Disclosure under Scrutiny: A Theory of Investigation and Information Withholding

Jie Ma*

August 2026

Abstract

Institutions that scrutinize wrongdoing are expected to bring more of it to light. We show that whether they do depends on what an investigation reveals, not on whether one occurs. We study a disclosure game in which an agent privately learns a damaging fact and chooses whether to reveal it, and in which silence may trigger an investigation that reports a verifiable finding. Absent investigation, everyone who knowingly conceals is worse than he is, so disclosure is his only way to separate from them. An investigation that establishes exactly what the agent knew lets him be judged as himself rather than alongside those concealers, and disclosure falls. An investigation that reports only that something was concealed does the opposite: being caught is no longer a shelter but a stigma, since it again groups him with concealers who are uniformly worse, and disclosure can rise above its no-investigation level. The design problem of scrutiny is therefore not how much to investigate, but how finely to report what is found.

Keywords: Information Disclosure; Investigation; Formal theory

*Ph.D. Student in Political Science, Washington University in St. Louis. Contact: jie.m@wustl.edu.

1 Introduction

Perhaps one of the most intriguing puzzles in political economy is why the very institution designed to elicit information ends up suppressing it. To list a few: in the standard model of Bayesian persuasion (Kamenica and Gentzkow 2011), the sender's commitment power leaves the receiver no better off than she would be absent commitment; transparency can eliminate a committee's ability to convey anything useful (Gradwohl and Feddersen 2018); and transparency over an agent's actions can induce conformity with prior beliefs, ultimately harming the principal (Prat 2005).

We study the institutional design problem in voluntary disclosure. Specifically, we study the investigation paradox in settings of voluntary disclosure, showing why and how outside scrutiny—the canonical institutional response to concealment—can *discourage* the disclosure it is meant to induce.

Voluntary disclosure of adverse information is pervasive in public policy. Firms report misconduct to the Department of Justice and the SEC, self-report pollution violations to environmental regulators, and come forward about collusion under antitrust leniency programs. Politically active nonprofits choose whether to reveal their donors, and how much of their spending to characterize as political. In each case, disclosure is verifiable through documents, emails, bank records, and accounting logs—the feature that motivates a large literature on the voluntary disclosure of verifiable information.

Yet the option to disclose is also an option to withhold, which is why these settings rarely rely on disclosure alone. Silence meets not only receiver skepticism but external investigation that independently reveals what was concealed: the SEC can establish what management knew and suppressed, an EPA inspection can document both a violation and the firm's awareness of it, antitrust authorities can expose undisclosed knowledge of a cartel, and FEC enforcement actions and IRS audits can trace contributions a group declined to report. The threat operates before any investigation occurs, shaping the disclosure decision itself—and, as we show, in ways that cut against conventional intuition. This paper develops a theory of voluntary disclosure in dynamic

settings where investigation is possible—a feature largely absent from classical disclosure models.

Using a dynamic disclosure model with investigation, we identify a distinct mechanism, operating through a purely informational channel, that explains why investigation backfires: it bounds the worst-case belief attached to silence, weakening the unraveling force that sustains voluntary disclosure.

The intuition is as follows. Consider a group of government officials who differ in their levels of corruption and who may voluntarily disclose that information to the public. Each official chooses either to withhold or to disclose; disclosure is verifiable, so whatever is disclosed must be true. The key observation is that officials do not always prefer to withhold. If an official discloses, verifiability guarantees that he is punished for what he reveals. If he stays silent, he escapes that direct punishment, but he also knows that every official more corrupt than he is will stay silent as well. What drives disclosure, then, is not the official's own information but the pooling behavior of others: silence is costly precisely because it lumps him together with worse types. This reasoning is formalized in Lemma 1.

Investigation does not change the nature of the choice, only its consequences. If the official withholds, he is either caught or he gets away. Being caught reveals his corruption; getting away pools him only with the worse types who also got away. Because investigation is probabilistic, the inference drawn from silence can never be harsher than under no investigation—the threat of pooling with worse types is diluted. Investigation can therefore paradoxically discourage disclosure.

We distinguish two features of an investigation technology: accuracy and precision. Accuracy is the probability that investigation detects concealment ("is he hiding something?"); precision is the fineness of the investigative outcome ("with whom is he hiding?"). Proposition 1 establishes the haunting force of pooling with worse types in a static benchmark without investigation. Proposition 2 then varies accuracy while holding precision at its finest level and compares disclosure behavior with and without investigation: for any level of accuracy, investigation discourages disclosure. High precision thus generates a quantity–quality tradeoff—the information that does come out is sharper, but less of it is volunteered. This raises a natural puzzle: if investigation discour-

ages disclosure, why investigate at all? Proposition 3 identifies a low-precision advantage. When accuracy is sufficiently high, coarse investigation can encourage disclosure. High accuracy makes concealment likely to be detected, and low precision means that, once detected, an official is likely to be pooled with types worse than his own. High accuracy and low precision together thus restore the very pooling threat that fine investigation dilutes.

Taken together, our results show that whether investigation curbs information withholding depends on a domino effect in beliefs. In deciding whether to speak, each individual reasons about the incentives of others, and the effectiveness of investigation hinges on how far it alters the hidden connotation of silence.

The literature closest to this paper concerns mechanisms that halt unraveling short of full revelation. In the classic disclosure game (Milgrom 1981), information endowment is certain and maximal skepticism about silence drives full unraveling. Subsequent work identifies frictions that leave some information withheld in equilibrium.

