, $u^A\left(\Pr(\theta = H|C)\right),$ where $u^A(\cdot) \geq 0$ is a concave increasing function of $\Pr(\theta = H|C)$.

Separated Institutional Structure The general form of $u^A(\cdot)$ does not significantly alter the properties of the equilibrium.

Proposition 7 *An equilibrium with separate institutions exists, and it has the following main properties:*

- ***(Effort strategy)*** *The equilibrium effort choices satisfies $e_H \geq e_L$, with strict inequality if $e_H > 0$.*
- ***(Intervention Strategy)*** *D intervenes in the intermediate stage after $m = B$ and does not intervene after $m = G$ whenever $e_\theta > 0$.*
- ***(Beliefs)*** *Equilibrium beliefs satisfy:*

$$\frac{\Pr(\theta = H|C = 0)}{\Pr(\theta = L|C = 0)} \geq \frac{\Pr(\theta = H|C_1)}{\Pr(\theta = L|C_1)} \geq \frac{\Pr(\theta = H|C = C_2)}{\Pr(\theta = L|C = C_2)}.$$

Proof. In Section A.3.14. ■

As before, the H-type exerts higher effort in equilibrium. Given Assumption 2, D always wants to choose ι^B after a bad signal and to choose ι^G after either a good signal or no information. As described below, A will only withhold information from D in case of a bad signal, which makes it optimal for D to respond by intervening whenever $m = B$. Finally, as before, the public's beliefs about A's type are highest after a good outcome ($y = G$) and lowest after a major crisis ($y = B$).

With separated institutions, the general form of the payoff from reputation now allows for multiple types of equilibria:

Proposition 8 *With separate institutions, there can be three types of equilibria in terms of effort and message strategies:*

(1) **A full revelation equilibrium**: A with $\theta = L$ takes e_L^* and sends $m = B$ after $s = B$; A with $\theta = H$ takes e_H^* and sends $m = B$ after $s = B$; and A always sends $m = G$ after $s = G$.

(2) **An equilibrium with partial information withholding:** A of type L puts in effort e_L^* and sends $m = G$ after $s = G$ and message $m = B$ after $s = B$. A of type H puts in effort e_H^* and sends $m = G$ after $s = G$, message $m = G$ with probability γ and message $m = B$ with probability $(1 - \gamma)$ after $s = B$, for $\gamma \in (0, 1)$.

(3) **An equilibrium with full information withholding:** A always sends $m = G$ after every signal.

Proof. In Section A.3.15. ■

In the equilibrium with full revelation, A sends a message that is fully informative of the signal. In the other two possible equilibria, A withholds information from D in order to avoid the reputational loss due to an intervention.

Centralized Institutional Structure In the centralized institution, the equilibrium properties described in Proposition 2 are maintained.

Proposition 9 *An equilibrium with centralized institutions exists, and it has the following main properties:*

- ***(Effort strategy)*** *The equilibrium effort choices satisfied $e_H \geq e_L$, with strict inequality if $e_H > 0$.*

- ***(Intervention Strategy)*** *D intervenes after $s = B$ and does not intervene after $s = G$.*

- ***(Beliefs)*** *Equilibrium beliefs satisfy:*

$$\frac{\Pr(\theta = H|y = G)}{\Pr(\theta = L|y = G)} \geq \frac{\Pr(\theta = H|C_1)}{\Pr(\theta = L|C_1)} \geq \frac{\Pr(\theta = H|y = B)}{\Pr(\theta = L|y = B)}.$$

Proof. In Section A.3.16. ■

For simplicity of exposition, consider the case in which $\varepsilon_G = \varepsilon_B = \varepsilon$. As before, A's effort is strictly higher for the H-type whenever positive effort is undertaken. With the general form of the utility function, the first order condition (10) becomes

$$q'(e_\theta)\left\{(1-\varepsilon)\left(u(b^\varepsilon_G) - u\left(b^\varepsilon_{C_1}\right)\right) - \varepsilon\left(u(b^\varepsilon_{C_2}) - u\left(b^\varepsilon_{C_1}\right)\right)\right\} = c'_\theta(e_\theta),$$

where

$$b^\varepsilon_{outcome} = \frac{\Pr(\theta = H|outcome)}{1 - \Pr(\theta = H|outcome)}. \tag{24}$$

This condition can then be used to derive the following comparative static of effort with respect to ε.

Proposition 10 *Effort e_H is decreasing in ε if*

$$MA_\varepsilon(e_H) \equiv q'(e_H)\left\{\begin{array}{c}(1-\varepsilon)\left(u^A(b^\varepsilon_G(e_H)) - u^A\left(b^\varepsilon_{C_1}(e_H)\right)\right) \\ -\varepsilon\left(u^A(b^\varepsilon_{C_2}(e_H)) - u^A\left(b^\varepsilon_{C_1}(e_H)\right)\right)\end{array}\right\} \tag{25}$$

is concave in e_H.

Proof. In Section A.3.17. ■

The above result shows that A's incentive to exert effort is higher when the precision of the signal is lower: when it becomes more likely to reach the bad outcome or to have an intermediate intervention (that costs C_1) when the outcome would have otherwise been $y = G$. In this case, observing $y = G$ increases the public's belief that A is an H-type, while observing the other two outcomes decreases the public's belief. This reputational effect in turn increases the incentive for A to exert effort.

Ex-ante and ex-post trade-offs. We can now show that the same trade-offs highlighted in the simple model from section 3 exist in the model with the general payoff from reputation. Define

$$b_{outcome}^{sep} \equiv \frac{\Pr(\theta = H | outcome,\ separated\ institutions)}{1 - \Pr(\theta = H | outcome,\ separated\ institutions)},$$

and $b_{outcome}^{\varepsilon}$ as in (24). We use these expressions to derive the conditions under which separate institutions generate higher regulatory effort and lower expected costs of the outcome.

First, we can conclude that the full revelation equilibrium in the case with separate institutions leads to the same equilibrium outcomes are the centralized institution.

Proposition 11 *The equilibrium regulatory effort and the expected cost of the outcome are the same under both institutional structures if the following conditions are satisfied:*

- *When $s = G$, $\iota = \iota^G$ gives A higher utility than $\iota = \iota^B$:*

$$\frac{q(e)(1-\varepsilon)}{q(e)(1-\varepsilon) + (1-q(e))\varepsilon} u^A\left(b_0^{sep}\right) + \frac{(1-q(e))\varepsilon}{q(e)(1-\varepsilon) + (1-q(e))\varepsilon} u^A\left(b_{C_2}^{sep}\right) \geq u^A(b_{C_1}^{sep}),$$

- *When $s = B$, $\iota = \iota^G$ gives A higher utility than $\iota = \iota^B$:*

$$\frac{q(e)\varepsilon}{q(e)\varepsilon + (1-q(e))(1-\varepsilon)} u^A(b_0^{sep}) + \frac{(1-q(e))(1-\varepsilon)}{q(e)\varepsilon + (1-q(e))(1-\varepsilon)} u^A(b_{C_2}^{sep}) \leq u^A(b_{C_1}^{sep}).$$

Proof. In Section A.3.18. ■

If the conditions of Proposition 11 are not satisfied, then the equilibrium with separate institutions features a different effort choice and intervention outcome compared to the centralized institution.

Proposition 12 *In the equilibrium with partial or full information withholding for the separated institutions, effort e_H is higher with separate institutions if the following 2 conditions are satisfied:*

1. $q'(e_H)\left\{u^A(b_0^{sep}(e_H)) - u^A(b_{C_2}^{sep}(e_H))\right\}$ *is concave;*

2. $MA_\varepsilon(e_H)$, *defined in (25), is concave.*

Proof. In Section A.3.19. ■

Proposition 12 shows that effort is higher under separated institutions whenever the marginal benefit of effort of concave in both institutional environments. This condition ensures that the public's beliefs do not change drastically as effort increases: they do not put much higher probability on A being an L-type if the outcome $y = G$ is not observed. If this were not the case, the reputational gains from more effort – or the reputational losses from an intervention or a bad outcome – would lead A to want to put in ever more effort to avoid the risk of his reputation decreasing. Therefore, these conditions ensure sufficient regularity to the payoff obtained from reputation. Assuming a sufficiently concave $q(e)$ or sufficiently low $q(e)$ and $\frac{c_L(e)}{c_H(e)}$ $\forall e \in [0,1]$ guarantee that the above conditions are satisfied.[25]

Proposition 13 *Suppose all the conditions in Propositions 10 and 12 are satisfied. Then the expected cost to the public is lower under separate institutions if ε and $\frac{C_1}{C_2}$ are sufficiently high.*

Proof. In Appendix Section A.3.20. ■

Proposition 13 shows that the conclusion derived from the simple model of Section 3 extends to a more general specification for the evolution of A's reputation. Under the regularity conditions described above, the same trade-off emerges: centralizing the institutional structure provides the benefit of using the information efficiently to save costs; however, regulatory effort decreases.

[25] The details are presented in the proof to the Proposition.

A.1.2 Integrated Objectives

So far, we have examined the case in which the objective of A and that of the central bank do not change with institutional structure. One may argue that the institutional centralization should bring about a change in the objective of A or D. In what follows, we study this case.

We consider the same setup as in Section 3, with the change that now the utility of A in the centralized institution is given by:

$$u^{com} = \alpha \left[u^A(\Pr(\theta = H|outcome)) - c_\theta\left(e\right)\right] + (1-\alpha)\left(-\iota(s)C_1 - (1-\iota(s))\mathbf{1}_{y=B}C_2\right),$$

while in the separate institution, it is as before,

$$u_\theta^{sep} = u^A(\Pr(\theta = H|outcome)) - c_\theta\left(e\right)$$

The objective of the decision maker remains that of minimizing the expected cost:

$$u = -C.$$

Under Assumption 2, this implies that D intervenes after $s = B$ and does not intervene after $s = G$.

The problem and the equilibrium characterization does not change in the case of separated institutions. In the case of institutional centralization, we can derive the following results.

Result 1 *In the centralized institution, regulatory effort decreases in α. Moreover, there exists $\alpha^* \in (0,1)$ such that A's effort is higher in the separated institution whenever $\alpha > \alpha^*$.*

The result from the analysis of A's problem,

$$\begin{aligned}&\max_e \alpha\left[\Pr(y = G, s = G|e) - c_\theta\left(e\right)\right] + \\ &(1-\alpha)\left(-\Pr(s = B|e)C_1 - \Pr(s = G, y = B|e)C_2\right),\end{aligned}$$

which leads to the first-order condition

$$q'(e)\left[(1-\varepsilon)+\frac{1-\alpha}{\alpha}\left(\varepsilon C_2-\varepsilon C_1+(1-\varepsilon)\,C_1\right)\right]=c'_\theta\,(e)\,.$$

Analyzing this condition, it follows that A's effort choice is higher whenever he places some positive weight on the central bank's objective. The distortions described in Section 3 persist, however, in the model with integrated regulatory objective, as long as A continues to place sufficient weight on reputation.

Result 2 *For $\alpha > \alpha^*$, the expected cost of a banking crisis is larger in the centralized institution than under separate institutions if ε and $\frac{C_1}{C_2}$ are sufficiently high.*

The result follows from the same analysis as in Proposition 3 whenever the equilibrium effort is higher with separate institutions – the case when A still places sufficient weight on his reputation relative to the objective of the decision maker.

A.2 Dynamic Analysis when the L-type also works

A.2.1 Best PBE without Communication

The equivalence of the best PBE to the best sustainable equilibrium is readily extended to the case in which L-type works. Hence, we are left to characterize the best sustainable equilibrium.

Once the principal decides the replacement probability after each event, it determines the best response of the $L-$type. In particular, once the principal decides the target effort for the $H-$type, e_z^H, and the continuation promised utility for the $H-$type, $V'(o_z)$, the $L-$type maximizes his own payoff:

$$\begin{aligned} V_L(\mu,V) &= \max_{\{e_z\}_z}\int_z (1-p_z)\left\{(1-\delta)\left(1-c_L(e_z)\right)\right. \\ &\quad \left.+\delta\sum_{o_z}\Pr\left(o_z|\psi_z,e_z,\iota_z\right)V_L\left(\mu'\left(\mu,\psi_z,e_z^L,e_z^H,\iota_z,o_z\right),V'(o_z)\right)\right\}dz. \end{aligned}$$

Here $V_L(\mu, V)$ is the value function of the regulator with type L when the belief is μ and the promised utility for the $H-$type is V. In the maximization problem, $\left\{e_z^L, e_z^H\right\}_z$ is the equilibrium effort. Since the principal believes that the agent follows the equilibrium effort when calculating the belief update, effort e_z^L is not controlled by the $L-$type agent.

The principal's problem is to maximize

$$\begin{aligned} J(\mu, V) &= \max_{(p_z, \psi_z, e_z^H, e_z^L, \iota_z, V_z')_z} \int_z [p_z \bar{J} + (1-p_z)\{(1-\delta)[\mu \mathbb{E}\left[-C|\psi_z, e_z^H, \iota_z\right] \\ &\quad + (1-\mu)\mathbb{E}\left[-C|\psi_z, e_z^L, \iota_z\right]] \\ &\quad + \delta \sum_{o_z} \Pr(o_z|\mu, \psi_z, e_z, \iota_z) J\left(\mu'\left(\mu, \psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V_z'(o_z)\right)\}] dz. \end{aligned}$$

At the same time, we keep track of type L's value function:

$$\begin{aligned} V_L(\mu, V) &= \int_z (1-p_z)\{(1-\delta)\left(1 - c_L\left(e_z^L\right)\right) \\ &\quad + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^L, \iota_z\right) V_L\left(\mu'\left(\mu, \psi_z, e_z^L, e_z^H, \iota_z, o_z\right), V'(o_z)\right)\} dz. \end{aligned}$$

1. Promise keeping constraint:

$$V = \int_z (1-p_z)\left\{(1-\delta)\left[1 - c_H\left(e_z^H\right)\right] + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^H, \iota_z\right) V_z'(o_z)\right\} dz; \qquad (26)$$

2. Incentive compatibility constraint for the high type: For each z with $p_z > 0$, we have

$$e_z^H \in \arg\max_e \left\{(1-\delta)[1 - c_H(e)] + \delta \sum_{o_z} \Pr(o_z|\psi_z, e, \iota_z) V_z'(o_z)\right\}; \qquad (27)$$

3. Incentive compatibility constraint for the high type: For each z with $p_z > 0$, we have

$$e_z^L \in \arg\max_e \left\{ \begin{array}{c} (1-\delta)(1 - c_L(e)) \\ +\delta \sum_{o_z} \Pr(o_z|\psi_z, e, \iota_z) V_L\left(\mu'\left(\mu, \psi_z, e_z^L, e_z^H, \iota_z, o_z\right), V'(o_z)\right) \end{array} \right\}. \qquad (28)$$

Hence now the problem is recursive on social welfare function $J(\mu, V)$ and type L's value function $V_L(\mu, V)$. The rest of the analysis is accordingly generalized.

A.2.2 Best PBE with Communication

The equivalence of the best PBE to the best mechanism design problem equilibrium is readily extended to the case in which L-type works. Hence, we are left to solve the mechanism design problem.

We will show that the above method is robust even when the low type works. If the low type works, then the social welfare $J_L(V_{LH})$ from the low type is not globally linear given V_{LH} (the value that the low type obtains from $\mathcal{C}_H^*$). One may wonder how to modify the step from (43) to (44).

