AI Washing or Substantive Adoption? A Game-Theoretic Model of Corporate AI Disclosure under Investor Scrutiny and Regulatory Enforcement

Chengcheng Wu[1], Zhiqiang Wang[1,*]

[1] School of Accounting, Guangzhou Huashang College, Guangzhou, Guangdong, China 511300

* Email: ynlpwzq@126.com (Corresponding Author)

Received 1 August 2025 · Revised 15 September 2025 · Accepted 30 October 2025 · Published 30 March 2026

Abstract

Firms increasingly advertise their use of artificial intelligence (AI), and securities regulators have begun to sanction "AI washing", that is, claims of AI capabilities that firms do not have. This study develops a game-theoretic model in which firms decide whether to adopt AI at a privately known cost and then choose between a substantive, verifiable AI claim and a vague one. Investors price firms on the basis of their disclosures, and a regulator audits substantive claims and sanctions false ones. Under the D1 refinement, the perfect Bayesian equilibrium is unique and moves through four regimes as the expected sanction rises: silence, pooling washing, partial washing, and truthful disclosure. Closed-form expressions are derived for the adoption rate, the share of firms making substantive claims, and the share of those claims that are false. Two results matter for business data analytics. First, in the partial-washing regime, measures that count substantive AI claims in corporate text overstate true adoption by a factor equal to the price stake of a claim divided by the expected sanction, and stronger enforcement raises true adoption while lowering measured claims. Second, the welfare-maximizing audit intensity is interior, has a closed form, and tolerates some washing; without commitment, the regulator does not audit and informative disclosure disappears. Numerical experiments, validated by an exhaustive equilibrium search and Monte Carlo simulation, show that the optimal policy raises adoption from 25% to 35% of firms and welfare by 7.0% relative to no enforcement, whereas full deterrence costs more than it gains. Sensitivity analysis and extensions to audit errors and sanction caps identify when these conclusions change.

Keywords: AI washing; Corporate disclosure; Signaling game; Regulatory enforcement; Information asymmetry; AI adoption measurement; Textual analysis; Game theory

AI Washing or Substantive Adoption? A Game-Theoretic Model of Corporate AI Disclosure under Investor Scrutiny and Regulatory Enforcement

1. Introduction

Artificial intelligence (AI) has become a central theme of corporate communication. Firms describe AI initiatives in annual reports, earnings calls, and investor presentations, and investors have reasons to listen: firms that invest in AI grow faster in sales, employment, and market value, largely through product innovation (Babina et al., 2024), and establishments exposed to AI change their hiring in measurable ways (Acemoglu et al., 2022). Actual use of AI, however, has been concentrated in a minority of mostly larger firms (McElheran et al., 2024), and its benefits take years to materialize (Brynjolfsson et al., 2021). When investors reward AI claims, non-adopters are tempted to claim that they have adopted AI.

Regulators now treat this temptation as a matter of securities enforcement. In March 2024, the U.S. Securities and Exchange Commission (SEC) settled charges against two investment advisers, Delphia (USA) Inc. and Global Predictions Inc., for false and misleading statements about their use of AI; the firms paid combined civil penalties of $400,000, and the SEC Chair described such conduct as AI washing that hurts investors (U.S. Securities and Exchange Commission [SEC], 2024). In January 2025, the SEC brought its first AI-washing action against a public company, Presto Automation Inc., which had claimed that its drive-through voice-ordering product eliminated the need for human order-taking when most orders in fact required human intervention (SEC, 2025). Such cases show that specific AI claims can be checked and punished, but also that enforcement is selective and costly.

The same problem confronts researchers and analysts who measure AI adoption. Because firm-level survey data are scarce, a growing literature measures firms' exposure to technologies and risks from corporate text (Gentzkow et al., 2019; Hassan et al., 2019; Loughran & McDonald, 2011), and counting AI claims in annual reports is a natural extension. Yet if firms choose what to say strategically, the count of claims is an equilibrium object that depends on the payoff to claiming and on the expected penalty for claiming falsely, as the experience with greenwashing suggests (Delmas & Burbano, 2011; Lyon & Maxwell, 2011; Marquis et al., 2016).

This study asks when non-adopters claim AI use, how far disclosure-based measures depart from true adoption, and what audit intensity a regulator should choose when audits are costly. In the signaling game developed here, firms with private adoption costs decide whether to adopt AI and then issue either a substantive claim, which can be audited and reveals proprietary information, or a costless vague claim. Investors price firms from their disclosures, and a regulator commits to the probability of auditing substantive claims, so the expected sanction acts as a policy-determined cost of lying.