Most of these operate on the Sender's side. Dye (1985) and Jung and Kwon (1988) show that when the Receiver is unsure whether the Sender is informed, silence becomes excusable. Timing supplies a second friction: Acharya, DeMarzo and Kremer (2011) establish that when information arrives stochastically, the anticipation of future bad news pulls disclosure forward, and Guttman, Kremer and Skrzypacz (2014) show that with multiple pieces of evidence arriving over time, late disclosure can be evaluated more favorably than early disclosure. A third strand introduces an investigation device directly. In Gratton, Holden and Kolotilin (2018), senders who choose when to disclose delay bad news to reduce exposure to outside scrutiny; in Zhou (2024), good types separate by committing to longer disclosure windows. Organizational structure supplies a fourth: Onuchic and Ramos (2023) study teams in which only some members hold decision rights, so that silence by a rank-and-file member escapes the worst-case inference precisely because his choice did not determine the collective outcome.

A separate strand locates the friction on the Receiver's side. Shishkin, Titova and Zhang (2025) show that when the Receiver chooses from finitely many actions, small belief changes need not

move her action, creating credible room for pooling even though the Sender can prove any true fact; sender-preferred equilibria then pool nonadjacent states. Gieczewski and Titova (2024) instead propose an equilibrium selection criterion for disclosure games with general message mappings.

We contribute a mechanism that operates through neither channel. Endowment uncertainty, timing, evidence multiplicity, team structure, and action coarseness all weaken unraveling by making silence *less* informative. Investigation does the opposite: it makes silence *more* informative, yet still weakens unraveling, because what it reveals is which types silence is consistent with. This caps the penalty the Receiver can attach to non-disclosure. The comparison with Shishkin, Titova and Zhang (2025); Callander, Lambert and Matouschek (2021) is instructive, since coarseness enters both models with opposite sign: there, a coarse action space shelters the Sender and sustains pooling; here, a coarse investigation report exposes him, and it is *fine* reports that shelter. Our results show that this distinction determines the direction of the effect. We characterize how the disclosure threshold changes across different investigation regimes and provide a uniform example throughout to better illustrate the intuition.

2 The Model

We have a group of Senders(S,he) and a Receiver(R, she). The Receiver wants to identify the type of each Sender, while all Senders want to be identified as low(er) types.

Every Sender has a type $t \in T \equiv [\underline{\theta}, \overline{\theta}]$ where $t \sim F$. We assume that F has full support. With probability p, each Sender privately receives an empty signal, which suggests that he does not get to learn his type ($s = \emptyset$) and with probability $1-p$ he receives a signal identical to type($s = t$). The value of p corresponds to the genuine possibility of not knowing one's type.

Formally, we have

$$Pr[s = \theta | t = \theta] = 1 - p$$

$$Pr[s = \emptyset | t = \theta] = p$$

After the signal realization, each Sender presents his **evidence** to the Receiver, which we denote as $e \in \{s, \emptyset\}$. There are two situations.

1. With probability p, $s = \emptyset$, $e \in \{\emptyset\}$ (The only evidence available is $\emptyset$)

2. With probability $1-p$, $s = t$, $e \in \{t, \emptyset\}$ (Both t and $\emptyset$ are available)

The Receiver chooses action a upon receiving the evidence e. All Senders submit evidence to the Receiver *simultaneously*.

2.1 Utility

The utility function of the Receiver is

$$u_R(a,t) : -(a-t)^2$$

and therefore $\forall t, u_R(\cdot,t)$ is strictly concave, continuously differentiable and admits a maximum at $a = t$. A simple interpretation of the utility function is that the Receiver wants to match the Senders' types.

The utility function of the Sender is

$$u_S(a,t) = -a$$

In other words, all Senders prefer lower actions.

This set-up belongs to the broader literature on information disclosure(disclosure games) and we consider both the static and dynamic versions of that. In the static benchmark, we consider only the static version of the game where there is a one-shot interaction between Senders and Receiver. The game ends after the first stage. This setup is identical to the disclosure model in Dye (1985). Next, we consider the dynamic version of the game by introducting a novel investigation process in the second stage.

2.2 Sequence of the Play

We consider two variants of the static benchmark:

Static Benchmark

1. At the beginning of the game, each Sender gets to know his type $t \in T$ with probability $1-p$
2. All Senders submit evidence to the Receiver simultaneously
3. The Receiver takes action a
4. Game ends and payoffs are realized

No Investigation

1. At the beginning of the game, each Sender gets to know his type $t \in T$ with probability $1-p$
2. If in the previous stage, the Sender does not know his type, he gets another opportunity to learn his type $t \in T$ with probability $1-p$
3. All Senders submit evidence to the Receiver simultaneously
4. The Receiver takes action a
5. Game ends and payoffs are realized

Investigation

1. At the beginning of the game, each Sender gets to know his type $t \in T$ with probability $1-p$
2. All Senders submit evidence to the Receiver simultaneously
3. Investigation on submitted evidence starts
4. The Receiver observes both the evidence and the investigation outcome and take action a
5. Game ends and payoffs are realized

Equilibrium Concept The equilibrium concept is Perfect Bayesian Equilibrium (PBE), which we refer to as equilibrium hereafter; a detailed description is deferred to Appendix A.