In particular, $J_L(V)$ solves the following problem: Since we condition that the type is low (and guess $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ is not binding), we have

$$J_L(V) = \max_{(p_z, \psi_z, e_z, \iota_z, V_z')} \int_z \left\{ \begin{array}{c} (1-\delta)(1-\mu)\mathbb{E}[-C|\psi_z, e_z, \iota_z] \\ +\delta \sum_{o_z} \Pr(o_z|\psi_z, e_z, \iota_z) J_L(V_z'(o_z)) \end{array} \right\} dz$$

subject to

1. Promise keeping constraint:

$$V = \int_z (1-p_z) \left\{ (1-\delta)[1-c_L(e_z)] + \delta \sum_{o_z} \Pr(o_z|\psi_z, e_z, \iota_z) V_z'(o_z) \right\} dz; \qquad (29)$$

2. Incentive compatibility constraint: for each z with $p_z > 0$, we have

$$e_z \in \arg\max_e \left\{ (1-\delta)[1-c_L(e)] + \delta \sum_{o_z} \Pr(o_z|\psi_z, e, \iota_z) V_z'(o_z) \right\}. \qquad (30)$$

By the same proof as Lemma 2, there exists $\bar{V}_L \geq 1-\delta$ such that $J_L(V)$ is linear for $V \in [0, \bar{V}_L]$:

Lemma 11 *For each $\bar{J}$, there exists $\bar{V}_L \geq 1-\delta$ such that $J_L(V)$ is linear for $V \in \left[0, \bar{V}_L\right]$.*

Suppose we relax the principal's problem further so that the value from the $L-$type is $\hat{J}_L(V)$, where $\hat{J}_L(V)$ is linear extension of $J_L(V)$ for $V > \bar{V}_L$:

$$\hat{J}_L(V) = \bar{J} - \frac{\bar{J} - J_L\left(\bar{V}_L\right)}{\bar{V}_L} V.$$

If $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ is not binding and $\hat{J}_L(V)$ is the true welfare from the $L-$type, then the method in Section 4.3 will go through. Relegating the algebraic derivation to Section A.3.21, the principal solves

$$\begin{aligned} J(\mu, V) \quad = \quad & \max_{\{p_z, \psi_z, e_z^H, e_z^L, \iota_z, V_z'\}_z} \int_z [p_z \bar{J} + (1-p_z)\{(1-\delta)\{\mu \mathbb{E}\left[-C|\psi_z, e_z^H, \iota_z\right] \\ & + (1-\mu)[\bar{J} - \frac{\bar{J} - J_L\left(\bar{V}_L\right)}{\bar{V}_L}\left[1 - c_L\left(e_z^L\right)\right]]\} \\ & + \delta \sum_{o_z} \Pr\left(o_z|\mu, \psi_z, e_z^H, e_z^L, \iota_z\right) J\left(\mu'\left(\mu, \psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V'(o_z)\right)\}]dz. \quad (31) \end{aligned}$$

At the same time, we keep track of

$$\begin{aligned} V_{LH}(\mu, V) \quad = \quad & \int_z (1-p_z)\{(1-\delta)\left[1 - c_L\left(e_z^L\right)\right] \\ & + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^H, e_z^L, \iota_z\right) V_{LH}\left(\mu'\left(\mu, \psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V'(o_z)\right)\}dz. \end{aligned}$$

The constraints are:

1. Promise keeping constraint:

$$V = \int_z (1-p_z)\left\{(1-\delta)\left[1 - c_H\left(e_z^H\right)\right] + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^H, \iota_z\right) V_z'(o_z)\right\} dz;$$

2. Incentive compatibility constraint for the high type: For each z with $p_z > 0$, we have

$$e_z^H \in \arg\max_e \left\{ (1-\delta)\left[1 - c_H(e)\right] + \delta \sum_{o_z} \Pr\left(o_z | \psi_z, e, \iota_z\right) V_z'\left(o_z\right) \right\};$$

3. Incentive compatibility constraint for the high type: For each z with $p_z > 0$, we have

$$e_z^L \in \arg\max_e \left\{ \begin{array}{c} (1-\delta)\left[1 - c_L(e)\right] \\ +\delta \sum_{o_z} \Pr\left(o_z | \psi_z, e, \iota_z\right) V_{LH}\left(\mu\left(\mu, \psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V'\left(o_z\right)\right) \end{array} \right\}.$$

Finally, $\bar{J}$ is the fixed point such that

$$\bar{J} = \max_V \left\{ \left(1-\mu_H\right) \hat{J}_L\left(V_{LH}\left(\mu_H, V\right)\right) + \mu_H J_H\left(\mu_H, V\right) \right\} = \max_V J\left(\mu_H, V\right).$$

Let us call this problem "relaxed problem $\mathcal{C}_L^* \succcurlyeq_L \mathcal{C}_H^*$."

To show that $J(\mu, V)$ is the solution for the original problem, we are left to verify that there exists $V^* \in \arg\max_{\hat{V}} J\left(\mu_H, \hat{V}\right)$ such that, given $\mathcal{C}_H^*$ associated with this dynamic programming with initial state $\left(\mu_H, V^*\right)$, (i) the value that the low type obtains from $\mathcal{C}_L^*$ is in $\left[0, \bar{V}_L\right]$ (and so indeed $J_L(V)$ and $\hat{J}_L(V)$ coincide) and (ii) $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ holds.

We can show that we can always find such a V^*:

Lemma 12 *There exists $V^* \in \arg\max_{\hat{V}} J\left(\mu_H, \hat{V}\right)$ such that, given $\mathcal{C}_H^*$ associated with this dynamic programming with initial state $\left(\mu_H, V^*\right)$, (i) $V_{LH}\left(\mathcal{C}_H^*\right) \in \left[0, \bar{V}_L\right]$ and (ii) $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ holds.*

Proof. In Appendix Section A.3.22. ■

Since the problem is parallel to the dynamic game, as in Lemma 2, we can show that $J\left(\mu_H, V\right)$ is constant for $V \in \left[0, \bar{V}\right]$ with $\bar{V} \geq 1 - \delta$. That is, $\arg\max_{\hat{V}} J\left(\mu_H, \hat{V}\right) = \left[0, \bar{V}\right]$. This means that we can take V^* sufficiently small so that $V_{LH}\left(\mathcal{C}_H^*\right)$ is no more than $\bar{V}_L$.

Moreover, unless $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ prevents the principal from taking $V^* \leq \bar{V}$, changing the value promised to the $H-$type does not change the ex-ante welfare. Hence, we do not need to worry about $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ unless this constraint prevents the principal from taking $V^* \leq \bar{V}$; however if this were binding, then the principal could improve its welfare by the following replacement, which yields a contradiction: upon arrival of the regulator, P replaces him regardless of his type with a positive probability (using the public randomization). Only if he is not replaced, the contracts $(\mathcal{C}_H^*, \mathcal{C}_L^*)$ are offered. Since the replacement does not depend on the reported type, it does not change the incentive compatibility, but reduces the promised value V^* for the high type.

One may wonder why it is optimal for the principal to choose V^* very small. A small V^* means that the $H-$types are replaced with a high probability, and it may seem inefficient; however, keeping the $H-$type agent comes with the cost: we cannot replace the $L-$type quickly enough. By committing to the following strategy, the principal can reduce V^* without hurting the ex-ante efficiency: no matter what the agent says, the principal replaces him with probability p. From the ex-ante perspective, the principal obtains

$$\underbrace{\bar{J}}_{\substack{\text{ex ante}\\ \text{social welfare}}} \quad = \quad p\underbrace{\bar{J}}_{\substack{\text{the principal}\\ \text{gets a new draw}}} \quad +(1-p)\underbrace{\begin{bmatrix} \mu_H \times \text{social welfare from type } H \\ +(1-\mu_H) \times \text{social welfare from type } L \end{bmatrix}}_{\substack{=\bar{J} \text{ by dynamic}\\ \text{programming}}}.$$

That is, regardless of this replacement, the principal obtains $\bar{J}$ and so is indifferent.

This argument of using the replacement relies on the fact that the replacement is not costly. If the replacement is costly, then the above solution gives us the approximation of the second best. In other words, the above solution is the bench mark when the replacement is costless.

A.3 Proofs

A.3.1 Proof of Proposition 1

Part 1. Effort strategies

It is sufficient to show that e_θ is weakly increasing in θ since if $e_H = e_L$, then $b_{C_2} = b_{C_1} = b_0^{sep} = \frac{\mu_H}{\mu_L}$, and each agent puts in zero effort.

Let $v(\theta)$ be the equilibrium payoff of A with type θ. Take two types θ, θ' with $\theta < \theta'$. When A with type θ' takes the same effort e_θ and message strategy as type θ, he induces the same distribution of the reputation as θ. For such a deviation not to be profitable, we have

$$v(\theta') \geq v(\theta) + c_\theta(e_\theta) - c_{\theta'}(e_\theta).$$

Note that $-(c_{\theta'}(e_\theta) - c_\theta(e_\theta)) = c_\theta(e_\theta) - c_{\theta'}(e_\theta)$ is the extra cost that type θ' has to pay to exert effort e_θ compared to type θ. Similarly, so that A with θ does not want to pretend to be type θ', we have

$$v(\theta) \geq v(\theta') + c_{\theta'}(e_{\theta'}) - c_\theta(e_{\theta'}).$$

In total, we have

$$c_\theta(e_{\theta'}) - c_{\theta'}(e_{\theta'}) \geq v(\theta') - v(\theta) \geq c_\theta(e_\theta) - c_{\theta'}(e_\theta).$$

Rearranging these inequalities yields

$$c_\theta(e_{\theta'}) - c_\theta(e_\theta) \geq c_{\theta'}(e_{\theta'}) - c_{\theta'}(e_\theta).$$

Since we assume that the marginal cost of e is decreasing in θ, we have $e_{\theta'} \geq e_\theta$, as desired.

Part 2. Beliefs

We focus on the equilibrium with $e_H > e_L$.

2.a) Proof of $b_G > b_{C_2}$

For the sake of contradiction, suppose $b_0 \leq b_{C_2}$.

We first show that $b_{C_2} \geq b_{C_1}$. Again, suppose otherwise: $b_{C_2} < b_{C_1}$. Then each agent wants to send the message $m^B \in M$ to maximize the probability of D's action ι^B. If D is not indifferent between ι^G and ι^B after m^B, then this gives each agent the payoff of $u\left(b_{C_1}\right)$, and no agent has the incentive to put a positive effort and so $b_0 = b_{C_2} = b_{C_1}$, which is a contradiction. If D is indifferent, then D induces the intermediate intervention with probability p (note that p is independent of s or e_θ since each agent sends the message m^* to maximize p). Hence A's payoff is

$$(1-p)\left[q\left(e_\theta\right)u\left(b_0\right) + \left(1-q\left(e_\theta\right)\right)u\left(b_{C_2}\right)\right] + pu\left(b_{C_1}\right) - c_\theta\left(e_\theta\right).$$

The marginal benefit of increasing e_θ is non-positive [recall we are assuming $b_0 \leq b_{C_2} < b_{C_1}$] while the marginal cost is positive. Hence, no agent puts in positive effort. Therefore, we have $b_{C_1} \leq b_{C_2}$.

Hence under the initial assumption of $b_0 \leq b_{C_2}$, there are following two cases: $b_{C_1} < b_0 \leq b_{C_2}$ or $b_0 \leq b_{C_1} \leq b_{C_2}$.

We now show that $b_{C_1} < b_0 \leq b_{C_2}$ cannot be the case. Suppose $b_{C_1} < b_0 \leq b_{C_2}$ for the sake of contradiction. Then each agent wants to send the message m^* to minimize the probability of D's action ι^B. Let p be the probability with which D induces ι^B (again p is independent of s or e_θ since each agent sends the message to minimize p). Hence A's reputation is

$$b_0 = \frac{\mu_H q\left(e_H\right)(1-p)}{\mu_L q\left(e_L\right)(1-p)} = \frac{\mu_H q\left(e_H\right)}{\mu_L q\left(e_L\right)}$$

and

$$b_{C_2} = \frac{\mu_H\left(1-q\left(e_H\right)\right)(1-p)}{\mu_L\left(1-q\left(e_H\right)\right)(1-p)} = \frac{\mu_H\left(1-q\left(e_H\right)\right)}{\mu_L\left(1-q\left(e_H\right)\right)}.$$

From Part 1 above, it follows that $b_{C_2} < b_0$, so we reach a contradiction.

Finally, the remaining case is $b_0 \leq b_{C_1} \leq b_{C_2}$. Let $p(m)$ be the probability that D takes

ι^B after message m. Then, A's payoff is

$$\Pr\left(s|e_\theta\right)\max_m\left\{\left[1-p\left(m\right)\right]\left[\Pr\left(y=G|e_\theta,s\right)u\left(b_0\right)+\Pr\left(y=B|e_\theta,s\right)u\left(b_{C_2}\right)\right]+p\left(m\right)u\left(b_{C_1}\right)\right\}.$$

Since $\Pr\left(y=G|e,s\right)$ is increasing in e and s (we say $G>B$ for s) and we assumed $b_0\leq b_{C_2}$, we have that

$$\Pr\left(y=G|e_\theta,s\right)u\left(b_0\right)+\Pr\left(y=G|e_\theta,s\right)u\left(b_{C_2}\right)$$

is decreasing in e_θ and s. Since $\Pr\left(s=G|e_\theta\right)$ is increasing in e_θ, we have

$$\Pr\left(s|e_\theta\right)\max_m\left\{\left[1-p\left(m\right)\right]\left[\Pr\left(y=G|e_\theta,s\right)u\left(b_0\right)+\Pr\left(y=B|e_\theta,s\right)u\left(b_{C_2}\right)\right]+p\left(m\right)u\left(b_{C_1}\right)\right\}$$

decreasing in e_θ. Therefore, no agent has an incentive to put a positive effort, which is a contradiction.

In total, we have contradictions unless we have $b_0>b_{C_2}$.

2.b) Proof for $b_0>\frac{\mu_H}{\mu_L}\geq b_{C_1},b_{C_2}$

In the separate institution, let $\gamma_{s,\theta}$ be the probability of ι^G given type θ and signal s. We have

$$\begin{array}{rcl}
b_0 & = & \dfrac{\mu_H\left[q\left(e_H\right)\left(1-\varepsilon_G\right)\gamma_{G,H}+q\left(e_H\right)\varepsilon_G\gamma_{B,H}\right]}{\mu_L\left[q\left(e_L\right)\left(1-\varepsilon_G\right)\gamma_{G,L}+q\left(e_L\right)\varepsilon_B\gamma_{B,L}\right]};\\
b_{C_1} & = & \dfrac{\mu_H\left\{\begin{array}{c}\left[q\left(e_H\right)\left(1-\varepsilon_G\right)+\left(1-q\left(e_H\right)\right)\varepsilon_B\right]\left(1-\gamma_{G,H}\right)\\+\left[q\left(e_H\right)\varepsilon_G+\left(1-q\left(e_H\right)\right)\left(1-\varepsilon_B\right)\right]\left(1-\gamma_{B,H}\right)\end{array}\right\}}{\mu_L\left\{\begin{array}{c}\left[q\left(e_L\right)\left(1-\varepsilon_G\right)+\left(1-q\left(e_L\right)\right)\varepsilon_B\right]\left(1-\gamma_{G,L}\right)\\+\left[q\left(e_L\right)\varepsilon_G+\left(1-q\left(e_L\right)\right)\left(1-\varepsilon_B\right)\right]\left(1-\gamma_{B,L}\right)\end{array}\right\}};\\
b_{C_2} & = & \dfrac{\mu_H\left[\left(1-q\left(e_H\right)\right)\varepsilon_B\gamma_{G,H}+\left(1-q\left(e_H\right)\right)\left(1-\varepsilon_B\right)\gamma_{B,H}\right]}{\mu_L\left[\left(1-q\left(e_L\right)\right)\varepsilon_B\gamma_{G,L}+\left(1-q\left(e_L\right)\right)\left(1-\varepsilon_B\right)\gamma_{B,L}\right]}.
\end{array}$$

A with type θ sends m to minimize the probability of ι^B if

$$\Pr\left(y=G|e_\theta,s\right)u\left(b_0\right)+\Pr\left(y=B|e_\theta,s\right)u\left(b_{C_2}\right)>u\left(b_{C_1}\right)$$

and only if

$$\Pr\left(y=G|e_\theta,s\right)u\left(b_0\right)+\Pr\left(y=B|e_\theta,s\right)u\left(b_{C_2}\right)\geq u\left(b_{C_1}\right).$$

Since $b_0 > b_{C_2}$, the left hand side is increasing in e and s. Since $e_H > e_L$, we have $\gamma_{G,\theta}\geq\gamma_{B,\theta}$ for each θ and $\gamma_{s,H}\geq\gamma_{s,L}$ for each s.