Four results emerge. First, under the D1 refinement (Cho & Kreps, 1987), the equilibrium is unique and passes through four regimes as the expected sanction rises: silence, pooling washing, partial washing, and truthful disclosure. Second, adoption rises with enforcement but stays below the first-best level. Third, in the partial-washing regime, measured substantive claims exceed true adoption by the price stake of a claim divided by the expected sanction, so stronger enforcement raises adoption while reducing measured claims. Fourth, the welfare-maximizing audit intensity has a closed form, is typically interior, and tolerates some false claims; without commitment, the regulator does not audit and the market reverts to silence.

The study contributes to disclosure theory by combining endogenous investment with claims that are verifiable only through costly enforcement, between the unraveling results of Grossman (1981) and Milgrom (1981) and the cheap-talk benchmark of Crawford and Sobel (1982); to business data analytics by mapping disclosure-based measures of AI adoption to true adoption; and to enforcement economics (Becker, 1968; Polinsky & Shavell, 2000) by showing how deterrence changes the information available to investors. Section

2 reviews the literature, Sections 3 to 5 present and solve the model, Sections 6 and 7 report numerical results and implications, and Section 8 concludes.

2. Literature Review

2.1 Voluntary Disclosure and Its Credibility

When disclosures are verifiable and costless, firms with favorable information disclose it and silence is read as bad news (Grossman, 1981; Milgrom, 1981). Unraveling fails when disclosure is costly or when investors are unsure whether a firm has information: a proprietary cost sustains a threshold equilibrium (Verrecchia, 1983), and uncertainty about the manager's information endowment has a similar effect (Dye, 1985). Reviews of the empirical literature document that disclosure responds to capital-market incentives, litigation risk, and proprietary costs (Beyer et al., 2010; Healy & Palepu, 2001). The present model adopts the proprietary-cost channel but differs in two respects: the information disclosed is the outcome of an investment decision that anticipates the disclosure environment, and claims are verifiable only probabilistically.

2.2 Cheap Talk, Costly Lying, and Signaling

With costless and unverifiable messages, communication is limited by conflicts of interest (Crawford & Sobel, 1982). Lying costs generate inflated but partly informative messages (Kartik et al., 2007; Kartik, 2009), uncertainty about reporting objectives determines how much bias investors can undo (Fischer & Verrecchia, 2000), the credibility of soft information depends on the possibility of verification (Bertomeu & Marinovic, 2016), and repeated interaction can make voluntary disclosure credible (Stocken, 2000). Type-dependent signal costs allow separation (Spence, 1973), and D1 disciplines off-path beliefs (Cho & Kreps, 1987). Here the cost of a false claim is a policy variable chosen before firms act, so the regulator resembles the designer of an information structure (Kamenica & Gentzkow, 2011), except that its instrument is costly and works through the senders' incentives.

2.3 Greenwashing as an Analogue

Greenwashing combines poor environmental performance with positive communication about it (Delmas & Burbano, 2011) and is typically identified through external accusation (Seele & Gatti, 2017). Lyon and Maxwell (2011) show that a firm facing an activist audit discloses selectively and that harsher punishment of greenwash can deter disclosure altogether, and Marquis et al. (2016) find that greater scrutiny is associated with less selective disclosure. The model is also related to quality disclosure and certification (Dranove & Jin, 2010) and to the lemons problem, in which unobservable quality depresses the reward for genuine quality (Akerlof, 1970).

2.4 Measuring AI Adoption

Existing measures of firm-level AI adoption use online job vacancies (Acemoglu et al., 2022), employee skills and job postings (Babina et al., 2024), or Census survey data (McElheran et al., 2024), each with limits of coverage, timing, or cost. Corporate text is available for large panels and can be processed at scale (Gentzkow et al., 2019), and dictionary-based methods have produced firm-level measures of tone and risk (Hassan et al., 2019; Loughran & McDonald, 2011). For AI claims, however, there is a strategic threat to validity in addition to classification error: whether a statement reflects adoption depends on what it costs a non-adopter to make the same statement.

2.5 Economics of Enforcement

Since Becker (1968), enforcement economics has analyzed how the probability of detection and the size of the sanction combine to deter harmful acts. When sanctions are costless transfers and detection is costly, the optimal sanction is maximal and detection correspondingly rare, while sanction limits and enforcement errors qualify this conclusion (Polinsky & Shavell, 2000). In the present setting, deterrence also has an

informational dividend: by making genuine claims credible, it raises the share of the value of adoption captured through the market price, and hence adoption itself, given the short-term price pressure studied by Stein (1989).

3. Model Setup, Notation, and Assumptions

3.1 Players and Timing

The economy consists of a continuum of firms of unit mass, competitive risk-neutral investors, and a regulator. Each firm i has baseline value V_0 and can adopt AI at a private cost K_i, drawn independently from the uniform distribution on $[0, \bar{K}]$. Adoption raises fundamental value by $\Delta > 0$. A firm that adopts is of type H and one that does not is of type L, so that $V_i = V_0 + \Delta$ for type H and $V_i = V_0$ for type L. Each firm then sends a message $m_i \in \{S, N\}$. A substantive claim S describes AI use in specific, checkable terms, such as the systems deployed or quantified effects; a vague claim N consists of generic statements or no statement at all. Because a substantive claim reveals details of the firm's technology to rivals, a genuine adopter that makes it bears a proprietary cost $c > 0$ (Verrecchia, 1983).