3 Analysis

3.1 Static Benchmark

To solve the model, first notice that for Senders who receive empty signals, they have to present null evidence and therefore we only need to consider the strategies of those who are aware of their types.

We provide this useful Lemma first.

Lemma 1. *When $s \neq \emptyset$ and type θ presents evidence $\emptyset$, then for any type $\theta' > \theta$, they present evidence $\emptyset$ with probability 1.*

Proof. Consider Sender at information set $(t = \theta', s = \emptyset)$. It is trivially true since they can not provide any evidence.

Consider Sender at information set $(t = \theta', s = \theta')$. If type θ prefers to hide, then $u_S(a^*(\theta), \theta) < u_S(a^*(\emptyset), \theta)$. Since evidence is verifiable, we have $u_S(a^*(\theta), \theta) = u_S(\theta, \theta)$ and $u_S(\theta, \theta) = -\theta$(by sender's utility function). If type θ' discloses, then $u_S(a^*(\theta'), \theta') = u_S(\theta', \theta') = -\theta'$. In previous steps, we know $u_S(a^*(\theta'), \theta') = -\theta' < -\theta = u_S(a^*(\theta), \theta) < u_S(a^*(\emptyset), \theta) = u_S(a^*(\emptyset), \theta')$. Therefore if θ presents evidence $\emptyset$, $\theta' > \theta$ presents evidence $\emptyset$ as well. ■

In plain words, when a Sender strategically chooses not to disclose, then he knows that all Senders who are worse than himself would not disclose either.

Therefore, for Sender of type θ, he knows that when choosing no disclosure, the posterior of the Receiver upon observing silence is:

$$G(\theta)^S = \frac{\overbrace{p\,\mathbf{E}[t]}^{\text{those who are unknown}} + \overbrace{(1-p)\big(1-F(\theta)\big)\,\mathbf{E}\big[t \mid t \geq \theta\big]}^{\text{those who are worse than himself}}}{p + (1-p)\big(1-F(\theta)\big)} \qquad \text{(Static Benchmark)}$$

Conveniently, if $G(\theta)^S > \theta$, then type θ would prefer to disclose; if instead, $G(\theta)^S < \theta$, then type θ would prefer to hide. The Proposition below characterizes the equilibrium behavior of all Sender types.

Proposition 1. *In the static model without investigation, we have the following cutoff strategy in disclosure*

1. *Senders who receive a signal of $s = \emptyset$ choose no disclosure $e = \emptyset$*
2. *Senders who receive a signal of $s > \theta^*$ choose no disclosure $e = \emptyset$*
3. *Senders who receive a signal of $s \leq \theta^*$ choose to disclose $e = s$*

The threshold θ^* is the solution to $G(\theta)^S = \theta$(i.e. it is a fixed point of $G(\cdot)^S$ function)

3.2 No Investigation

When there is no investigation, the game play remains the same as in the static benchmark, the only difference being that it is played repeatedly. From Receiver's perspective, ex-ante, it is equivalent to the benchmark model with prior p modified to p^2. Therefore, for Sender of type θ, he knows that when choosing no disclosure, the posterior of the Receiver upon observing silence is:

$$G(\theta)^{NI} = \frac{\overbrace{p^2 \mathbf{E}[t]}^{\text{those who are unknown}} + \overbrace{(1-p^2)\big(1-F(\theta)\big)\mathbf{E}\big[t \mid t \geq \theta\big]}^{\text{those who are worse than himself}}}{p^2 + (1-p^2)\big(1-F(\theta)\big)} \qquad \text{(No Investigation)}$$

3.3 Investigation

In this section, we analyze the case when the investigation is available in the second stage.

Definition 1. The accuracy of investigation process is defined as

$$Pr[d=\theta|e=\emptyset,s=\theta,t=\theta]=1-g$$

$$Pr[d=\emptyset|e=\emptyset,s=\theta,t=\theta]=g$$

$$Pr[d=\theta|e=\emptyset,s=\emptyset,t=\theta]=0$$

$$Pr[d=\emptyset|e=\emptyset,s=\emptyset,t=\theta]=1$$

This definition renders some implicit assumptions. First of all, the investigation device does not punish honest reporting behavior. When Sender does know his type and claims so, the investigation will not mark him as dishonest(i.e. $Pr[d=\theta|e=\emptyset,s=\emptyset,t=\theta]=0, Pr[d=\emptyset|e=\emptyset,s=\emptyset,t=\theta]=1$). Second, the investigation device can only detect evidence withholding imperfectly. When Sender knows his type but chooses to hide, the investigation will catch that with probability $1-g$(i.e. $Pr[d=\theta|e=\emptyset,s=\theta,t=\theta]=1-g, Pr[d=\emptyset|e=\emptyset,s=\theta,t=\theta]=g$). Therefore, $1-g$ is the accuracy of the investigation device. The static benchmark is a special case with $g=1$. Also, once detected, the revealed evidence d is identical to Sender's type t.

To summarize, the investigation process defined above models the fact that the investigation device can only detect evidence withholding imperfectly while preserves the verifiability of the evidence structure.