If $\gamma_{B,H}=0$, then $\gamma_{B,L}=0$ and so

$$\begin{aligned}
b_0-\frac{\mu_H}{\mu_L} &= \frac{\mu_H}{\mu_L}\left(\frac{q\left(e_H\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_H\right)\varepsilon\gamma_{B,H}}{q\left(e_L\right)\left(1-\varepsilon\right)\gamma_{G,L}+q\left(e_L\right)\varepsilon\gamma_{B,L}}-1\right)\\
&= \frac{\mu_H}{\mu_L}\left(\frac{q\left(e_H\right)\gamma_{G,H}}{q\left(e_L\right)\gamma_{G,L}}-1\right)>0
\end{aligned}$$

since $e_H > e_L$ and $\gamma_{G,H}\geq\gamma_{G,L}$. On the other hand, if $\gamma_{B,H}>0$, then we have

$$\begin{aligned}
b_0-\frac{\mu_H}{\mu_L} &= \frac{\mu_H}{\mu_L}\left(\frac{q\left(e_H\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_H\right)\varepsilon\gamma_{B,H}}{q\left(e_L\right)\left(1-\varepsilon\right)\gamma_{G,L}+q\left(e_L\right)\varepsilon\gamma_{B,L}}-1\right)\\
&\geq \frac{\mu_H}{\mu_L}\left(\frac{q\left(e_H\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_H\right)\varepsilon\gamma_{B,H}}{q\left(e_L\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_L\right)\varepsilon\gamma_{B,H}}-1\right)\\
&= \frac{\mu_H}{\mu_L}\left(\frac{\left(q\left(e_H\right)-q\left(e_L\right)\right)\left(\left(1-\varepsilon\right)\gamma_{G,H}+\varepsilon\gamma_{B,H}\right)}{q\left(e_L\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_L\right)\varepsilon\gamma_{B,H}}\right)\\
&\geq \frac{\mu_H}{\mu_L}\left(\frac{\left(q\left(e_H\right)-q\left(e_L\right)\right)\left(\left(1-\varepsilon\right)\gamma_{B,H}+\varepsilon\gamma_{B,H}\right)}{q\left(e_L\right)\left(1-\varepsilon\right)\gamma_{G,H}+q\left(e_L\right)\varepsilon\gamma_{B,H}}\right)\\
&= \frac{\mu_H}{\mu_L}\left(\frac{\left(q\left(e_H\right)-q\left(e_L\right)\right)\left(1-2\varepsilon\right)\gamma_{B,H}}{q\left(e_L\right)\left(\left(1-\varepsilon\right)\gamma_{G,H}+\varepsilon\gamma_{B,H}\right)}\right)>0.
\end{aligned}$$

In total, we have $b_0-\frac{\mu_H}{\mu_L}>0$.

Suppose $b_{C_2}-\frac{\mu_H}{\mu_L}\geq 0$. Then since the belief is updated positively after both final $y=G$ and $y=B$, in order for the belief to be consistent with the prior, we have $b_{C_1}-\frac{\mu_H}{\mu_L}<0$. Hence the agent wants to minimize the probability of ι^B. By Assumption (5), sending the same message between $s=G$ and $s=B$ allows the agent to achieve $\gamma_{B,H}=\gamma_{G,H}=\gamma_{G,L}=$

$\gamma_{G,L} = 1$ and so

$$\begin{aligned} b_{C_2} - \frac{\mu_H}{\mu_L} &= \frac{\mu_H}{\mu_L}\left(\frac{1-q(e_H)}{1-q(e_L)} - 1\right) \\ &= -\frac{\mu_H}{\mu_L}\left(\frac{q(e_H)-q(e_L)}{1-q(e_L)}\right) < 0, \end{aligned}$$

which is a contradiction. Therefore, we have $b_{C_2} - \frac{\mu_H}{\mu_L} < 0$.

Finally, if $\gamma_{G,L} = \gamma_{B,L}$, then since we have $\gamma_{G,H} \geq \gamma_{G,L}$ and $\gamma_{B,H} \geq \gamma_{B,L}$, we have

$$\begin{aligned} b_{C_1} - \frac{\mu_H}{\mu_L} &= \frac{\mu_H}{\mu_L}\left(\frac{\left\{\begin{array}{c} [q(e_H)(1-\varepsilon) + (1-q(e_H))\varepsilon]\big(1-\gamma_{G,H}\big) \\ +[q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon)]\big(1-\gamma_{B,H}\big) \end{array}\right\}}{\left\{\begin{array}{c} [q(e_L)(1-\varepsilon) + (1-q(e_L))\varepsilon]\big(1-\gamma_{G,L}\big) \\ +[q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)]\big(1-\gamma_{B,L}\big) \end{array}\right\}} - 1\right) \\ &\leq \frac{\mu_H}{\mu_L}\left(\frac{\left\{\begin{array}{c} [q(e_H)(1-\varepsilon) + (1-q(e_H))\varepsilon]\big(1-\gamma_{G,L}\big) \\ +[q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon)](1-\gamma_{GL}) \end{array}\right\}}{\left\{\begin{array}{c} [q(e_L)(1-\varepsilon) + (1-q(e_L))\varepsilon]\big(1-\gamma_{G,L}\big) \\ +[q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)]\big(1-\gamma_{G,L}\big) \end{array}\right\}} - 1\right) < 0 \end{aligned}$$

since $e_H > e_L$. On the other hand, if $\gamma_{G,L} > \gamma_{B,L}$, then

$$
\begin{aligned}
b_{C_1} - \frac{\mu_H}{\mu_L} &= \frac{\mu_H}{\mu_L}\left(\frac{\left\{ \begin{array}{c} \left[q\left(e_H\right)\left(1-\varepsilon\right) + \left(1-q\left(e_H\right)\right)\varepsilon\right]\left(1-\gamma_{G,H}\right) \\ +\left[q\left(e_H\right)\varepsilon + \left(1-q\left(e_H\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,H}\right) \end{array} \right\}}{\left\{ \begin{array}{c} \left[q\left(e_L\right)\left(1-\varepsilon\right) + \left(1-q\left(e_L\right)\right)\varepsilon\right]\left(1-\gamma_{G,L}\right) \\ +\left[q\left(e_L\right)\varepsilon + \left(1-q\left(e_L\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,L}\right) \end{array} \right\}} - 1 \right) \\
&\leq \frac{\mu_H}{\mu_L}\left(\frac{\left\{ \begin{array}{c} \left[q\left(e_H\right)\left(1-\varepsilon\right) + \left(1-q\left(e_H\right)\right)\varepsilon\right]\left(1-\gamma_{G,L}\right) \\ +\left[q\left(e_H\right)\varepsilon + \left(1-q\left(e_H\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,L}\right) \end{array} \right\}}{\left\{ \begin{array}{c} \left[q\left(e_L\right)\left(1-\varepsilon\right) + \left(1-q\left(e_L\right)\right)\varepsilon\right]\left(1-\gamma_{G,L}\right) \\ +\left[q\left(e_L\right)\varepsilon + \left(1-q\left(e_L\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,L}\right) \end{array} \right\}} - 1 \right) \\
&= \frac{\mu_H}{\mu_L}\left(q\left(e_H\right) - q\left(e_L\right)\right) \frac{\left[\left(1-2\varepsilon\right)\left(1-\gamma_{G,L}\right) - \left(1-2\varepsilon\right)\left(1-\gamma_{B,L}\right)\right]}{\mu_L \left\{ \begin{array}{c} \left[q\left(e_L\right)\left(1-\varepsilon\right) + \left(1-q\left(e_L\right)\right)\varepsilon\right]\left(1-\gamma_{G,L}\right) \\ +\left[q\left(e_L\right)\varepsilon + \left(1-q\left(e_L\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,L}\right) \end{array} \right\}} \\
&< \frac{\mu_H}{\mu_L}\left(q\left(e_H\right) - q\left(e_L\right)\right) \frac{\left[\left(1-2\varepsilon\right)\left(1-\gamma_{G,L}\right) - \left(1-2\varepsilon\right)\left(1-\gamma_{G,L}\right)\right]}{\mu_L \left\{ \begin{array}{c} \left[q\left(e_L\right)\left(1-\varepsilon\right) + \left(1-q\left(e_L\right)\right)\varepsilon\right]\left(1-\gamma_{G,L}\right) \\ +\left[q\left(e_L\right)\varepsilon + \left(1-q\left(e_L\right)\right)\left(1-\varepsilon\right)\right]\left(1-\gamma_{B,L}\right) \end{array} \right\}} \\
&= 0.
\end{aligned}
$$

In total, we have $b_{C_1} < \frac{\mu_H}{\mu_L}$ as desired.

Part 3. Messaging strategy

If $e_H = e_L = 0$, since $b_{C_1} = b_{C_2} = b_0 = \frac{\mu}{1-\mu}$, the agent is indifferent between any message. Hence any $m\left(s\right) : \left\{G, B\right\} \rightarrow M$ is an equilibrium strategy.

If $e_H > 0$, given the results shown in part 2 above, and the form of the utility function $u(b)$, the agent A receives payoff 0 if the decision maker takes ι^B. Therefore, each agent sends the message m which minimizes the probability of ι^B.

Part 4. Intervention strategy

If $e_H = e_L = 0$, depending on the message strategy, there is multiplicity of the intervention strategy. For example, if $m\left(s\right)$ is such that $M = \left\{G, B\right\}$ and $m\left(s\right) = s$ (that is, message G signifies the signal G and message B signifies the signal B), then (4) implies that D takes ι^B if and only if she receives message B. Another example is that $m(s)$ is such that $M = \{\emptyset\}$ and $m(s) = \{\emptyset\}$ for each $s \in \{G, B\}$ (babbling equilibrium). Then (5) implies that D takes ι^G all the time.

If $e_H > 0$, given the messaging strategy, each agent sends $m \in M$ to minimize the probability of ι^B. Given (5), D takes ι^G on equilibrium path.

A.3.2 Proof of Proposition 2

Part 1. Effort strategies

The proof that $e_H \geq e_L$, with strict inequality if $e_H > 0$, is the same as the one in the proof to Proposition 1, Part 1.

Part 2. Beliefs

In the combined institution, we have

$$\begin{aligned} b_0 &= \frac{\mu_H q\left(e_H\right)\left(1-\varepsilon_G\right)}{\mu_L q\left(e_L\right)\left(1-\varepsilon_G\right)} = \frac{\mu_H q\left(e_H\right)}{\mu_L q\left(e_L\right)}; \\ b_{C_1} &= \frac{\mu_H\left[q\left(e_H\right)\varepsilon_G + \left(1-q\left(e_H\right)\right)\left(1-\varepsilon_B\right)\right]\left(1-\gamma_{B,H}\right)}{\mu_L\left[q\left(e_L\right)\varepsilon_G + \left(1-q\left(e_L\right)\right)\left(1-\varepsilon_B\right)\right]\left(1-\gamma_{B,L}\right)}; \\ b_{C_2} &= \frac{\mu_H\left(1-q\left(e_H\right)\right)\varepsilon_B}{\mu_L\left(1-q\left(e_L\right)\right)\varepsilon_B} = \frac{\mu_H\left(1-q\left(e_H\right)\right)}{\mu_L\left(1-q\left(e_L\right)\right)}. \end{aligned}$$

This is essentially the same as in the case with separate institutions described in Part 2 of the proof to Proposition 1, with $\gamma_{G,H} = \gamma_{G,L} = 1$ and $\gamma_{B,H} = \gamma_{B,L} = 0$. So we have $b_0 > \frac{\mu_H}{\mu_L} > b_{C_1}, b_{C_2}$.

Part 3. Intervention Strategy

The expected cost of i^G after signal $s = B$ is

$$C_2 \Pr(y = B|s = B) = C_2 \frac{(1-q(e))(1-\varepsilon)}{q(e)\varepsilon + (1-q(e))(1-\varepsilon)},$$

while the expected cost of ι^B is C_1. Since $q\left(e\right)$ is increasing in e, we have

$$\frac{1}{\frac{q(e)\varepsilon}{(1-q(e))(1-\varepsilon)}+1} \geq \frac{1}{\frac{q(1)\varepsilon}{(1-q(1))(1-\varepsilon)}+1},$$

and so

$$C_2\frac{\left(1-q(e)\right)\left(1-\varepsilon\right)}{q(e)\varepsilon+\left(1-q(e)\right)\left(1-\varepsilon\right)} \geq C_2\left(\frac{q(1)\varepsilon}{\left(1-q(1)\right)\left(1-\varepsilon\right)}+1\right)^{-1} \tag{32}$$

Assumption 2 states that

$$C_2\left(\frac{q(1)\varepsilon}{\left(1-q(1)\right)\left(1-\varepsilon\right)}+1\right)^{-1} > C_1, \tag{33}$$

so (32) and (33) imply

$$C_2\frac{\left(1-q(e)\right)\left(1-\varepsilon\right)}{q(e)\varepsilon+\left(1-q(e)\right)\left(1-\varepsilon\right)} > C_1.$$

Therefore, the optimal strategy is ι^B if $s=B$.

If $s=G$, the expected cost of ι^G is

$$C_2\Pr(y=B|s=G) = C_2\frac{\left(1-q(e)\right)\varepsilon}{\left(1-q(e)\right)\varepsilon+q(e)\left(1-\varepsilon\right)},$$

while the cost of ι^B is C_1. Since $q\left(e\right)$ is increasing, we have

$$\frac{1}{1+\frac{q(e)(1-\varepsilon)}{(1-q(e))\varepsilon}} \leq \frac{1}{1+\frac{q(0)(1-\varepsilon)}{(1-q(0))\varepsilon}},$$

and so

$$C_2\left(1+\frac{q(e)\left(1-\varepsilon\right)}{\left(1-q(e)\right)\varepsilon}\right)^{-1} \leq C_2\left(1+\frac{q(0)\left(1-\varepsilon\right)}{\left(1-q(0)\right)\varepsilon}\right)^{-1}. \tag{34}$$

Given Assumption 2,

$$C_2\left(1+\frac{q(0)\left(1-\varepsilon\right)}{\left(1-q(0)\right)\varepsilon}\right)^{-1} < C_1, \tag{35}$$

so (34) and (35) imply

$$C_2\left(1+\frac{q(e)\left(1-\varepsilon\right)}{\left(1-q(e)\right)\varepsilon}\right)^{-1} < C_1,$$

and ι^G is therefore optimal when $s = G$.

Part 4. What if we allow the message exchange?

As in Parts 2 and 3 of Appendix A.3.1, each agent sends the message to minimize the probability of ι^B, and so no meaningful message exchange is made.

A.3.3 Proof of Lemma 1

Part 1: Concavity with respect to V.

We show that $J(\mu, V)$ is concave in V with a fixed μ. Suppose $V = \beta V_1 + (1-\beta) V_2$ for V_1, V_2; and let $\alpha[V_1]$ and $\alpha[V_2]$ be the optimal policy for (μ, V_1) and (μ, V_2), respectively.

Suppose the principal take $\alpha[V_1]$ with probability β and $\alpha[V_2]$ with probability $1-\beta$, according to the realization of the public randomization device.

We need to check (i) we satisfy the promise keeping and incentive compatibility; and the principal's value is at least $\beta J(\mu, V_1) + (1-\beta) J(\mu, V_2)$.