The timing is shown in Figure 1. At Stage 0, the regulator commits to an audit intensity $q \in [0, 1]$, the probability that a substantive claim is audited, while the sanction F for a false claim is set by law. At Stage 1, firms observe their costs and decide whether to adopt; at Stage 2, they choose messages; at Stage 3, investors price them; and at Stage 4, the regulator audits substantive claims, and a firm whose claim is found false pays F. Vague claims are not audited because they contain nothing verifiable. An analyst who reads annual reports observes the messages but not the types.

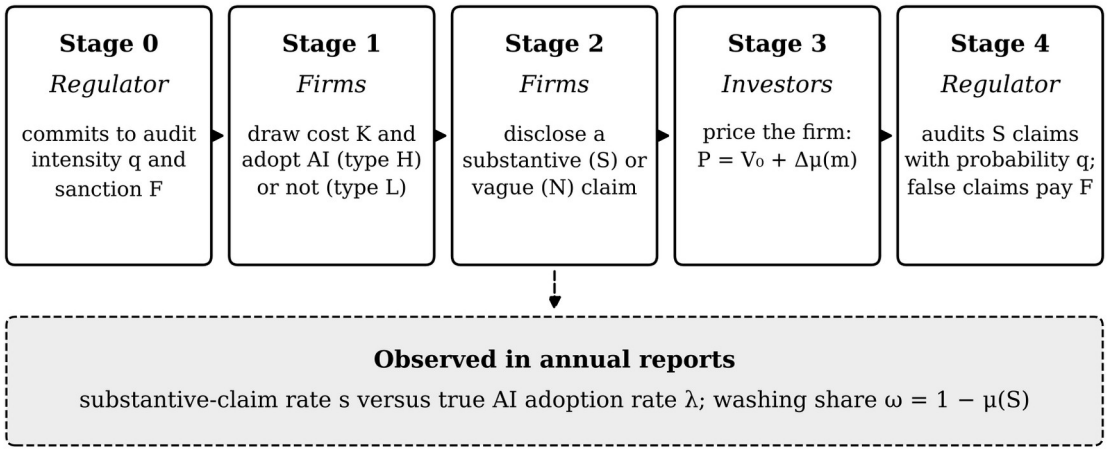


Figure 1. Timing of the Disclosure and Enforcement Game

3.2 Payoffs and Prices

Let $\mu(m)$ denote investors' posterior probability that a firm sending message m is of type H. Competitive pricing gives

$$P(m) = V_0 + \Delta\mu(m). \quad (1)$$

Following Stein (1989), the manager places weight $\alpha \in (0, 1)$ on the current price and $1 - \alpha$ on fundamental value, reflecting equity issuance, takeover pressure, and stock-based pay. Let $e = qF$ be the expected sanction for a false substantive claim and $t_i \in \{H, L\}$ the type of firm i. The payoff of firm i is

$$U_i = \alpha P(m_i) + (1 - \alpha)V_i - K_i a_i - c \cdot 1[m_i = S, t_i = H] - e \cdot 1[m_i = S, t_i = L], \quad (2)$$

where $a_i \in \{0, 1\}$ indicates adoption and $1[\cdot]$ is the indicator function. Eq. (2) shows the source of credibility: a substantive claim costs a genuine adopter c and a non-adopter e, so it can separate the two types only when e exceeds c. Section 5.4 adds audit errors.

3.3 Equilibrium Concept

A strategy specifies an adoption rule and the probabilities σ_H and σ_L with which each type sends S. A firm adopts if and only if $K_i \le G$, where G is the equilibrium gain from adoption, so the adoption rate is $\lambda = G/\bar{K}$. The solution concept is perfect Bayesian equilibrium with beliefs off the equilibrium path restricted by the D1 criterion (Cho & Kreps, 1987), which attributes an unexpected message to the type that would gain from it for the widest range of investor responses.