In the benchmark model, if Sender at information set $(t=\theta,s=\theta)$ chooses no disclosure, then he can send $e=\emptyset$ without being detected. With investigation, Sender at information set $(t=\theta,s=\theta)$'s expected payoff in the first stage becomes:

$$\mathbf{E}[u_S(a,t)]=-\left[(1-g)t+gG^I(\theta)\right]$$

That is, with probability $1-g$, his true type will be revealed and with probability g, he can still

hide. The type that is indifferent between disclosing and not disclosing satisfies:

$$\theta = (1-g)\theta + gG^I(\theta) \qquad \text{(Indifference Condition)}$$
$$\theta = G^I(\theta)$$

$$G(\theta)^I = \frac{\overbrace{p\mathbf{E}[t]}^{\text{those who are unknown}} + \overbrace{g(1-p)\big(1-F(\theta)\big)\mathbf{E}\big[t \mid t \geq \theta\big]}^{\text{those who are worse than himself}}}{p + g(1-p)\big(1-F(\theta)\big)} \qquad \text{(Investigation)}$$

Proposition 2. *(investigation discourages disclosure) For any $(p,g) \in (0,1)^2$, denote $\theta^S = G(\theta^S)$, $\theta^{NI} = G(\theta^{NI})^{NI}$ and $\theta^I = G(\theta^I)^I$, we have*

$$\theta^I < \theta^S < \theta^{NI}$$

That is, Senders disclose more without investigation, disclose less with investigation.

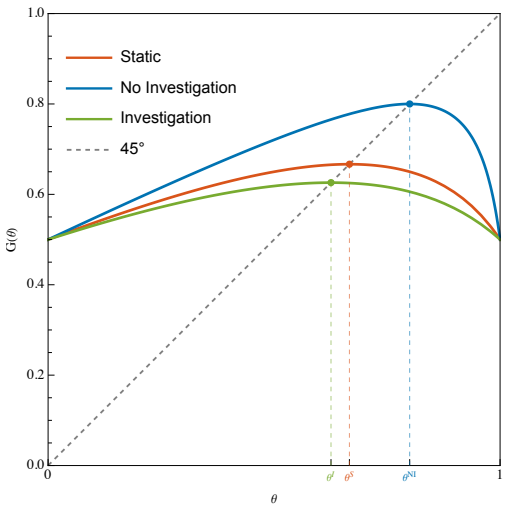
So far we have established the main result that investigation ***discourages*** disclosure. The intuition is that with investigation device, each Sender gets to pool with types that are strictly worse than themselves only with probability g. With probability $1-g$, though they are caught by the investigation device, they also get to *pool only with themselves*. As accuracy gets lower, g increases, which puts a heavier weight on the penalty associated with the pooling behavior. As a consequence of that, the threshold increases and more Senders choose to disclose to separate from worse types.

Example 1. Uniform Case

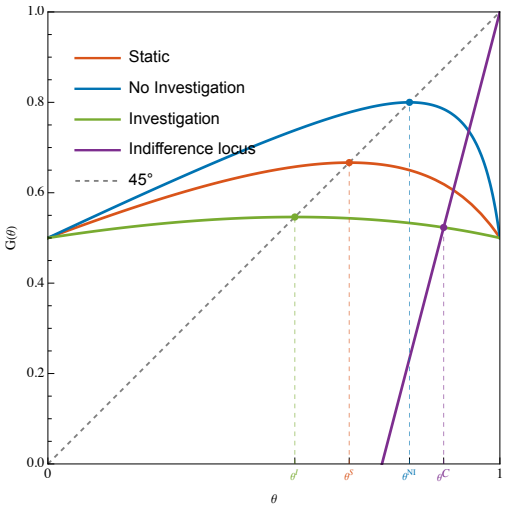
Assume that every Sender has type $t \sim \text{Unif}[0,1]$ and plot G^S, G^{NI} and G^I functions accordingly. The fixed points of the functions can be easily verified by intersecting $G(\cdot)$ and the 45-degree line. See Figure 1a.

3.4 Low Precision Advantage

Corollary 1. *As g increases(accuracy gets lower), the fixed point of $G(\theta)^I$ increases. That is, when accuracy of investigation gets lower, Senders disclose more.*



(a) $g = 0.6$. Three regimes, with $\theta^I < \theta^S < \theta^{NI}$.



(b) $g = 0.15$. Four regimes, with $\theta^I < \theta^S < \theta^{NI} < \theta^C$.

Figure 1: $G(\theta)$ under alternative regimes, $p = 0.25$.

Notice that Receiver's posteriors after observing null evidence in the static benchmark and investigation regime differ by g and when $g = 1$, they are identical. T

$$G(\theta)^S = \frac{\overbrace{p\,\mathbf{E}[t]}^{\text{those who are unknown}} + \overbrace{(1-p)\big(1-F(\theta)\big)\,\mathbf{E}\big[t \mid t \geq \theta\big]}^{\text{those who are worse than himself}}}{p + (1-p)\big(1-F(\theta)\big)} \qquad \text{(Static Benchmark)}$$

$$G(\theta)^I = \frac{\overbrace{p\,\mathbf{E}[t]}^{\text{those who are unknown}} + \overbrace{g\,(1-p)\big(1-F(\theta)\big)\,\mathbf{E}\big[t \mid t \geq \theta\big]}^{\text{those who are worse than himself}}}{p + g\,(1-p)\big(1-F(\theta)\big)} \qquad \text{(Investigation)}$$

Surprisingly, when investigation is inaccurate, Senders are more willing to disclose rather than conceal information. This reflects a tradeoff between the quality and quantity of information: greater investigative accuracy improves the quality of information by allowing the hidden state to be inferred more accurately from null evidence, but reduces the quantity of information disclosed, as fewer Senders choose to reveal it.