1. Since both $\alpha[V_1]$ and $\alpha[V_2]$ satisfy promise keeping

$$\begin{aligned} V_1 &= \int_{z_1} [(1 - p_{z_1}[V_1]) \{(1-\delta)[1 - c_H(e_{z_1}[V_1])] \\ &\quad + \delta \sum_{o_z} \Pr\left(o_z | \psi_{z_1}[V_1], e_{z_1}[V_1], \iota_{z_1}[V_1]\right) V'_z[V_1](o_z)\}] dz_1 \end{aligned}$$

and

$$\begin{aligned} V_2 &= \int_{z_2} [(1 - p_{z_2}[V_2]) \{(1-\delta)[1 - c_H(e_{z_2}[V_2])] \\ &\quad + \delta \sum_{o_z} \Pr\left(o_z | \psi_{z_2}[V_2], e_{z_2}[V_2], \iota_{z_2}[V_2]\right) V'_z[V_2](o_z)\}] dz_2, \end{aligned}$$

we have

$$\begin{aligned}V &= \beta V_1 + (1-\beta) V_2 \\ &= \beta[\int_{z_1} [(1-p_{z_1}[V_1]) \{(1-\delta)[1-c_H(e_{z_1}[V_1])] \\ &\quad +\delta \sum_{o_z} \Pr\left(o_z|\psi_{z_1}[V_1], e_{z_1}[V_1], \iota_{z_1}[V_1]\right) V_z'[V_1](o_z)\}]dz_1] \\ &\quad +(1-\beta)[\int_{z_2} [(1-p_{z_2}[V_2]) \{(1-\delta)[1-c_H(e_{z_2}[V_2])] \\ &\quad +\delta \sum_{o_z} \Pr\left(o_z|\psi_{z_2}[V_2], e_{z_2}[V_2], \iota_{z_2}[V_2]\right) V_z'[V_2](o_z)\}]dz_2.\end{aligned}$$

Hence promise keeping is satisfied.

2. Conditional on the realization of the public randomization device, since both $\alpha[V_1]$ and $\alpha[V_2]$ are incentive compatible, A with type H will take $e[V_1]$ and $e[V_2]$ according to the realization of the public randomization device. Hence incentive compatibility is satisfied.

Therefore, we are left to verify that this policy gives the principal at least the value $\beta J(\mu, V_1) + (1-\beta) J(\mu, V_2)$. With probability β, we achieve $J(\mu, V_1)$ and with probability $1-\beta$, we achieve $J(\mu, V_2)$ since we fix μ. Hence we achieve

$$\beta J(\mu, V_1) + (1-\beta) J(\mu, V_2),$$

as desired.

Part 2: Convexity with respect to μ.

Let $J(\mu, V, H)$ be the social welfare when the principal follows the optimal strategy given (μ, V), and the current type is actually H; and let $J(\mu, V, L)$ be the social welfare when the current type is actually L. We have

$$\begin{aligned}J(\mu, V) &= \mu J(\mu, V, H) + (1-\mu) J(\mu, V, L) \\ &= J(\mu, V, L) + \mu[J(\mu, V, H) - J(\mu, V, L)].\end{aligned}$$

Take μ, μ_1, μ_2 and β such that $\mu = \beta\mu_1 + (1-\beta)\,\mu_2$. By taking the strategy given (μ, V) when the belief is μ_1, the principal guarantees herself

$$\begin{aligned} & \mu_1 J(\mu, V, H) + (1-\mu_1)\, J(\mu, V, L) \\ = \; & J(\mu, V, L) + \mu_1 \left[J(\mu, V, H) - J(\mu, V, L)\right] \\ \leq \; & J(\mu_1, V)\,. \end{aligned}$$

Similarly, we have

$$J(\mu, V, L) + \mu_2 \left[J(\mu, V, H) - J(\mu, V, L)\right] \leq J(\mu_2, V)\,.$$

Hence $J(\mu, V)$ is on the line with y-intercept being equal to $J(\mu, V, L)$ and the slope being equal to $[J(\mu, V, H) - J(\mu, V, L)]$. On the other hand, $J(\mu_1, V)$ and $J(\mu_2, V)$ are above this line. Hence $J(\mu, V)$ is convex.

By the envelop theorem, this result also implies that

$$J_\mu(\mu, V) = J(\mu, V, H) - J(\mu, V, L)$$

since $J(\mu, V, H)$ and $J(\mu, V, L)$ depends on μ only through the strategy [recall that $J(\mu, V, \theta)$ is the welfare conditional on the type θ].

Monotonicity with respect to μ.

Suppose that $J(\mu, V) = J$ for some μ and V. Then, for higher value $\mu' > \mu$ and the same promised utility V, we have $J(\mu', V) \geq J$. To see why, if $(p_z, \psi_z, e_z, \iota_z, V_z')_z$ satisfies the promise keeping and incentive compatibility for μ, then it satisfies them for μ' as well since none of these constraints depends on μ.

Since $u^P(\mu, \psi_z, e_z, \iota_z)$ is increasing in μ, the usual result in dynamic programming –see Stokey (1989)– implies that $J(\mu, V)$ is increasing in μ.

A.3.4 Proof of Lemma 3

By feasibility, $V \in [0,1]$ since 1 is the maximum average payoff that the principal can deliver by letting $p_z = e_z = 0$ for each z. Given $V = 1$, since A does not work and stay forever, we have $J(\mu, V) = \underline{SW}$ for each μ. Hence we have $\bar{J} \geq \underline{SW}$.

On the other hand, if $V = 0$, then the principal replaces A right away, and so $J(\mu, 0) = \bar{J} \geq \underline{SW}$.

By Lemma 1, for each μ and $V \in [0,1]$, we have $J(\mu, V) \geq \underline{SW}$. Hence by the threat of switching to the no effort equilibrium, the principal has an incentive to follow the equilibrium institution design and replacement decision (note that the randomization device z makes the mixed strategy observable).

A.3.5 Proof of Lemma 2

We have $J(\mu, 0) = \bar{J}$ for each μ since the principal has to replace A right away. Hence we are left to prove the other two properties:

There exists $V(\mu)$ such that $J(\mu, V)$ is linear for $V \in [0, V(\mu)]$.

Suppose such $V(\mu)$ does not exist. By Lemma 1, this means that $J(\mu, V)$ is strictly concave near $V = 0$.

Take $V \in (0, (1-\delta))$. This means that the principal needs to stochastically replace A since otherwise A receives $1-\delta$ by not working. Let β be the probability of replacement. We have to have

$$\beta \times 0 + (1-\beta) \times \hat{V} = V,$$

where $\hat{V} \geq 1-\delta$ is the promised utility conditional on that A is not replaced.

By concavity of $J(\mu, \cdot)$, conditional on that the principal promises $\hat{V}$, the highest social welfare is $J\left(\mu, \hat{V}\right)$. Hence the social welfare is

$$\begin{aligned} & \beta \bar{J} + (1-\beta) J\left(\mu, \hat{V}\right) \\ = \quad & \bar{J} + (1-\beta)\left(J\left(\mu, \hat{V}\right) - \bar{J}\right) \end{aligned}$$

under the constraint that

$$\beta \times 0 + (1-\beta) \times \hat{V} = V \text{ and } \hat{V} \geq 1-\delta.$$

Substituting the constraint, the social welfare is

$$\begin{aligned} & \bar{J} + \frac{V}{\hat{V}} \left(J \left(\mu, \hat{V} \right) - \bar{J} \right) \\ = \; & \bar{J} + \frac{V}{\hat{V}} \left(J \left(\mu, \hat{V} \right) - J(\mu, 0) \right) \end{aligned}$$

Taking the derivative with respect to $\hat{V}$ (the differentiability follows from the envelop theorem), we have

$$\begin{aligned} & V \frac{J_{\hat{V}} \left(\mu, \hat{V} \right) \hat{V} - \left(J \left(\mu, \hat{V} \right) - J(\mu, 0) \right)}{\left(\hat{V} \right)^2} \\ = \; & V \frac{J(\mu, 0) + \left[J_{\hat{V}} \left(\mu, \hat{V} \right) \hat{V} - J \left(\mu, \hat{V} \right) \right]}{\left(\hat{V} \right)^2}. \end{aligned}$$

At $\hat{V} = 0$, this is zero. And taking the derivative of the numerator, we have

$$\begin{aligned} & \frac{d}{d\hat{V}} \left\{ J(\mu, 0) + \left[J_{\hat{V}} \left(\mu, \hat{V} \right) \hat{V} - J \left(\mu, \hat{V} \right) \right] \right\} \\ = \; & J_{\hat{V}\hat{V}} \left(\mu, \hat{V} \right) \hat{V} + J_{\hat{V}} \left(\mu, \hat{V} \right) - J_{\hat{V}} \left(\mu, \hat{V} \right) \\ = \; & J_{\hat{V}\hat{V}} \left(\mu, \hat{V} \right) \hat{V}. \end{aligned}$$

Since we assume $J(\mu, \cdot)$ is strictly concave, this is negative. Hence, the smallest $\hat{V} = 1 - \delta$ is optimal. Then we have

$$J(\mu, V) = \beta \bar{J} + (1-\beta) J \left(\mu, \hat{V} \right)$$

for $V \in [0, 1-\delta]$, which is linear in V.

For $\mu \geq \mu_H$, we have $V(\mu) > 1 - \delta$.

Suppose $\mu \geq \mu_H$. For the sake of contradiction, assume that $V(\mu) \leq 1 - \delta$. Then, in the above problem, $\hat{V} = 1 - \delta$ is the unique optimal. Since the principal cannot replace A in the current period after $\hat{V}$, this $1 - \delta$ is equal to what A can obtain without working at all and then replaced. Hence we have to make sure that $V'[\hat{V}](o_z) = 0$ for each o. Hence we have to implement $e = 0$.

Therefore, the effort has to be equal to 0. Hence the instantaneous social welfare is $(1 - \delta)\underline{SW}$. Moreover, since $V'[\hat{V}](o) = 0$ for each o, A will be replaced in the next period with probability one. Hence the continuation social welfare is $\delta \bar{J}$. In total,

$$J(\mu, 1 - \delta) = (1 - \delta)\underline{SW} + \delta \bar{J}. \tag{36}$$

For $\mu = \mu_H$, this implies that $J(\mu, 0) = \bar{J}$ and $J(\mu, V)$ is linear and less than $\bar{J}$ for each $V \in (0, 1 - \delta]$. By concavity, this means that $J(\mu, V) < \bar{J}$ for each $V > 0$. Hence $\arg\max_V J(\mu, V) = 0$. This means that $\bar{J}$ is uniquely obtained by always replacing A. However, this implies that A puts no effort, which is a contradiction. Hence $V(\mu_H) > 1 - \delta$. Moreover, since $\bar{J} = \max_V J(\mu_H, V)$, we have $J(\mu_H, V) = \bar{J}$ for $V \in [0, V(\mu_H)]$.

For $\mu > \mu_H$, by Lemma 1, we have $J(\mu, 1 - \delta) > J(\mu_H, 1 - \delta) \geq \bar{J}$, which contradicts to (36). Hence we have $V(\mu) > 1 - \delta$ as well.

The Slope of the Linear Part.

Finally, since $J(\mu, V)$ is strictly increasing in $\mu \in (0, 1)$ and $J(\mu, 0) = \bar{J}$ for each μ, "$J(\mu_H, V) = \bar{J}$ for $V \in [0, V(\mu_H)]$" implies the slope of the linear part is negative for $\mu < \mu_H$ and positive for $\mu > \mu_H$.

Property of $V \in \arg\max_{\hat{V}} J\left(\mu, \hat{V}\right)$ Without loss, we can take $V \in \arg\max_{\hat{V}} J\left(\mu, \hat{V}\right)$ such that V is the extreme point of the graph $\left\{\hat{V}, J\left(\mu, \hat{V}\right)\right\}_{\hat{V}}$. This means that there is no mixture to implement V. Hence the social welfare $J(\mu, V)$ at $V \in \arg\max_{\hat{V}} J\left(\mu, \hat{V}\right)$, denoted by $J(\mu)$, is determined by the dynamic programming without mixture:

$$\begin{aligned} J(\mu) &= \max_{(\psi,e,\iota,V')} \{(1-\delta)\, u^P(\mu,\psi,e,\iota) \\ &\quad +\delta \sum_o \Pr(o|\mu,\psi,e,\iota)\, J(\mu'(\mu,\psi,e,\iota,o), V'(o))\} \end{aligned}$$

subject to incentive compatibility constraint:

$$e \in \arg\max \left\{ (1-\delta)\,[1-c_H(e)] + \delta \sum_o \Pr(o|\psi,e,\iota)\, V'(o) \right\}.$$

Note that we do not impose the promise keeping constraint since we are free to choose V to maximize $J\left(\mu,\hat{V}\right)$. Moreover, since the first order condition for e is always necessary and sufficient by the assumption of the cost function c, we can see the above dynamic programming as deciding $(V'(o))_o$, and then e is determined by the first order condition.

In this problem, we first show that $V'(o) \leq \arg\max_{\hat{V}} J\left(\mu'(\mu,\psi,e,\iota,o), \hat{V}\right)$ after $\mu'(\mu,\psi,e,\iota,o) \leq \mu$. Suppose otherwise: There exists $\bar{o}$ such that $V'(\bar{o}) > \arg\max_{\hat{V}} J\left(\mu'(\mu,\psi,e,\iota,\bar{o}), \hat{V}\right)$ after $\mu'(\mu,\psi,e,\iota,\bar{o}) \leq \mu$.

Since

$$\begin{aligned} \mu'(\mu,\psi,e,\iota,\bar{o}) &= \frac{\mu \Pr(\bar{o}|\psi,e,\iota)}{\mu \Pr(\bar{o}|\psi,e,\iota) + (1-\mu)\Pr(\bar{o}|\psi,e=0,\iota)} \\ &= \frac{\mu}{\mu + (1-\mu)\frac{\Pr(\bar{o}|\psi,e=0,\iota)}{\Pr(\bar{o}|\psi,e,\iota)}} \leq \mu, \end{aligned} \tag{37}$$

we have $\Pr(\bar{o}|\psi,0,\iota) \geq \Pr(\bar{o}|\psi,e,\iota)$. Fixing ι, since $\Pr(o|\psi,e,\iota)$ is monotone in e for each o and ψ, the probability $\Pr(\bar{o}|\psi,e,\iota)$ is decreasing in e.

Then, the first order condition is

$$\begin{aligned}
0 &= \frac{d}{dV'(\bar{o})}\{(1-\delta)[1-c_H(e)] \\
&\quad +\delta\sum_o \Pr(o|\mu,\psi,e,\iota)\, J(\mu'(\mu,\psi,e,\iota,o), V'(o))\} \\
&= \{(1-\delta)[-c_H'(e)] \\
&\quad +\delta\sum_o [\frac{d}{de}\Pr(o|\mu,\psi,e,\iota)] J(\mu'(\mu,\psi,e,\iota,o), V'(o)) \\
&\quad +\delta\sum_o \Pr(o|\mu,\psi,e,\iota)\, J_1(\mu'(\mu,\psi,e,\iota,o), V'(o))\,[\frac{d}{de}\mu'(\mu,\psi,e,\iota,o)]\}\frac{de}{dV'(\bar{o})} \\
&\quad +\delta\Pr(\bar{o}|\mu,\psi,e,\iota)\, J_2(\mu'(\mu,\psi,e,\iota,\bar{o}), V'(\bar{o})).
\end{aligned}$$

In general, J_n denote the derivative of J with respect to the nth element. Since $\Pr(\bar{o}|\psi,e,\iota)$ is decreasing in e, we have $\frac{de}{dV'(\bar{o})} < 0$. Moreover $J_2(\mu'(\mu,\psi,e,\iota,\bar{o}), V'(\bar{o})) < 0$ since $V'(\bar{o}) > \arg\max_{\hat{V}} J\left(\mu'(\mu,\psi,e,\iota,\bar{o}), \hat{V}\right)$ and J is concave. Hence, we have

$$\{(1-\delta)[-c_H'(e)] + \delta\sum_o [\frac{d}{de}\Pr(o|\mu,\psi,e,\iota)] J(\mu'(\mu,\psi,e,\iota,o), V'(o)) \tag{38}$$

$$+\delta\sum_o \Pr(o|\mu,\psi,e,\iota)\, J_1(\mu'(\mu,\psi,e,\iota,o), V'(o))\,[\frac{d}{de}\mu'(\mu,\psi,e,\iota,o)]\} < 0. \tag{39}$$

Similarly, if there exists $\hat{o}$ such that $\Pr(\hat{o}|\psi,e,\iota)$ is decreasing in e but

$$V'(\hat{o}) \leq \arg\max_{\hat{V}} J\left(\mu'(\mu,\psi,e,\iota,\hat{o}), \hat{V}\right),$$

then the symmetric argument implies that

$$\begin{aligned}
&\{(1-\delta)[-c_H'(e)] + \delta\sum_o \left[\frac{d}{de}\Pr(o|\mu,\psi,e,\iota)\right] J(\mu'(\mu,\psi,e,\iota,o), V'(o)) \\
&+\delta\sum_o \Pr(o|\mu,\psi,e,\iota)\, J_1(\mu'(\mu,\psi,e,\iota,o), V'(o))\,[\frac{d}{de}\mu'(\mu,\psi,e,\iota,o)]\} \geq 0,
\end{aligned}$$

which is a contradiction.

Therefore, letting O_- be the set of outcomes o such that $\Pr(o|\psi, e, \iota)$ is decreasing in e, we have that, for each $o \in O_-$, we have $V'(o) > \arg\max_{\hat{V}} J\left(\mu'(\mu, \psi, e, \iota, o), \hat{V}\right)$. Symmetrically, letting O_+ be the set of o such that $\Pr(o|\psi, e, \iota)$ is increasing in e, we have that, for each $o \in O_+$, we have $V'(o) < \arg\max_{\hat{V}} J\left(\mu'(\mu, \psi, e, \iota, o), \hat{V}\right)$.

Now we set $V^*(o) = \arg\max_{\hat{V}} J\left(\mu'(\mu, \psi, e, \iota, o), \hat{V}\right)$ for each o, and let e^* be the new optimal effort (fixing ψ and ι throughout). Since $V^*(o) < V'(o)$ for $o \in O_-$ and $V^*(o) > V'(o)$ for $o \in O_+$, we have $e^* > e$ (here, e is the original effort). Hence we have

$$(1-\delta) u^P(\mu, \psi, e^*, \iota) > (1-\delta) u^P(\mu, \psi, e, \iota). \tag{40}$$

In addition, we adjust $V^*(o)$ so that the continuation welfare increases with fixed e:

$$\begin{aligned} & \sum_o \Pr(o|\mu, \psi, e, \iota) J(\mu'(\mu, \psi, e, \iota, o), V'(o)) \\ < & \sum_o \Pr(o|\mu, \psi, e, \iota) J(\mu'(\mu, \psi, e, \iota, o), V^*(o)). \end{aligned} \tag{41}$$

Moreover, since $\max_{\hat{V}} J\left(\mu', \hat{V}\right)$ is increasing in μ', we have

$$J(\mu'(\mu, \psi, e, \iota, o), V^*(o)) < J(\mu'(\mu, \psi, e, \iota, \hat{o}), V^*(\hat{o}))$$

for each $o \in O_-$ and $\hat{o} \in O_+$. Since e^* increases the probability of event o if and only if $o \in O_+$, we have

$$\begin{aligned} & \sum_o \Pr(o|\mu, \psi, e, \iota) J(\mu'(\mu, \psi, e, \iota, o), V^*(o)) \\ < & \sum_o \Pr(o|\mu, \psi, e^*, \iota) J(\mu'(\mu, \psi, e, \iota, o), V^*(o)). \end{aligned} \tag{42}$$

Finally, learning (difference between $\mu'(\mu, \psi, e, \iota, o)$ and $\mu'(\mu, \psi, e^*, \iota, o)$) further increases the continuation payoff. To see why, the following claim is useful:

Claim 1 *For $\mu_1 < \mu_2$, for $V^*(\mu_1) \in \arg\max_{\hat{V}} J\left(\mu_1, \hat{V}\right)$ and $V^*(\mu_2) \in \arg\max_{\hat{V}} J\left(\mu_2, \hat{V}\right)$,* *we have*

$$J_1\left(\mu_1, V^*\left(\mu_1\right)\right) \leq J_1\left(\mu_2, V^*\left(\mu_2\right)\right).$$

Proof. Since J is convex, we have

$$\begin{aligned} & J\left(\mu_1, V^*\left(\mu_1\right)\right) + J_1\left(\mu_1, V^*\left(\mu_1\right)\right)\left[\mu_2 - \mu_1\right] \\ \leq \; & J\left(\mu_2, V^*\left(\mu_1\right)\right) \\ \leq \; & J\left(\mu_2, V^*\left(\mu_2\right)\right) \text{ since } V^*\left(\mu_2\right) \text{ is } \arg\max \text{ for } \mu_2. \end{aligned}$$

At the same time, we have

$$\begin{aligned} J\left(\mu_1, V^*\left(\mu_1\right)\right) & \geq J\left(\mu_2, V^*\left(\mu_2\right)\right) - J_1\left(\mu_2, V^*\left(\mu_2\right)\right)\left[\mu_2 - \mu_1\right] \\ & \geq J\left(\mu_1, V^*\left(\mu_1\right)\right) + J_1\left(\mu_1, V^*\left(\mu_1\right)\right)\left[\mu_2 - \mu_1\right] \\ & \quad - J_1\left(\mu_2, V^*\left(\mu_2\right)\right)\left[\mu_2 - \mu_1\right] \text{ from the previous inequality.} \end{aligned}$$

Hence we have

$$0 \geq \left[J_1\left(\mu_1, V^*\left(\mu_1\right)\right) - J_1\left(\mu_2, V^*\left(\mu_2\right)\right)\right]\left[\mu_2 - \mu_1\right],$$

as desired. ■

Given this claim, since $J_1\left(\mu'\left(\mu, \psi, e, \iota, o\right), V^*\left(o\right)\right)$ is larger for o with $\mu'\left(\mu, \psi, e, \iota, o\right) > \mu$ than for o with $\mu'\left(\mu, \psi, e, \iota, o\right) < \mu$, increasing (decreasing) $\mu'\left(\mu, \psi, e^*, \iota, o\right) > \mu'\left(\mu, \psi, e, \iota, o\right)$ for o with $\mu'\left(\mu, \psi, e, \iota, o\right) > \mu$ (for o with $\mu'\left(\mu, \psi, e, \iota, o\right) < \mu$) increases the continuation payoff since the distribution of $\left\{\mu'\left(\mu, \psi, e^*, \iota, o\right)\right\}_o$ given e^* is the mean-preserving spread of that of $\left\{\mu'\left(\mu, \psi, e, \iota, o\right)\right\}_o$ given e. Together with (41) and (42), we have

$$\begin{aligned} & \sum_o \Pr\left(o \mid \mu, \psi, e, \iota\right) J\left(\mu'\left(\mu, \psi, e, \iota, o\right), V'\left(o\right)\right) \\ < \; & \sum_o \Pr\left(o \mid \mu, \psi, e^*, \iota\right) J\left(\mu'\left(\mu, \psi, e^*, \iota, o\right), V^*\left(o\right)\right). \end{aligned}$$

Together with (40), we have proven that the social welfare is increased.

The proof for $V'(o) \geq \arg\max_{\hat{V}} J\left(\mu'(\mu,\psi,e,\iota,o), \hat{V}\right)$ after $\mu'(\mu,\psi,e,\iota,o) \geq \mu$ is completely symmetric, and so is omitted.

A.3.6 Proof of Lemma 4

With Proposition 4 and Assumption 4: if $\psi_z = 1$, then once $s = B$ happens, we have $\iota_z = \iota^B$; on the other hand, if $s = G$, then we have $C = 0$ (implying $y = G$) for sure by Assumption 4. On the other hand, if $\psi_z = 0$, then the outcome is either $C = 0$ (implying $y = G$) or $C = C_2$ (implying $y = B$). Hence, we can write $o_z \in \{0, C_1, C_2\}$, suppressing the information of ψ_z, where $o_z = C_1$ means that $\psi = 1$ and $s = B$, $o_z = 0$ means either "$\psi = 1$ and $y = s = G$" or "$\psi = 0$ and $y = G$," and $o_z = C_2$ means that $\psi = 0$ and $y = B$.

A.3.7 Proof of Lemma 5

1. With separated institutions:

From (14), A's choice of effort satisfies

$$c'_H(e_z) = \delta q'(e_z)\left[V'_z(0|\psi = 0) - V'_z(C_2|\psi = 0)\right].$$

It then follows that

$$V'_z(0|\psi = 0) - V'_z(C_2|\psi = 0) = \frac{1}{\delta}\frac{c'_H(e_z)}{q'(e_z)} > 0,$$

Given D's strategy,(13) becomes:

$$V_z = [u - c(e_z)] + \delta q(e_z)\left[V'_z(0|\psi = 0) - V'_z(C_2|\psi = 0)\right] + \delta V'_z(C_2|\psi = 0).$$

Then, given (16), $V'_z(C_2|\psi = 0)$ is a function of both e_z and V_z,

$$\delta V'_z(C_2|\psi = 0) = V_z - \left[u - c(e_z) + q(e_z)\frac{c'_H(e_z)}{q'(e_z)}\right],$$

so

$$\frac{\partial V_z'\left(C_2|\psi=0\right)}{\partial e_z}<0.$$

So,

$$\frac{\partial\left[V_z'\left(0|\psi=0\right)-V_z'\left(C_2|\psi=0\right)\right]}{\partial e_z}>0,$$

and therefore

$$\frac{\partial V_z'\left(0|\psi=0\right)}{\partial e_z}>0,$$

2. With the centralized institution:

From (14),

$$c_H'\left(e_z\right)=\delta\left(1-\varepsilon_G\right)q'\left(e_z\right)\left[V_z'\left(0\right)-V_z'\left(C_1\right)\right],$$

so

$$V_z'\left(0|\psi=1\right)-V_z'\left(C_1|\psi=1\right)=\frac{1}{\delta}\frac{c_H'\left(e_z\right)}{\left(1-\varepsilon\right)q'\left(e_z\right)}.$$

From Proposition 4, D's strategy implies that (13) is

$$V_z=\left[u-c\left(e_z\right)\right]+\delta\left(1-\varepsilon_G\right)q\left(e_z\right)\left[V_z'\left(0\right)-V_z'\left(C_1\right)\right]+\delta V_z'\left(C_1\right),$$

so

$$\delta V_z'\left(C_1|\psi=1\right)=V_z-\left[u-c\left(e_z\right)+q\left(e_z\right)\frac{c_H'\left(e_z\right)}{q'\left(e_z\right)}\right].$$

Then,

$$\frac{\partial V_z'\left(C_1|\psi=1\right)}{\partial e_z}<0,$$

and

$$\begin{aligned}\frac{\partial\left[V_z'\left(0|\psi=1\right)-V_z'\left(C_1|\psi=1\right)\right]}{\partial e_z} &> 0,\\ \frac{\partial V_z'\left(0|\psi=1\right)}{\partial e_z} &> 0.\end{aligned}$$

A.3.8 Proof of Corollary 3

Consider the difference in belief updates after the worst outcome for each institution:

$$\Delta(\mu) = \frac{\mu\left[1 - q\left(e\right)\right]}{1 - \left[\mu q\left(e\right) + \left(1 - \mu\right) q\left(0\right)\right]} - \frac{\mu\left[1 - \left(1 - \varepsilon_G\right) q\left(e\right)\right]}{1 - \left(1 - \varepsilon_G\right)\left[\mu q\left(e\right) + \left(1 - \mu\right) q\left(0\right)\right]}.$$

Taking the second derivative with respect to μ:

$$\frac{\partial^2 \Delta(\mu)}{\partial \mu^2} < 0,$$

showing that Δ is a concave function of μ. At $\mu = 0$, $\Delta = 0$, and at $\mu = 1$, $\Delta = 0$, so $\Delta > 0$ $\forall \mu \in (0, 1)$. This also implies that $\exists \mu^* \in (0, 1)$ such that

$$\mu^* = \arg\max \Delta(\mu).$$

A.3.9 Proof of Lemma 7

Suppose $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$ is not binding. Then the following relaxed problem is equivalent to the original problem: maximizing the ex ante social welfare subject to $\mathcal{C}_H \succcurlyeq_H \mathcal{C}_L$.

Further relax the problem so that the principal maximizes the ex ante social welfare. Then the solution is to replace the regulator immediately when the regulator declares the low type; and to solve a usual principal agent problem given $\theta = H$ when he declares the high type. In particular, after declaring H, the principal keeps the high type regulator after a good event with a positive probability to incentivize him to put a positive effort. Since the high type regulator receives the lowest value 0 by declaring that he is a low type, his incentive compatibility constraint $\mathcal{C}_H \succcurlyeq_H \mathcal{C}_L$ is satisfied. Hence this is also the solution for maximizing the ex ante social welfare subject to $\mathcal{C}_H \succcurlyeq_H \mathcal{C}_L$, and so is the solution for the original problem.

However this contract does not satisfy the incentive compatibility constraint for the low type $\mathcal{C}_L \succcurlyeq_L \mathcal{C}_H$ (L can declare H and obtain a positive utility), which is a contradiction.

A.3.10 Proof of Lemma 8

Since the problem is a regular dynamic programming with discounting, it suffices to show that, with the guess

$$J_L(V) = \bar{J} - \left(\bar{J} - \underline{SW}\right) V,$$

the solution for the problem is actually equal to this guess.

Since A of type L does not work, the principal's instantaneous welfare is

$$\begin{aligned} u^P(\theta = L, \psi_z, \iota_z, e_z = 0) &= \max\{-(1 - q(0)) C_2, -(1 - (1 - \varepsilon) q(0)) C_1\} \\ &= \underline{SW}. \end{aligned}$$

Moreover, since our guess of $J_L(V)$ is linear, we have

$$\begin{aligned} \sum_o \Pr(o|\psi_z, e_z = 0) J_L(V_z'(o)) &= J_L(\mathbb{E}[V_z'(o)|\psi_z, e_z = 0]) \\ &= \bar{J} - \left(\bar{J} - \underline{SW}\right) \mathbb{E}[V_z'(o)|\psi_z, e_z = 0]. \end{aligned}$$

Defining $\mathbb{E}[V_z'(o)|\psi_z, e_z = 0] \equiv EV_z'$, the problem becomes

$$\max_{(p_z, EV_z')_z} \int_z \left\{p_z \bar{J} + (1 - p_z)\left[(1 - \delta)\underline{SW} + \delta\left(\bar{J} - \left(\bar{J} - \underline{SW}\right) EV_z'\right)\right]\right\} dz,$$

subject to promise keeping constraint:

$$V = \int_z (1 - p_z)\left[(1 - \delta) + \delta EV_z'\right] dz.$$

Substituting the constraint, we have

$$\begin{aligned} J_L(V) &= \max_{(p_z)_z} \left(\int_z p_z dz\right) \bar{J} + \left(\int_z (1 - p_z) dz\right)\left[(1 - \delta)\underline{SW} + \delta \bar{J}\right] \\ &\quad + \left[\left(\int_z (1 - p_z) dz\right)(1 - \delta) - V\right]\left(\bar{J} - \underline{SW}\right). \end{aligned}$$

Letting $\left(\int_z p_z dz\right) = \beta$, we have

$$\begin{aligned} J_L(V) &= \max_{\beta} \beta \bar{J} + (1-\beta)\left[(1-\delta)\underline{SW} + \delta \bar{J}\right] + \left[(1-\beta)(1-\delta) - V\right]\left(\bar{J} - \underline{SW}\right) \\ &= \bar{J} - \left(\bar{J} - \underline{SW}\right) V, \end{aligned}$$

as desired.