In Figure 1, as g increases, the green line approaches the orange line, and the two coincide when $g = 1$. Intuitively, the fixed point of $G(\cdot)^I$ is at most θ^S, which is lower than θ^{NI}. This raises

a natural question: if investigation discourages disclosure, why would a Receiver ever choose to investigate? In particular, can investigation ever induce more disclosure than in the absence of investigation? It turns out that investigation can indeed induce greater disclosure when its precision varies.

Definition 2. The precision of investigation process is defined as

$$Pr[d = \mathscr{D}^{\theta} | e = \emptyset, s = \theta, t = \theta] = 1 - g$$

$$Pr[d = \emptyset | e = \emptyset, s = \theta, t = \theta] = g$$

$$Pr[d = \mathscr{D}^{\theta} | e = \emptyset, s = \emptyset, t = \theta] = 0$$

$$Pr[d = \emptyset | e = \emptyset, s = \emptyset, t = \theta] = 1$$

$\mathscr{D}^t$ denotes the partition that t is identified with. Furthermore, we assume that partitions have the following properties

1. Verifiability: $\{t\} \subseteq \mathscr{D}^t \subseteq T$
2. Uniqueness: $\forall t \in T, \mathscr{D}^t$ is unique

Notice that in Definition 1, the investigation device either reveals the Sender's exact type ($d = t$) or returns no information ($d = \emptyset$). While this formulation is analytically convenient, it is restrictive in practice, as few investigative devices can perfectly identify the Sender's type. We therefore extend the model to allow for a more general investigation process. Specifically, in the following sections, we allow the investigation device to return a partition of the type space, so that the information revealed upon detection need not fully identify the Sender's type. That is, the investigation may be imprecise.

Proposition 3. *(low precision advantage) For any $(p,g) \in (0,1)^2$, denote $\theta^S = G(\theta^S)$, $\theta^{NI} = G(\theta^{NI})^{NI}$ and $\theta^I = G(\theta^I)^I$. With the coarsest investigation device, the threshold exceeds θ^I and exceeds θ^{NI} iff $g < g^*(p)$.*

This proposition demonstrates the *low-precision advantage*: when an investigation becomes more accurate, lower precision can make individuals more willing to disclose overall. This highlights a fundamental distinction between accuracy and precision. Accuracy determines how likely individuals are to be pooled with other types, whereas precision determines which types they are pooled with. Paradoxically, low precision can therefore become an advantage for the Receiver, as it increases the risk that Senders will be pooled with worse types.

Example 2. Uniform Case with Coarsest Partition

Assume that every Sender has type $t \sim \text{Unif}[0,1]$ and consider the coarsest partition $\mathscr{D}^t \equiv T, \forall t \in T$. Then the indifference condition becomes:

$$\theta = (1-g)\frac{\theta+1}{2} + gG^I(\theta) \qquad \text{(Indifference Condition-P)}$$
$$G^I(\theta) = \frac{(1+g)\theta - (1-g)}{2g}.$$

The order of fixed points is illustrated in Figure 1b.

4 Conclusion

This paper has developed a theory of voluntary disclosure in dynamic settings where investigation may follow silence. The central finding is that investigation, despite its reputation as a tool for surfacing information, can reduce disclosure in equilibrium. This poses a trade-off for institutional design. Scrutiny does improve the credibility of what is disclosed—but the same force that improves credibility weakens the incentive to disclose at all. Whether investigation increases information overall depends on how that trade-off resolves.

References

Acharya, Viral V, Peter DeMarzo and Ilan Kremer. 2011. "Endogenous information flows and the clustering of announcements." *American Economic Review* 101(7):2955–2979.

Callander, Steven, Nicolas S Lambert and Niko Matouschek. 2021. "The power of referential advice." *Journal of Political Economy* 129(11):3073–3140.

Dye, Ronald A. 1985. "Disclosure of nonproprietary information." *Journal of accounting research* pp. 123–145.

Gieczewski, Germán and Maria Titova. 2024. "Coalition-Proof Disclosure.".

Gradwohl, Ronen and Timothy Feddersen. 2018. "Persuasion and transparency." *The Journal of Politics* 80(3):903–915.

Gratton, Gabriele, Richard Holden and Anton Kolotilin. 2018. "When to drop a bombshell." *The Review of Economic Studies* 85(4):2139–2172.

Guttman, Ilan, Ilan Kremer and Andrzej Skrzypacz. 2014. "Not only what but also when: A theory of dynamic voluntary disclosure." *American Economic Review* 104(8):2400–2420.

Jung, Woon-Oh and Young K Kwon. 1988. "Disclosure when the market is unsure of information endowment of managers." *Journal of Accounting research* pp. 146–153.

Kamenica, Emir and Matthew Gentzkow. 2011. "Bayesian persuasion." *American Economic Review* 101(6):2590–2615.

Milgrom, Paul R. 1981. "Good news and bad news: Representation theorems and applications." *The Bell Journal of Economics* pp. 380–391.

Onuchic, Paula and Joao Ramos. 2023. "Disclosure and incentives in teams." *arXiv preprint arXiv:2305.03633* .

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95(3):862–877.

Shishkin, Denis, Maria Titova and Kun Zhang. 2025. "Withholding Verifiable Information.".

Zhou, Beixi. 2024. Optimal Disclosure Windows. Technical report Working paper.