A.3.11 Proof of Lemma 9

Given $\mathcal{C}_L^*(\mathcal{C}_H)$, the optimal strategy for A of type H is not to work since the outcome of the initial mixture is either being replaced or guaranteed value of 1. Hence the payoff from $\mathcal{C}_L^*(\mathcal{C}_H)$ for type H is the same as the payoff from $\mathcal{C}_L^*(\mathcal{C}_H)$ for type L, which is by construction equal to $V_{LH}(\mathcal{C}_H)$.

Since the high type has the lower cost of effort, from $\mathcal{C}_H$, the high type obtains a weakly higher payoff than the low type. Hence his payoff from $\mathcal{C}_H$ is no less than $V_{LH}(\mathcal{C}_H)$. Hence we have $\mathcal{C}_H \succcurlyeq_L \mathcal{C}_L^*(\mathcal{C}_H)$.

A.3.12 Proof of Proposition 6

With ex ante probability μ_H, the agent is of type H, and he brings the ex-ante welfare

$$\sum_{\tau=1}^{t} (1-\delta)\delta^{\tau-1}\mathbb{E}\left[-C_\tau | H\right] + \delta^t \sum_{h^t} \Pr\left(h^t | H\right) J_H\left(h^t\right)$$

to the principal. Here, $\mathbb{E}\left[-C_\tau|H\right]$ is the expected cost in period τ given type H, and so $\sum_{\tau=1}^{t}(1-\delta)\delta^{\tau-1}\mathbb{E}\left[-C_\tau|H\right]$ is the instantaneous welfare for the principal up to period t. On the other hand, with ex ante probability $1-\mu_H$, the agent is of type L, and when he obtains the ex ante payoff of

$$V_{LH} = \sum_{\tau=1}^{t}(1-\delta)\delta^{\tau-1}\left[u - \mathbb{E}\left[c_L(e_L)|L\right]\right] + \delta^t \sum_{h^t}\Pr\left(h^t|L\right) V_{LH}\left(h^t\right)$$

from the contract $\mathcal{C}_H^*$, the ex ante welfare from the L-type is $J_L\left(V_{LH}\right)$. Here $\mathbb{E}\left[c_L\left(e_L\right)|L\right]$ is the expected cost of effort (in our special case in which L-type does not work, $\mathbb{E}\left[c_L\left(e_L\right)|L\right] = 0$), and so $\sum_{\tau=1}^{t}\left(1-\delta\right)\delta^{\tau-1}\left[u-\mathbb{E}\left[c_L\left(e_L\right)|L\right]\right]$ is the instantaneous payoff of L type up to t.

In total, P's ex ante utility is

$$\begin{aligned}&\mu_H\left[\sum_{\tau=1}^{t}\left(1-\delta\right)\delta^{\tau-1}\mathbb{E}\left[-C_\tau|H\right]+\delta^t\sum_{h^t}\Pr\left(h^t|H\right)J_H\left(h^t\right)\right]\\&+\left(1-\mu_H\right)J_L\left(\begin{array}{c}\sum_{\tau=1}^{t}\left(1-\delta\right)\delta^{\tau-1}\left[u-\mathbb{E}\left[c_L\left(e_L\right)|L\right]\right]\\+\delta^t\sum_{h^t}\Pr\left(h^t|L\right)V_{LH}\left(h^t\right)\end{array}\right).\end{aligned}\tag{43}$$

Since J_L is linear, maximizing the social welfare is equivalent to maximizing

$$\mu_H\delta^t\Pr\left(h^t|H\right)J_H\left(h^t\right)+\left(1-\mu_H\right)\delta^t\Pr\left(h^t|L\right)\left[\bar{J}-\left(\bar{J}-\underline{SW}\right)V_{LH}\left(h^t\right)\right]\tag{44}$$

for each h^t. Dividing both sides by

$$\mu_H\delta^t\Pr\left(h^t|H\right)+\left(1-\mu_H\right)\delta^t\Pr\left(h^t|L\right)\ \ \text{(a constant at } h^t\text{)},$$

it is equivalent to maximizing

$$\mu\left(h^t\right)J_H\left(h^t\right)+\left(1-\mu\left(h^t\right)\right)\left[\bar{J}-\left(\bar{J}-\underline{SW}\right)V_{LH}\left(h^t\right)\right]\tag{45}$$

with

$$\mu\left(h^t\right)=\frac{\mu_H\Pr\left(h^t|H\right)}{\mu_H\Pr\left(h^t|H\right)+\left(1-\mu_H\right)\Pr\left(h^t|L\right)}.$$

Using the same notation as in the best PBE without communciation, we can express the first three terms of (45) recursively. First, given the current belief μ, the targeted effort e_z, and the public outcome o_z, the next belief $\mu_z'\left(\mu,e_z,o_z\right)$ is determined by

$$\mu_z'\left(\mu,e_z,o_z\right)=\frac{\mu\Pr\left(o_z|\psi_z,e_z\right)}{\mu\Pr\left(o_z|\psi_z,e_z\right)+\left(1-\mu\right)\Pr\left(o_z|\psi_z,e_z\right)}.$$

Second, P's welfare from the high type is

$$\begin{aligned} J_H\left(\mu'_z\left(\mu, e_z, o_z\right), V\right) &= \int_z [p_z \bar{J} + \left(1-p_z\right)\left\{-\left(1-\delta\right)\mathbb{E}\left[C|\psi_z, \iota_z, e_z\right]\right. \\ &\quad \left. +\delta \sum_{o_z} \Pr\left(o_z|\psi_z, \iota_z, e_z\right) J_H\left(\mu'_z\left(\mu, e_z, o_z\right), V'_z\left(o_z\right)\right)\right\}]dz. \end{aligned} \tag{46}$$

Third, the continuation payoff that the L-type obtains if he takes the contract $\mathcal{C}_H$ and reaches history (μ, V) is

$$\begin{aligned} & V_{LH}\left(\mu'_z\left(\mu, e_z, o_z\right), V\right) \\ = & \int_z \left(1-p_z\right)\left\{\left(1-\delta\right) + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, \iota_z, e=0\right) V_{LH}\left(\mu'_z\left(\mu, e_z, o_z\right), V'_z\left(o_z\right)\right)\right\}dz. \end{aligned} \tag{47}$$

Given (46) and (47), we have

$$\begin{aligned} & \mu J_H\left(\mu, V\right) + \left(1-\mu\right)\left[\bar{J} - \left(\bar{J} - \underline{SW}\right) V_{LH}\left(\mu, V\right)\right] \\ = & \int_z [p_z \bar{J} + \left(1-p_z\right)\{\left(1-\delta\right)\left[\mu \mathbb{E}\left[-C|\psi_z, e_z, \iota_z\right] + \left(1-\mu\right)\underline{SW}\right] \\ & +\delta \sum_{o_z}\{[\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + \left(1-\mu\right)\Pr\left(o_z|\psi_z, e=0, \iota_z\right)] \\ & \times[\frac{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right)}{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + \left(1-\mu\right)\Pr\left(o_z|\psi_z, e_z=0, \iota_z\right)} \\ & \times J_H\left(\mu'_z\left(\mu, e_z, o_z\right), V'_z\left(o_z\right)\right) && (48) \\ & + \frac{\left(1-\mu\right)\Pr\left(o_z|\psi_z, e_z=0, \iota_z\right)}{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + \left(1-\mu\right)\Pr\left(o_z|\psi_z, e_z=0, \iota_z\right)} && (49) \\ & \times[\bar{J} - \left(\bar{J} - \underline{SW}\right) V_{LH}\left(\mu'_z\left(\mu, e_z, o_z\right), V'_z\left(o_z\right)\right)]]\}\}]dz && (50) \end{aligned}$$

Let $J\left(\mu, V\right)$ denote the maximized value of (49):

$$J\left(\mu, V\right) = \mu J\left(\mu, V\right) + \left(1-\mu\right)\left[\bar{J} - \left(\bar{J} - \underline{SW}\right) V_{LH}\left(\mu, V\right)\right]. \tag{51}$$

Given this notation, since

$$\begin{aligned}
\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + (1-\mu) \Pr\left(o_z|\psi_z, e_z = 0, \iota_z\right) &= \Pr\left(o_z|\mu, e_z, \iota_z\right); \\
\frac{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right)}{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + (1-\mu) \Pr\left(o_z|\psi_z, e_z = 0, \iota_z\right)} &= \mu'\left(\mu, \psi_z, e_z, \iota_z, o_z\right); \\
\frac{(1-\mu) \Pr\left(o_z|\psi_z, e_z = 0, \iota_z\right)}{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + (1-\mu) \Pr\left(o_z|\psi_z, e_z = 0, \iota_z\right)} &= 1 - \mu'\left(\mu, \psi_z, e_z, \iota_z, o_z\right),
\end{aligned}$$

(49) becomes

$$\begin{aligned}
&\int_z [p_z \bar{J} + (1-p_z) \{(1-\delta) [\mu \mathbb{E}\left[-C|\psi_z, e_z, \iota_z\right] + (1-\mu) \underline{SW}] \\
&+ \delta \sum_{o_z} \Pr\left(o_z|\mu, \psi_z, e_z, \iota_z\right) J\left(\mu'\left(\mu, \psi_z, e_z, \iota_z\right), V_z'\left(o_z\right)\right)\}] dz.
\end{aligned}$$

Hence the dynamic programming boils down to

$$\begin{aligned}
J(\mu, V) = & \max_{(p_z, \iota_z, e_z, W_z)_z} \int_z [p_z \bar{J} + (1-p_z) \{(1-\delta) [\mu \mathbb{E}\left[-C|\psi_z, e_z, \iota_z\right] + (1-\mu) \underline{SW}] \\
& + \delta \sum_{o_z} \Pr\left(o_z|\mu, \psi_z, e_z, \iota_z\right) J\left(\mu'\left(\mu, \psi_z, e_z, \iota_z\right), V_z'\left(o_z\right)\right)\}] dz, \qquad (52)
\end{aligned}$$

with

$$\mu'\left(\mu, \psi_z, e_z, \iota_z, o_z\right) = \frac{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right)}{\mu \Pr\left(o_z|\psi_z, e_z, \iota_z\right) + (1-\mu) \Pr\left(o_z|\psi_z, e_z = 0, \iota_z\right)}. \qquad (53)$$

In addition, the principal is subject to the following two constraints:

1. Promise keeping constraint:

$$V = \int_z (1-p_z) \left\{ (1-\delta)\left(1 - c_H\left(e_z\right)\right) + \delta \sum_{o_z} \Pr\left(o|\psi_z, e_z, \iota_z\right) V_z'\left(o_z\right) \right\} dz; \qquad (54)$$

2. Incentive compatibility constraint: For each z with $p_z > 0$, we have

$$e_z \in \arg\max \left\{ (1-\delta)\left(1 - c_H\left(e_z\right)\right) + \delta \sum_{o_z} \Pr\left(o|\psi_z, e_z, \iota_z\right) V_z'\left(o_z\right) \right\}. \qquad (55)$$

Finally, $\bar{J}$ is the fixed point such that

$$\begin{aligned}\bar{J} &= \max_{V}\left\{(1-\mu_H)\, J_L\left(V_{LH}\left(\mu_H, V\right)\right) + \mu_H J_H\left(\mu_H, V\right)\right\} \\ &= \max_{V} J\left(\mu_H, V\right) \text{ by definition of (51).} \end{aligned} \tag{56}$$

A.3.13 Proof of Lemma 10

After any history, the continuation social welfare from the dynamic mechanism design problem is no less than $\underline{SW}$. To see why, after the agent declares that he is L-type, the worst outcome is to keep him forever, which gives the welfare of $\underline{SW}$. After the agent declares that he is H-type, the "welfare" $J(\mu, V)$ is the convex combination of the welfare from H-type and that from L-type by (44). By Lemma 1, for each μ and $V \in [0,1]$, we have $J(\mu, V) \geq \underline{SW}$. Since the welfare from L-type is $\underline{SW}$, the welfare from H-type is no less than $\underline{SW}$.

Therefore, by the threat of switching to the no effort equilibrium, the principal has an incentive to follow the solution of the dynamic mechanism design problem (note that the randomization device z makes the mixed strategy observable).

A.3.14 Proof of Proposition 7

Part 1. Effort strategies

The proof is the same as the Part 1 of the proof to Proposition 1.

Part 2. Beliefs

The proof is the same as the Part 3 of the proof to Proposition 1.

Part 3. Intervention strategy

Given the ordering of beliefs $b_0 > \frac{\mu_H}{\mu_L} > b_{C_1}, b_{C_2}$, it is a weakly dominant strategy for A to send message $m = G$ if $s = G$. Then, $m = B$ is sent only after $s = B$. Given Assumption 2, the expected cost for D is higher without intervention than with intervention after $s = B$. Therefore, intervention is chosen in equilibrium whenever $m = B$.

A.3.15 Proof of Proposition 8

(1) **The full revelation equilibrium**: A with $t = L$ takes e_L^* and sends $m = B$ after $s = B$; A with $t = H$ takes e_H^* and sends $m = B$ after $s = B$; and A always sends $m = G$ after $s = G$.

After $m = G$, D believes that $s = G$ for sure, and so no intervention happens, since the expected cost of an intervention would be higher than the expected cost of not intervening:

$$C_2 \frac{(1 - q(e))\,\varepsilon_B}{q(e)\,(1 - \varepsilon_G) + (1 - q(e))\,\varepsilon_B} < C_1,$$

$\forall e$, by Assumption 2.

After $m = B$, since D believes that $s = B$, she intervenes since

$$C_2 \frac{(1 - q(e))\,(1 - \varepsilon_B)}{q(e)\varepsilon_G + (1 - q(e))\,(1 - \varepsilon_B)} > C_1,$$

$\forall e$, by Assumption 2.

Given this intervention and message strategy, the equilibrium belief is determined as follows:

$$\begin{aligned}
\mathbb{E}\left[t|y = G\right] &= \frac{\mu \Pr\left(y = G, s = G|e_H^*\right)}{\mu \Pr\left(y = G|s = G, e_H^*\right) + (1 - \mu) \Pr\left(s = G, y = G|e_L^*\right)} \\
&= \frac{\mu q\left(e_H^*\right)(1 - \varepsilon_G)}{\mu q\left(e_H^*\right)(1 - \varepsilon_G) + (1 - \mu)\, q\left(e_L^*\right)(1 - \varepsilon_G)};
\end{aligned}$$

$$\mathbb{E}\left[t|y = B\right] = \frac{\mu\left(1 - q\left(e_H^*\right)\right)\varepsilon_B}{\mu\left(1 - q\left(e_H^*\right)\right)\varepsilon_B + (1 - \mu)\left(1 - q\left(e_L^*\right)\right)\varepsilon_B};$$

and

$$\begin{aligned}
\mathbb{E}\left[t|C_1\right] &= \frac{\mu \Pr\left(s = B|e_H^*\right)}{\mu \Pr\left(s = B|e_H^*\right) + (1 - \mu) \Pr\left(s = B|e_L^*\right)} \\
&= \frac{\mu\left[\left(1 - q\left(e_H^*\right)\right)(1 - \varepsilon_B) + q\left(e_H^*\right)\varepsilon_G\right]}{\left\{\begin{array}{c} \mu\left[\left(1 - q\left(e_H^*\right)\right)(1 - \varepsilon_B) + q\left(e_H^*\right)\varepsilon_G\right] \\ + (1 - \mu)\left[\left(1 - q\left(e_L^*\right)\right)(1 - \varepsilon_B) + q\left(e_L^*\right)\varepsilon_G\right] \end{array}\right\}}.
\end{aligned}$$

So

$$\begin{aligned}
\frac{\mathbb{E}\left[H|y=G\right]}{\mathbb{E}\left[L|y=G\right]} &= \frac{\mu q\left(e_H^*\right)}{\left(1-\mu\right)q\left(e_L^*\right)}, \\
\frac{\mathbb{E}\left[H|y=B\right]}{\mathbb{E}\left[L|y=B\right]} &= \frac{\mu\left(1-q\left(e_H^*\right)\right)}{\left(1-\mu\right)\left(1-q\left(e_L^*\right)\right)}, \\
\frac{\mathbb{E}\left[H|C_1\right]}{\mathbb{E}\left[L|C_1\right]} &= \frac{\mu\left[\left(1-q\left(e_H^*\right)\right)\left(1-\varepsilon_B\right)+q\left(e_H^*\right)\varepsilon_G\right]}{\left(1-\mu\right)\left[\left(1-q\left(e_L^*\right)\right)\left(1-\varepsilon_B\right)+q\left(e_L^*\right)\varepsilon_G\right]}.
\end{aligned}$$

Given the result from Part 2,

$$\frac{\mathbb{E}\left[H|y=G\right]}{\mathbb{E}\left[L|y=G\right]} \geq \frac{\mathbb{E}\left[H|C_1\right]}{\mathbb{E}\left[L|C_1\right]} \geq \frac{\mathbb{E}\left[H|y=B\right]}{\mathbb{E}\left[L|y=B\right]}, \tag{57}$$

and

$$\frac{\mathbb{E}\left[H|C_1\right]}{\mathbb{E}\left[L|C_1\right]} = \varrho\frac{\mathbb{E}\left[H|y=G\right]}{\mathbb{E}\left[L|y=G\right]} + \left(1-\varrho\right)\frac{\mathbb{E}\left[H|y=B\right]}{\mathbb{E}\left[L|y=B\right]},$$

where

$$\varrho = \frac{q\left(e_L^*\right)\varepsilon_G}{q\left(e_L^*\right)\varepsilon_G + \left(1-q\left(e_L^*\right)\right)\left(1-\varepsilon_B\right)}. \tag{58}$$

After signal $s = B$, the relative gain for A of type i from deviating to message $m = G$ is:

$$\begin{aligned}
&\Bigg\{u^A\left(b_0\right)\frac{q\left(e_i^*\right)\varepsilon_G}{q\left(e_i^*\right)\varepsilon_G+\left(1-q\left(e_i^*\right)\right)\left(1-\varepsilon_B\right)} \\
&+ u^A\left(b_{C_2}\right)\frac{\left(1-q\left(e_i^*\right)\right)\left(1-\varepsilon_B\right)}{q\left(e_i^*\right)\varepsilon_G+\left(1-q\left(e_i^*\right)\right)\left(1-\varepsilon_B\right)} - u\left(b_{C_1}\right)\Bigg\}.
\end{aligned} \tag{59}$$

After signal $s = G$, the relative gain for A of type i from deviating to message $m = B$ is:

$$\begin{aligned}
&\Bigg\{u^A\left(b_{C_1}\right) - u^A\left(b_0\right)\frac{q\left(e_i^*\right)\left(1-\varepsilon_G\right)}{q\left(e_i^*\right)\left(1-\varepsilon_G\right)+\left(1-q\left(e_i^*\right)\right)\varepsilon_B} \\
&- u^A\left(b_{C_2}\right)\frac{\left(1-q\left(e_i^*\right)\right)\varepsilon_B}{q\left(e_i^*\right)\left(1-\varepsilon_G\right)+\left(1-q\left(e_i^*\right)\right)\varepsilon_B}\Bigg\}.
\end{aligned} \tag{60}$$

The condition for this to be an equilibrium is that (59) and (60) are both negative.

From $u^A(b)$ concave and (58), it follows that

$$u^A(b_{C_1}) \geq u^A(b_0)\varrho + u^A(b_{C_2})(1-\varrho), \tag{61}$$

so (59) is negative for A of type L.

Let $\varrho^* \geq \varrho$ denote the value defined implicitly by

$$u^A(b_{C_1}) = \varrho^* u^A(b_0) + (1-\varrho^*)u^A(b_{C_2}). \tag{62}$$

Then, the (59) is negative for the H-type if the following condition holds:

$$\frac{q(e_H^*)\varepsilon_G}{q(e_H^*)\varepsilon_G + (1-q(e_H^*))(1-\varepsilon_B)} \leq \varrho^* \tag{63}$$

Finally, for $\varepsilon_B \to 0$ and $\varepsilon_G \to 0$, we have

$$\begin{aligned} \frac{q(e_i^*)(1-\varepsilon_G)}{q(e_i^*)(1-\varepsilon_G)+(1-q(e_i^*))\varepsilon_B} &\to 1, \\ \frac{(1-q(e_i^*))\varepsilon_B}{q(e_i^*)(1-\varepsilon_G)+(1-q(e_i^*))\varepsilon_B} &\to 0, \end{aligned}$$

so (60) is negative.

Therefore, an equilibrium with full revelation exists whenever (63) is satisfied.

(2) **An equilibrium with partial information withholding:** A of type L puts in effort e_L^* and sends $m = G$ after $s = G$ and message $m = B$ after $s = B$. A of type H puts in effort e_H^* and sends $m = G$ after $s = G$, and sends message $m = G$ with probability γ and message $m = B$ with probability $(1-\gamma)$ after $s = B$, for $\gamma \in (0,1)$.

The equilibrium is characterized by

$$q'(e_t)\left\{ \begin{array}{c} (1-\varepsilon)\left(u^A(b_0(e_H,e_L)) - u^A\left(b_{C_1}^{\gamma}(e_H,e_L)\right)\right) \\ -\varepsilon\left(u^A(b_{C_2}(e_H,e_L)) - u^A\left(b_{C_1}^{\gamma}(e_H,e_L)\right)\right) \end{array} \right\} = c_t'(e_t)$$

with

$$\begin{aligned}
u^A(b^{\gamma}_{C_1}(e_H,e_L)) &= u^A\left(\frac{\mu(1-\gamma)(q(e_H)\varepsilon_G+(1-q(e_H))(1-\varepsilon_B))}{(1-\mu)(q(e_L)\varepsilon_G+(1-q(e_L))(1-\varepsilon_B))}\right)\\
u^A(b^{\gamma}_{0}(e_H,e_L)) &= u^A\left(\frac{\mu q(e_H)((1-\varepsilon_G)+\gamma\varepsilon_G)}{(1-\mu)q(e_L)(1-\varepsilon_G)}\right);\\
u^A(b_{C_2}(e_H,e_L)) &= u^A\left(\frac{\mu(1-q(e_H))(\varepsilon_B+(1-\gamma)(1-\varepsilon_B))}{(1-\mu)(1-q(e_L))\varepsilon_B}\right),
\end{aligned}$$

and type H's indifference

$$\begin{aligned}
u^A\left(b^{\gamma}_{C_1}(e_H,e_L)\right) &= \frac{q(e_H)\varepsilon_G}{q(e_H)\varepsilon_G+(1-q(e_H))(1-\varepsilon_B)}u^A\left(b^{\gamma}_{0}(e_H,e_L)\right)\\
&\quad+\frac{(1-q(e_H))(1-\varepsilon_B)}{q(e_H)\varepsilon_G+(1-q(e_H))(1-\varepsilon_B)}u^A\left(b_{C_2}(e_H,e_L)\right). \qquad (64)
\end{aligned}$$

Therefore, (64) gives the condition for the equilibrium with partial information withholding.

(3) **An equilibrium with full information withholding:** A always sends $m=G$ after every signal.

The equilibrium effort choice is characterized by

$$q'(e_t)\{u(b_0)-u(b_{C_2})\}=c'_t(e_t)$$

with

$$\begin{aligned}
u(b_0) &= \frac{\mu q(e_H)}{(1-\mu)q(e_L)};\\
u(b_{C_2}) &= \frac{\mu(1-q(e_H))}{(1-\mu)(1-q(e_L))}.
\end{aligned}$$

Given (63), the condition for the equilibrium with full information withholding is

$$\frac{q(e^*_H)\varepsilon_G}{q(e^*_H)\varepsilon_G+(1-q(e^*_H))(1-\varepsilon_B)}>\varrho^*,$$

where ϱ^* is defined in (62).

A.3.16 Proof of Proposition 9

The proof is the same as in the proof to Proposition 2.

A.3.17 Proof of Proposition 10

The equilibrium is characterized by

$$q'(e_t)\left\{\begin{array}{c}(1-\varepsilon)\left(u(b^{\varepsilon}_G(e_H,e_L))-u\left(b^{\varepsilon}_{C_1}(e_H,e_L)\right)\right)\\ -\varepsilon\left(u(b_{C_2}(e_H,e_L))-u\left(b^{\varepsilon}_{C_1}(e_H,e_L)\right)\right)\end{array}\right\}=c'_t(e_t)$$

with

$$\begin{aligned}u(b^{\varepsilon}_{C_1}(e_H,e_L)) &= u\left(\frac{\mu_H\left(q(e_H)\varepsilon+(1-q(e_H))(1-\varepsilon)\right)}{\mu_L\left(q(e_L)\varepsilon+(1-q(e_L))(1-\varepsilon)\right)}\right)\\ u(b^{\varepsilon}_G(e_H,e_L)) &= u\left(\frac{\mu_H q(e_H)(1-\varepsilon)}{\mu_L q(e_L)(1-\varepsilon)}\right)=u\left(\frac{\mu_H q(e_H)}{\mu_L q(e_L)}\right);\\ u(b_{C_2}(e_H,e_L)) &= u\left(\frac{\mu_H(1-q(e_H))\varepsilon}{\mu_L(1-q(e_L))\varepsilon}\right)=u\left(\frac{\mu_H(1-q(e_H))}{\mu_L(1-q(e_L))}\right).\end{aligned}$$

We have

$$\frac{c'_H(e_H)}{q'(e_H)}=\frac{c'_L(e_L)}{q'(e_L)}.$$

Solving this, we have

$$e_L=\phi(e_H),$$

where $\phi(x)=\left(\frac{c'_H}{q'}\right)^{-1}\left(\frac{c'_L(x)}{q'(x)}\right)$. Plugging this into the equilibrium condition, we have

$$\begin{aligned}&q'(e_H)\left\{\begin{array}{c}(1-\varepsilon)\left(u(b^{\varepsilon}_G(e_H,\phi(e_H)))-u\left(b^{\varepsilon}_{C_1}(e_H,\phi(e_H))\right)\right)\\ -\varepsilon\left(u(b_{C_2}(e_H,\phi(e_H)))-u\left(b^{\varepsilon}_{C_1}(e_H,\phi(e_H))\right)\right)\end{array}\right\}\\ =\;&c'_H(e_H).\end{aligned}$$

Define

$$MA_{\varepsilon}(e_H) \equiv q'(e_H) \left\{ \begin{array}{c} (1-\varepsilon)\left(u(b^{\varepsilon}_{G}(e_H, \phi(e_H))) - u\left(b^{\varepsilon}_{C_1}(e_H, \phi(e_H))\right)\right) \\ -\varepsilon\left(u(b^{\varepsilon}_{C_2}(e_H, \phi(e_H))) - u\left(b^{\varepsilon}_{C_1}(e_H, \phi(e_H))\right)\right) \end{array} \right\}.$$

If $MA_{\varepsilon}(e_H)$ is concave and $u(\cdot) \geq 0$, these conditions guarantee that there are only two solutions: $e_H = 0$ and $e_H > 0$. Hence we are left to show that $\frac{d}{d\varepsilon} MA_{\varepsilon}(e_H) \leq 0$

We have

$$\begin{aligned} \frac{d}{d\varepsilon} MA_{\varepsilon}(e_H) &= q'(e_H) \left\{ \begin{array}{c} -u(b^{\varepsilon}_{G}(e_H, \phi(e_H))) + u\left(b^{\varepsilon}_{C_1}(e_H, \phi(e_H))\right) \\ +u(b^{\varepsilon}_{C_2}(e_H, \phi(e_H))) - u\left(b^{\varepsilon}_{C_1}(e_H, \phi(e_H))\right) \end{array} \right\} \\ & \quad -q'(e_H)(1-2\varepsilon) u'\left(b^{\varepsilon}_{C_1}\right) \frac{d}{d\varepsilon} \frac{\mu_H (q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon))}{\mu_L (q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon))}. \end{aligned}$$

Since

$$\begin{aligned} & \frac{d}{d\varepsilon} \frac{\mu_H (q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon))}{\mu_L (q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon))} \\ = & \frac{\mu_H}{\mu_L} \frac{\begin{array}{c} [q(e_H) - (1-q(e_H))][q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)] \\ -[q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon)][q(e_L) - (1-q(e_L))] \end{array}}{[q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)]^2} \\ = & \frac{q(e_H) - (1-q(e_H))}{q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)} \frac{\mu_H}{\mu_L} \\ & - \frac{q(e_L) - (1-q(e_L))}{q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)} b^{\varepsilon}_{C_1} \end{aligned}$$

and

$$\frac{q(e_L) - (1-q(e_L))}{q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)} \leq \frac{q(e_H) - (1-q(e_H))}{q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)},$$

we have

$$\begin{aligned} & \frac{d}{d\varepsilon} \frac{\mu_H (q(e_H)\varepsilon + (1-q(e_H))(1-\varepsilon))}{\mu_L (q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon))} \\ \geq & \frac{q(e_H) - (1-q(e_H))}{q(e_L)\varepsilon + (1-q(e_L))(1-\varepsilon)} \left(\frac{\mu_H}{\mu_L} - b^{\varepsilon}_{C_1} \right). \end{aligned}$$

Since $q\left(e_H\right) \geq q\left(e_L\right)$ and $\varepsilon \leq \frac{1}{2}$, we have

$$\begin{aligned}
\frac{\mu_H}{\mu_L} - b^{\varepsilon}_{C_1} &= \frac{\mu_H}{\mu_L}\left(1 - \frac{q\left(e_H\right)\varepsilon + \left(1 - q\left(e_H\right)\right)\left(1 - \varepsilon\right)}{q\left(e_L\right)\varepsilon + \left(1 - q\left(e_L\right)\right)\left(1 - \varepsilon\right)}\right) \\
&= \frac{\mu_H}{\mu_L}\frac{\left(q\left(e_H\right) - q\left(e_L\right)\right)\left(1 - 2\varepsilon\right)}{q\left(e_L\right)\varepsilon + \left(1 - q\left(e_L\right)\right)\left(1 - \varepsilon\right)} \geq 0.
\end{aligned}$$

In total, we have

$$\frac{d}{d\varepsilon}\frac{\mu_H\left(q\left(e_H\right)\varepsilon + \left(1 - q\left(e_H\right)\right)\left(1 - \varepsilon\right)\right)}{\mu_L\left(q\left(e_L\right)\varepsilon + \left(1 - q\left(e_L\right)\right)\left(1 - \varepsilon\right)\right)} \geq 0.$$

Hence,

$$\begin{aligned}
\frac{d}{d\varepsilon}MA_{\varepsilon}\left(e_H\right) &= q'\left(e_H\right)\left\{\begin{array}{c} -u(b^{\varepsilon}_{G}\left(e_H, \phi\left(e_H\right)\right)) + u\left(b^{\varepsilon}_{C_1}\left(e_H, \phi\left(e_H\right)\right)\right) \\ +u(b^{\varepsilon}_{C_2}\left(e_H, \phi\left(e_H\right)\right)) - u\left(b^{\varepsilon}_{C_1}\left(e_H, \phi\left(e_H\right)\right)\right) \end{array}\right\} \\
&\quad -q'\left(e_H\right)\left(1 - 2\varepsilon\right)u'\left(b^{\varepsilon}_{C_1}\right)\frac{d}{d\varepsilon}\frac{\mu_H\left(q\left(e_H\right)\varepsilon + \left(1 - q\left(e_H\right)\right)\left(1 - \varepsilon\right)\right)}{\mu_L\left(q\left(e_L\right)\varepsilon + \left(1 - q\left(e_L\right)\right)\left(1 - \varepsilon\right)\right)} \\
&\leq 0,
\end{aligned}$$

as desired.