Online Supporting Information

Table of Contents

A Equilibrium Concept

Equilibrium Concept The equilibrium concept is Perfect Bayesian Equilibrium (PBE), which we refer to as equilibrium hereafter.

Let $d \in \mathscr{D} \cup \{\emptyset\}$ denote the investigation report, where $d = \emptyset$ records that the investigation returned nothing. In the static benchmark and in the no-investigation game, $d \equiv \emptyset$ throughout, so the Receiver's information set reduces to the evidence alone. Write $\mathscr{I} = \big(\{T \cup \{\emptyset\}\}\big) \times \big(\mathscr{D} \cup \{\emptyset\}\big)$ for the set of the Receiver's information sets, with generic element (e,d).

An assessment is a triple (σ, α, β) consisting of

- a *disclosure rule* $\sigma : T \cup \{\emptyset\} \to \Delta\big(T \cup \{\emptyset\}\big)$, mapping the Sender's private signal s to a distribution over admissible evidence, subject to the verifiability restriction

$$\operatorname{supp} \sigma(\cdot \mid s) \subseteq \begin{cases} \{\emptyset\}, & s = \emptyset, \\ \{s, \emptyset\}, & s \in T; \end{cases}$$

- an *evaluation rule* $\alpha : \mathscr{I} \to \mathbb{R}$, assigning an action to every pair of evidence and investigation report;
- a *belief system* $\beta : \mathscr{I} \to \Delta T$, assigning a posterior over the Sender's type at every information set.

An assessment (σ, α, β) is a *Perfect Bayesian Equilibrium* if:

1. *(Sender optimality)* For every signal s, every $e \in \operatorname{supp} \sigma(\cdot \mid s)$ maximizes the Sender's expected payoff,

$$e \in \arg\max_{e' \in E(s)} -\mathbf{E}_d\big[\alpha(e', d)\big],$$

   where $E(s)$ is the admissible evidence set and the expectation is taken over the investigation report induced by (e', s, t).

2. *(Receiver optimality)* For every $(e,d) \in \mathscr{I}$,

$$\alpha(e,d) \in \arg\max_{a \in \mathbb{R}} \int_T u_R(a,t)\, d\beta(t \mid e,d) = \mathbf{E}_\beta\big[t \mid e,d\big],$$

   the equality following from strict concavity of u_R.

3. *(Bayesian consistency)* At every information set reached with positive probability under σ and the prior F, the belief β is derived from F and σ by Bayes' rule. In particular, the

posterior conditions on *both* components of (e,d), so that an investigation report is evaluated jointly with the disclosure decision that preceded it. Throughout, beliefs respect hard evidence; this is standard in verifiable-disclosure games.

We restrict attention throughout to symmetric equilibria in which all Senders of the same signal play the same disclosure rule, and to equilibria in pure strategies on the path.

B Proof for Lemma 1

Lemma 1 *When $s \neq \emptyset$ and type θ presents evidence $\emptyset$, then for any type $\theta' > \theta$, they present evidence $\emptyset$ with probability 1.*

Proof. The proof is in the main text. ■

C Proof for Proposition 1

Proposition 1 *In the static model without investigation, we have the following cutoff strategy in disclosure*

1. *Senders who receive a signal of $s = \emptyset$ choose no disclosure $e = \emptyset$*
2. *Senders who receive a signal of $s > \theta^*$ choose no disclosure $e = \emptyset$*
3. *Senders who receive a signal of $s \leq \theta^*$ choose to disclose $e = s$*

The threshold θ^* is the solution to $G(\theta)^S = \theta$(i.e. it is a fixed point of $G(\cdot)^S$ function)

Proof. In this proof, we show several things. First of all, we give the derivation for the $G^S(\cdot)$ function. Then we show that the fixed point of G^S exists, is unique and is the global maximum of the G^S function(single-peakedness). The last property is also known as the Maximum Principle, proved in Acharya, DeMarzo and Kremer (2011).

C.1 Derivation of G^S

t is distributed on $T = [\underline{\theta}, \overline{\theta}]$ according to distribution $F(\cdot)$.

Type	Signal	Evidence Submitted	Prob
$t \in T$	$s = \emptyset$	$\emptyset$	p
$t \geq \theta^*$	$s = t$	$\emptyset$	$(1-p)(1-F(\theta^*))$
$t < \theta^*$	$s = t$	t	$(1-p)F(\theta^*)$

Table 1: Probability Distribution of Revelation Decision

$$G(\theta)^S = \frac{p\mathbf{E}(t) + (1-p)(1-F(\theta))(\mathbf{E}[t|t \geq \theta])}{p + (1-p)(1-F(\theta))}$$

$G(\theta)^S$ is the posterior mean after observing evidence $\emptyset$, conditioning on the conjecture that the threshold for no disclosure is θ. The solution solves

$$\theta = G(\theta)^S$$

That is, type θ is indifferent between disclosing and not disclosing.
Denote

$$H(\theta) = G(\theta)^S - \theta$$

$H(\theta)$ is continuous.

C.2 Existence

First notice that $\underline{\theta} < \mathbf{E}[t] < \bar{\theta}$

1. When $\theta = \underline{\theta}$, then $G(\underline{\theta})^S = \mathbf{E}[t]$ since $F(\underline{\theta}) = 0$ and $\mathbf{E}[t|t \geq \underline{\theta}] = \mathbf{E}[t]$. When $\theta = \underline{\theta}$, $H(\underline{\theta}) > 0$.