A.3.18 Proof of Proposition 11

If the two conditions are satisfied, then the equilibrium with separate institutions is a full revelation equilibrium, as shown in the proof to proposition 8. Then, D has access to the same information in both institutional setups, and so the intervention decision, the beliefs and the effort choices are derived from the same maximization problem.

A.3.19 Proof of Proposition 12

Proof. If, for each fixed e_H,

$$\begin{aligned}
& q'\left(e_H\right)\{(1-\varepsilon)\left(u(b_G^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u\left(b_{C_1}^{sep}\left(e_H,\phi\left(e_H\right)\right)\right)\right) \\
& -\varepsilon\left(u(b_{C_2}^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u\left(b_{C_1}^{sep}\left(e_H,\phi\left(e_H\right)\right)\right)\right)\} \\
\leq\ & q'\left(e_H\right)\left\{u(b_G^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u(b_{C_2}^{sep}\left(e_H,\phi\left(e_H\right)\right))\right\}, \qquad (65)
\end{aligned}$$

then e_H is higher in the separate institutions.

Condition (65) is equivalent to

$$\begin{aligned}
& \{(1-\varepsilon)\left(u(b_G^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u\left(b_{C_1}^{sep}\left(e_H,\phi\left(e_H\right)\right)\right)\right) \\
& -\varepsilon\left(u(b_{C_2}^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u\left(b_{C_1}^{sep}\left(e_H,\phi\left(e_H\right)\right)\right)\right)\} \\
\leq\ & \left\{u(b_G^{sep}\left(e_H,\phi\left(e_H\right)\right)) - u(b_{C_2}^{sep}\left(e_H,\phi\left(e_H\right)\right))\right\}.
\end{aligned}$$

Since q is increasing, ε is no more than $\frac{1}{2}$, and $e_H \geq e_L$ (by the definition of ϕ), this condition is automatically satisfied. ■

Since

$$\begin{aligned}
b_G^{sep}\left(e_H\right) &= \frac{\mu_H}{\mu_L}\frac{q\left(e_H\right)}{q\left(\phi\left(e_H\right)\right)}; \\
b_{C_2}^{sep}\left(e_H\right) &= \frac{\mu_H}{\mu_L}\frac{1-q\left(e_H\right)}{1-q\left(\phi\left(e_H\right)\right)},
\end{aligned}$$

we have

$$\begin{aligned}
\frac{d}{de_H}b_0^{sep}\left(e_H\right) &= \frac{\mu_H}{\mu_L}\frac{q'\left(e_H\right)q\left(\phi\left(e_H\right)\right) - q\left(e_H\right)q'\left(\phi\left(e_H\right)\right)\phi'\left(e_H\right)}{\left[q\left(\phi\left(e_H\right)\right)\right]^2} \\
&= \frac{\mu_H}{\mu_L}\frac{q'\left(e_H\right)}{q\left(\phi\left(e_H\right)\right)} - \frac{q'\left(\phi\left(e_H\right)\right)\phi'\left(e_H\right)}{q\left(\phi\left(e_H\right)\right)}b_0^{sep}\left(e_H\right);
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{(de_H)^2} b_0^{sep}(e_H) &= \frac{\mu_H}{\mu_L} \frac{q''(e_H)\, q(\phi(e_H)) - q'(e_H)\, q'(\phi(e_H))\, \phi'(e_H)}{[q(\phi(e_H))]^2} \\
&\quad - \frac{q''(\phi(e_H))\, \phi'(e_H)\, q(\phi(e_H))}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad - \frac{q'(\phi(e_H))\, \phi''(e_H)\, q(\phi(e_H))}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad + \frac{(q'(\phi(e_H))\, \phi'(e_H))^2}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad - \frac{q'(\phi(e_H))\, \phi'(e_H)}{q(\phi(e_H))} \frac{d}{de_H} b_0^{sep}(e_H).
\end{aligned}$$

Since

$$\begin{aligned}
& -\frac{q'(\phi(e_H))\, \phi'(e_H)}{q(\phi(e_H))} \frac{d}{de_H} b_0^{sep}(e_H) \\
= {} & -\frac{q'(\phi(e_H))\, \phi'(e_H)}{q(\phi(e_H))} \left(\frac{\mu_H}{\mu_L} \frac{q'(e_H)}{q(\phi(e_H))} - \frac{q'(\phi(e_H))\, \phi'(e_H)}{q(\phi(e_H))} b_0^{sep}(e_H) \right) \\
= {} & -\frac{q'(\phi(e_H))\, q'(e_H)\, \phi'(e_H)}{[q(\phi(e_H))]^2} \frac{\mu_H}{\mu_L} + \frac{[q'(\phi(e_H))\, \phi'(e_H)]^2}{[q(\phi(e_H))]^2} b_0^{sep}(e_H),
\end{aligned}$$

we have

$$\begin{aligned}
\frac{d^2}{(de_H)^2} b_0^{sep}(e_H) &= \frac{\mu_H}{\mu_L} \frac{q''(e_H)\, q(\phi(e_H)) - q'(e_H)\, q'(\phi(e_H))\, \phi'(e_H)}{[q(\phi(e_H))]^2} \\
&\quad - \frac{q''(\phi(e_H))\, \phi'(e_H)\, q(\phi(e_H))}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad - \frac{q'(\phi(e_H))\, \phi''(e_H)\, q(\phi(e_H))}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad + \frac{(q'(\phi(e_H))\, \phi'(e_H))^2}{[q(\phi(e_H))]^2} b_0^{sep}(e_H) \\
&\quad - \frac{q'(\phi(e_H))\, q'(e_H)\, \phi'(e_H)}{[q(\phi(e_H))]^2} \frac{\mu_H}{\mu_L} \\
&\quad + \frac{[q'(\phi(e_H))\, \phi'(e_H)]^2}{[q(\phi(e_H))]^2} b_0^{sep}(e_H).
\end{aligned}$$

Simplifying, we have

$$\begin{aligned}\frac{d^2}{(de_H)^2}b_0^{sep}(e_H) &= \frac{q''(e_H)\,q(\phi(e_H))}{[q(\phi(e_H))]^2}\left(\frac{\mu_H}{\mu_L}-\phi'(e_H)\,b_0^{sep}(e_H)\right)\\ &\quad -\frac{q'(\phi(e_H))\,\phi'(e_H)}{[q(\phi(e_H))]^2}\left(q'(e_H)\frac{\mu_H}{\mu_L}-q'(\phi(e_H))\,\phi'(e_H)\,b_0^{sep}(e_H)\right)\\ &\quad -\frac{q'(\phi(e_H))\,\phi'(e_H)}{[q(\phi(e_H))]^2}\left(q'(e_H)\frac{\mu_H}{\mu_L}-q'(\phi(e_H))\,\phi'(e_H)\,b_0^{sep}(e_H)\right)\\ &\quad -\phi''(e_H)\frac{q'(\phi(e_H))}{q(\phi(e_H))}b_0^{sep}(e_H).\end{aligned}$$

A.3.20 Proof of Proposition 13

Define $c = C_2/C_1$, and denote A's type by $\theta \in \{H, L\}$

Note that

$$\begin{aligned}V(\varepsilon(c),c) &= -\sum_t \mu_t\left(1-q\left(e_t^{sep}\right)\right)c\\ &\quad +\sum_t \mu_t\underbrace{\left((1-\varepsilon(c))\left(1-q\left(e_t^{\varepsilon(c)}\right)\right)+\varepsilon(c)\,q\left(e_t^{\varepsilon(c)}\right)\right)}_{\Pr(s=B|e)=-q\left(e_t^{\varepsilon(c)}\right)(1-2\varepsilon(c))+1-\varepsilon(c)}\\ &\quad +\sum_t \mu_t\varepsilon(c)\left(1-q\left(e_t^{\varepsilon(c)}\right)\right)c.\end{aligned}$$

By the implicit function theorem, we have

$$\begin{aligned}&-\sum_t \mu_t\left(1-q\left(e_t^{sep}\right)\right)+\sum_t \mu_t\begin{pmatrix} -q'\left(e_t^{\varepsilon(c)}\right)(1-2\varepsilon(c))\frac{de_t^{\varepsilon}}{d\varepsilon}\varepsilon'(c)\\ +q\left(e_t^{\varepsilon(c)}\right)2\varepsilon'(c)-\varepsilon'(c)\end{pmatrix}\\ &+\sum_t \mu_t\left(\varepsilon'(c)\left(1-q\left(e_t^{\varepsilon(c)}\right)\right)-\varepsilon(c)\,q'\left(e_t^{\varepsilon(c)}\right)\frac{de_t^{\varepsilon}}{d\varepsilon}\varepsilon'(c)\right)c\\ &+\sum_t \mu_t\varepsilon(c)\left(1-q\left(e_t^{\varepsilon(c)}\right)\right)=0.\end{aligned}$$

$\Leftrightarrow$

$$\begin{aligned} & \sum_t \mu_t \left(\varepsilon\left(c\right) \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) - \left(1 - q\left(e_t^{sep}\right)\right) \right) \\ = & \sum_t \mu_t \begin{pmatrix} \left(1 - 2\varepsilon\left(c\right) + \varepsilon\left(c\right) c\right) q'\left(e_t^{\varepsilon(c)}\right) \frac{de_t^{\varepsilon}}{d\varepsilon} \\ + \left(1 - 2q\left(e_t^{\varepsilon(c)}\right)\right) - \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) c \end{pmatrix} \varepsilon'\left(c\right). \end{aligned}$$

Since $V\left(\varepsilon\left(c\right), c\right) = 0$, we have

$$\begin{aligned} & \sum_t \mu_t \varepsilon\left(c\right) \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) - \sum_t \mu_t \left(1 - q\left(e_t^{sep}\right)\right) \\ = & -\frac{1}{c} \sum_t \mu_t \left(\left(1 - \varepsilon\left(c\right)\right) \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) + \varepsilon\left(c\right) q\left(e_t^{\varepsilon(c)}\right) \right), \end{aligned}$$

and so

$$\sum_t \mu_t \left(\varepsilon\left(c\right) \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) - \left(1 - q\left(e_t^{sep}\right)\right) \right) < 0.$$

Moreover, we have

$$\left(1 - 2\varepsilon\left(c\right) + \varepsilon\left(c\right) c\right) q'\left(e_t^{\varepsilon(c)}\right) \frac{de_t^{\varepsilon}}{d\varepsilon} \leq 0$$

by Proposition 10, and

$$\begin{aligned} \left(1 - 2q\left(e_t^{\varepsilon(c)}\right)\right) - \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) c & \leq \left(1 - 2q\left(e_t^{\varepsilon(c)}\right)\right) - \left(1 - q\left(e_t^{\varepsilon(c)}\right)\right) \\ & \leq 0. \end{aligned}$$

Therefore, we have

$$\varepsilon'\left(c\right) \geq 0,$$

as desired.

A.3.21 Derivation of (31)

Given that we reach history h^t, let $J_H(h^t)$ be the social welfare from the high type and $V_{LH}(h^t)$ be the value that type L obtains by pretending to be high-type. Since $\hat{J}_L(V)$ is linear, it is optimal to maximize

$$\mu_H \delta^t \Pr\left(h^t|H\right) J_H\left(h^t\right) + (1-\mu_H)\,\delta^t \Pr\left(h^t|L\right)\left[\bar{J} - \frac{\bar{J} - J_L\left(\bar{V}_L\right)}{\bar{V}_L} V_{LH}\left(h^t\right)\right].$$

Equivalently, dividing both sides by

$$\mu_H \delta^t \Pr\left(h^t|H\right) + (1-\mu_H)\,\delta^t \Pr\left(h^t|L\right),$$

the principal maximizes

$$\mu\left(h^t\right) J_H\left(h^t\right) + \left(1-\mu\left(h^t\right)\right)\left[\bar{J} - \frac{\bar{J} - J_L\left(\bar{V}_L\right)}{\bar{V}_L} V_{LH}\left(h^t\right)\right].$$

Hence the state variables are belief μ and promised utility for the high type V, and we have

$$\begin{aligned}
& \mu J_H(\mu, V) + (1-\mu)\left[\bar{J} - \frac{\bar{J} - J_L(\bar{V}_L)}{\bar{V}_L} V_{LH}(\mu, V)\right] \\
= & \ \mu \int_z [p_z \bar{J} + (1-p_z)\{(1-\delta)\mathbb{E}\left[-C|\psi_z, e_z^H, \iota_z\right] \\
& + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^H, \iota_z\right) J_H\left(\mu_z'\left(\psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V_z'(o_z)\right)\}]dz \\
& + (1-\mu)\bar{J} - (1-\mu)\frac{\bar{J} - J_L(\bar{V}_L)}{\bar{V}_L}\int_z (1-p_z)\{(1-\delta)\left[1 - c_H\left(e_z^L\right)\right] \\
& + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^L, \iota_z\right) V_{LH}\left(\mu_z'\left(\psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V_z'(o_z)\right)\}dz \\
= & \int_z [p_z \bar{J} + (1-p_z)\{(1-\delta)\{\mu \mathbb{E}\left[-C|\psi_z, e_z^H, \iota_z\right] \\
& + (1-\mu)[\bar{J} - \frac{\bar{J} - J_L(\bar{V}_L)}{\bar{V}_L}\left[1 - c_H\left(e_z^L\right)\right]\} \\
& + \delta \sum_{o_z} \Pr\left(o_z|\psi_z, e_z^H, e_z^L, \iota_z\right) J\left(\mu_z'\left(\psi_z, e_z^H, e_z^L, \iota_z, o_z\right), V_z'(o_z)\right)\}]dz.
\end{aligned}$$

A.3.22 Proof of Lemma 12

We first show that $V_{LH}(\mathcal{C}_H^*) \in \left[0, \bar{V}_L\right]$. As in Lemma 2, there exists $\bar{V} \geq (1-\delta)u$ such that $J(\mu_H, V)$ is constant for $V \in \left[0, \bar{V}\right]$ and strictly decreasing for $V > \bar{V}$. Hence $\left[0, \bar{V}\right] = \arg\max_{\hat{V}} J\left(\mu_H, \hat{V}\right)$.

By taking $V^* \leq (1-\delta)u$, we have

$$\begin{aligned}
V_{LH} & \leq V^* \text{ since low type needs to pay more cost for the same effort} \\
& \leq (1-\delta)u \\
& < \bar{V}_L \text{ by Lemma 11,}
\end{aligned}$$

as desired.

Further, we can show that the $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ is not binding: Suppose otherwise. Then when we solve $J_L(V)$, we need to take into account the following effect: if we increase V, then it change V_{HL}, the value that the high type obtains from $\mathcal{C}_L^*$, which may decrease the social welfare since we need to increase V^* to incentivize the high type to tell the truth.

However, since $J\left(\mu_H, V\right)$ is constant for $V \in \left[0, \bar{V}\right]$, as long as V_{HL} is sufficiently small that $V^* \leq \bar{V}$, changing V^* does not affect the ex ante social welfare. Hence solving the problem taking into account V_{HL} but with the guess that $V^* \leq \bar{V}$ does not change $J_L\left(V\right)$.

This implies that $\mathcal{C}_H^* \succcurlyeq_H \mathcal{C}_L^*$ is binding only if $V^* > \bar{V}$, which means $V^* \notin \arg\max_{\hat{V}} J\left(\mu_H, \hat{V}\right)$, that is, we cannot pick V^* to maximize the ex ante welfare since $V^* \geq V_{HL}$ is binding.

However, this implies that the following uniform increase in the replacement probability is welfare improving, which is a contradiction: Upon the arrival of the regulator, regardless of his type, we replace him with probability $p > 0$ by public randomization. If the replacement does not happen, then the principal offers contracts $\{\mathcal{C}_H^*, \mathcal{C}_L^*\}$. Since the replacement does not depend on the declared types, this does not change the incentives while decreases V^*.