2. When $\theta = \overline{\theta}$, then $G(\overline{\theta})^S = \mathbf{E}[t]$ since $F(\overline{\theta}) = 1$. When $\theta = \bar{\theta}$, $H(\bar{\theta}) < 0$.

By Intermediate Value Theorem, there must be at least one θ such $H(\theta) = 0$

C.3 Uniqueness

If $H(\theta)$ should intersect with 0 multiple times, then it must intersect with 0 for odd number of times. Suppose it intersects with 0 for $2n+1$ times and denote each intersection point by θ^i.
Then

1. $H'(\theta^i) < 0$, i is odd

2. $H'(\theta^i) > 0$, i is even

$$H'(\theta) = G'(\theta)^S - 1$$

$$(1-F(\theta))\mathbf{E}[t|t \geq \theta] = \int_\theta^{\bar{t}} t f(t) dt$$

$$\frac{d\int_\theta^{\bar{t}} t f(t) dt}{d\theta} = -\theta f(\theta) \text{ By Leibniz Rule}$$

$$\frac{\partial G(\theta)^S}{\partial \theta} = \frac{(1-p)(-\theta f(\theta))[p+(1-p)(1-F(\theta))] - [p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])](1-p)(-f(\theta))]}{[p+(1-p)(1-F(\theta)]^2}$$

FOC is

$$(1-p)(-\theta f(\theta))[p+(1-p)(1-F(\theta))] = [p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])](1-p)(-f(\theta))]$$

Rearrange to get

$$\theta = \frac{p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])}{p+(1-p)(1-F(\theta))} = G(\theta)^S$$

That is, at every intersection point, we must have
$G'(\theta)^S = 0$ and thus

$$H'(\theta) - 1 < 0$$

Therefore, $H(\theta)$ intersects with 0 only once and thus the fixed point of $G(\theta)$ is unique.

C.4 Single-peakedness

$$\frac{\partial G(\theta)^S}{\partial \theta} = \frac{(1-p)(-\theta f(\theta))[p+(1-p)(1-F(\theta))] - [p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])](1-p)(-f(\theta))]}{[p+(1-p)(1-F(\theta)]^2}$$

Let

$$A = (1-p)(-\theta f(\theta))[p+(1-p)(1-F(\theta))] - [p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])](1-p)(-f(\theta))]$$
$$B = [p+(1-p)(1-F(\theta)]^2$$

If a continuous function has only one critical point on its domain and it is a local maximum, then it is the global maximum. Therefore we check the SOC at the critical point, where $A = 0$.

$$\frac{\partial G(\theta)^S}{\partial \theta} = \frac{A}{B}$$

$$\frac{\partial^2 G(\theta)^S}{\partial \theta^2} = A'B^{-1} - AB^{-2}B'$$

$$A' = (1-p)(-f(\theta)-\theta f'(\theta))[p+(1-p)(1-F(\theta))]+(1-p)(-\theta f(\theta))[(1-p)(-f(\theta))]$$
$$+(1-p)\theta f(\theta)(1-p)(-f(\theta))+[p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])](1-p)f'(\theta)$$
$$B' = 2[p+(1-p)(1-F(\theta)](1-p)(-f(\theta))$$

$$Sgn(A'B^{-1}) = Sgn(A') = Sgn([p+(1-p)(1-F(\theta))](-f(\theta))) < 0$$

Therefore the fixed point of $G(\theta)^S$ is both the local and global maximum. ■

D Proof for Proposition 2

Proposition 2 *(investigation discourages disclosure) For any $(p,g)\in(0,1)^2$, denote $\theta^S = G(\theta^S)$, $\theta^{NI} = G(\theta^{NI})^{NI}$ and $\theta^I = G(\theta^I)^I$, we have*

$$\theta^I < \theta^S < \theta^{NI}$$

That is, Senders disclose more without investigation, disclose less with investigation.

Proof. We complete the proof in two steps. We first show that $\theta^S < \theta^{NI}$ is true and then we show $\theta^I < \theta^S$ is true.

D.1 $\theta^S < \theta^{NI}$

Notice that

$$G(\theta)^S = \frac{p\mathbf{E}(t)+(1-p)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])}{p+(1-p)(1-F(\theta))}$$

$$G(\theta)^{NI} = \frac{p^2\mathbf{E}(t)+(1-p^2)(1-F(\theta))(\mathbf{E}[t|t\geq\theta])}{p^2+(1-p^2)(1-F(\theta))}$$

$$\frac{\partial G(\theta)^S}{\partial p} = \frac{(1-F(\theta))\Big(\mathbb{E}[t]-\mathbb{E}[t\mid t\geq\theta]\Big)}{\big(p+(1-p)(1-F(\theta))\big)^2} \leq 0$$

This is essentially an application of the comparative statics result of $\frac{\partial G(\theta)^S}{\partial p}$. let $p' = p^2$. Since $p\in(0,1)$, then $p' = p^2 < p$.

$$G(\theta)^{NI} = \frac{p'\mathbf{E}(t)+(1-p')(1-F(\theta))(\mathbf{E}[t|t\geq\theta])}{p'+(1-p')(1-F(\theta))}$$

That is, $\forall\theta \in [\underline{\theta}, \overline{\theta}]$, we have

$$G(\theta)^{NI} > G(\theta)^{S}$$

Notice that θ^{NI} is the fixed point of $G(\theta^{NI})^{NI}$. Therefore by a similar logic as in the proof for Proposition 1, θ^{NI} is also unique and is the global maximum of G^{NI}.
Therefore $\theta^{S} < \theta^{NI}$.

D.2 $\theta^{I} < \theta^{S}$

Type	Signal	Evidence	Prob
$t \in T$	$s = \emptyset$	$(\emptyset, \emptyset)$	p
$t \geq \theta_0^*$	$s = t$	$(\emptyset, d)$	$(1-g)(1-p)(1-F(\theta_0^*))$
$t \geq \theta_0^*$	$s = t$	$(\emptyset, \emptyset)$	$g(1-p)(1-F(\theta_0^*))$
$t < \theta_0^*$	$s = t$	$(t, \cdot)$	$(1-p)F(\theta_0^*)$

Table 2: Probability Distribution of Revelation Decision

$$G(\theta)^{I} = \frac{p\mathbf{E}(t) + g(1-p)(1-F(\theta))(\mathbf{E}[t|t \geq \theta])}{p + g(1-p)(1-F(\theta))}$$

$G(\theta)^{I}$ is the posterior mean after observing evidence $(\emptyset, \emptyset)$, conditioning on the conjecture that the threshold for no disclosure is θ.
If type θ is indifferent between disclosing and not disclosing, then

$$\theta = gG(\theta)^{I} + (1-g)\theta$$

Organize to get

$$\theta^{I} = G(\theta^{I})^{I}$$

All the previous properties of fixed points follow. That is, θ^{NI} is still unique and is the global maximum. Notice that

$$G(\theta)^{I} = \frac{p\mathbf{E}(t) + g(1-p)(1-F(\theta))(\mathbf{E}[t|t \geq \theta])}{p + g(1-p)(1-F(\theta))}$$

$$G(\theta)^{S} = \frac{p\mathbf{E}(t) + (1-p)(1-F(\theta))(\mathbf{E}[t|t \geq \theta])}{p + (1-p)(1-F(\theta))}$$

$$G(\theta)^{S} - G(\theta)^{I} > 0, \forall\theta \in T$$

Therefore the fixed point of $G(\theta)^{S}$ must be larger than that of $G(\theta)^{I}$

■

E Proof for Corollary 1

Corollary 1 *As g increases(accuracy gets lower), the fixed point of $G(\theta)^I$ increases. That is, when accuracy of investigation gets lower, Senders disclose more.*

Proof. Notice that when $g = 1$, then $\theta^I = \theta^S = \theta^{NI}$. We show the comparative statics results to complete the proof.

$$\frac{\partial G^I}{\partial g} = \frac{p(1-p)(1-F(\theta))\left[\mathbb{E}[t|t \geq \theta] - \mathbb{E}[t]\right]}{(p+g(1-p)(1-F(\theta)))^2} \geq 0$$

Therefore, as g increases, $G(\theta)^I$ increases. Given that θ^I is the global maximum of G^I, θ^I also increases.

■

F Proof for Proposition 3

Proposition 3 *(low precision advantage) For any $(p,g) \in (0,1)^2$, denote $\theta^S = G(\theta^S)$, $\theta^{NI} = G(\theta^{NI})^{NI}$ and $\theta^I = G(\theta^I)^I$. With the coarsest investigation device, the threshold exceeds θ^I and exceeds θ^{NI} iff $g < g^*(p)$.*

Proof. The technology is a partition of the type space T. If type θ is caught, the technology returns a report $\mathscr{D}^\theta$, which satisfies $\theta \in \mathscr{D}^\theta \subseteq T$. The correct conjectures of the population that hides is therefore a truncated type space, which is $\mathscr{D}^\theta \cap [\theta, \overline{\theta}]$.
Consider the coarsest partition. Then the indifference condition becomes

$$\theta = (1-g)\mathbf{E}[t|t \geq \theta] + gG(\theta)^I$$

As $g \to 0$, $\theta \to \overline{\theta}$ and as $g \to 1$, $\theta \to \theta^I$, which is smaller than θ^{NI}.
Let $\theta = \theta^{NI}$ to have

$$\theta^{NI} = (1-g)\mathbf{E}[t|t \geq \theta^{NI}] + gG(\theta^{NI})^I$$

Organize to get

$$g[G(\theta^{NI})^I - \mathbf{E}[t|t \geq \theta^{NI}]] = \theta^{NI} - \mathbf{E}[t|t \geq \theta^{NI}]$$

$$g^*(p) = \frac{\mathbf{E}[t|t \geq \theta^{NI}] - \theta^{NI}}{\mathbf{E}[t|t \geq \theta^{NI}] - G(\theta^{NI})^I}$$

Notice that G^I function has a global maximum at θ^I and therefore $G(\theta^{NI})^I < G(\theta^I)^I = \theta^I$.
Then

$$g^*(p) > \frac{\mathbf{E}[t|t \geq \theta^{NI}] - \theta^{NI}}{\mathbf{E}[t|t \geq \theta^{NI}] - \theta^I}$$

From Proposition 2, we know that $\theta^{NI} > \theta^{I}$, which confirms that

$$\frac{\mathbf{E}[t|t \geq \theta^{NI}] - \theta^{NI}}{\mathbf{E}[t|t \geq \theta^{NI}] - \theta^{I}} < 1$$

■