3.4 Assumptions and Measured Quantities

Two assumptions are maintained. A1 (interior adoption): $\bar{K} > \Delta$, so that some firms do not adopt even under full credibility. A2 (moderate proprietary cost): $c < \min\{\alpha(1-\alpha)\Delta^2/\bar{K}, \alpha\Delta(1-\lambda_{max})\}$, where $\lambda_{max} = (\Delta - c)/\bar{K}$. The first bound ensures that all four regimes below exist, and the second that a genuine adopter always gains from a claim that investors believe; when only the first bound fails, the pooling-washing regime disappears. For an equilibrium with adoption rate λ and strategies (σ_H, σ_L), an analyst who observes disclosures cares about

$$s = \lambda\sigma_H + (1-\lambda)\sigma_L, \quad \omega = (1-\lambda)\sigma_L/s, \quad R = s/\lambda. \quad (3)$$

The substantive-claim rate s is what a text-based measure counting substantive AI claims records; the washing share $\omega = 1 - \mu(S)$ is the fraction of substantive claims that are false; and the measurement ratio R equals one for an unbiased measure and exceeds one when the measure overstates adoption. Table 1 summarizes the notation.

Table 1. Notation and Definitions

Symbol	Definition
Δ, V_0	Increase in fundamental value from AI adoption; baseline firm value
K_i, $\bar{K}$	Adoption cost of firm i, distributed uniformly on $[0, \bar{K}]$
H, L; S, N	Types (adopter, non-adopter); messages (substantive claim, vague or no claim)
α, c	Weight of the current price in the manager's objective; proprietary cost of a substantive claim
q, F, e	Audit intensity; sanction for a detected false claim; expected sanction $e = qF$
σ_H, σ_L	Probability that a type H or type L firm sends S
$\mu(m)$, λ	Posterior that a firm sending m is of type H; true AI adoption rate
s, ω, R	Substantive-claim rate, washing share, and measurement ratio s/λ
M	Expected squared mispricing, $E[(P - V)^2]$
θ, κ, ε	Welfare weight of mispricing; enforcement cost parameter; audit error rate

4. Equilibrium Analysis

4.1 Disclosure Regimes

For a given adoption rate λ, sending S rather than N changes a firm's payoff by $\alpha\Delta[\mu(S) - \mu(N)]$ minus the cost of the claim, which is c for type H and e for type L. When e > c, any belief that makes S attractive for type L makes it strictly attractive for type H, so D1 assigns an unexpected substantive claim to type H and an unexpected vague claim to type L; when e ≤ c, the ranking is reversed. These observations determine the equilibrium.

Proposition 1 (Disclosure regimes). Under A1 and A2, for every expected sanction e the D1 equilibrium is unique and belongs to one of four regimes, separated by the thresholds $c < \hat{e} < \alpha\Delta$, where

$$\hat{e} = \alpha\Delta[(1 - \alpha)\Delta - c]/(\bar{K} - \alpha\Delta). \quad (4)$$

(I) Silence ($e \le c$): $\sigma_H = \sigma_L = 0$, and all firms are priced at $\mu(N) = \lambda$. (II) Pooling washing ($c < e < \hat{e}$): $\sigma_H = \sigma_L = 1$ and $\mu(S) = \lambda$. (III) Partial washing ($\hat{e} \le e < \alpha\Delta$): $\sigma_H = 1$, $\mu(N) = 0$, $\mu(S) = e/(\alpha\Delta)$, and

$$\sigma_L = \lambda(\alpha\Delta - e)/[(1 - \lambda)e]. \quad (5)$$

(IV) Truthful disclosure ($e \ge \alpha\Delta$): $\sigma_H = 1$, $\sigma_L = 0$, $\mu(S) = 1$, and $\mu(N) = 0$.

Proof sketch. If e ≤ c, any profile in which type H sends S makes S strictly attractive to type L, so S cannot carry a premium that covers c. If e > c, D1 sets $\mu(S) = 1$ when S is unused, which type H would exploit because A2 implies $\alpha\Delta[1 - \mu(N)] > c$. Type L also sends S while $\alpha\Delta\lambda > e$; evaluating λ by Proposition 2 gives ê, and A2 guarantees ê > c. In Regime III, type L is indifferent, $\alpha\Delta\mu(S) = e$, and Bayes' rule gives Eq. (5), with $\sigma_L \le 1$ exactly when $e \ge \hat{e}$. For $e \ge \alpha\Delta$, the sanction exceeds the largest price gain from a claim. Within each interval of e, the conditions of the other regimes fail, which gives uniqueness.

4.2 Adoption

Proposition 2 (Adoption). The equilibrium gain from adoption and the adoption rate are

$$G(e) = (1 - \alpha)\Delta + \max\{\min(e, \alpha\Delta) - c, 0\}, \quad \lambda^*(e) = G(e)/\bar{K}. \quad (6)$$

Adoption is constant in Regime I, rises with e at rate $1/\bar{K}$ in Regimes II and III, and reaches λ_{max} in Regime IV, which is below the first-best rate $\lambda^{FB} = \Delta/\bar{K}$. Without credible disclosure, an adopter captures only the share $1 - \alpha$ of the value it creates. In Regimes II and III, adopters and washers receive the same price for S, so the adopter's advantage is the washer's expected sanction e minus its own cost c; under full credibility, it obtains the full premium αΔ, net of c. Enforcement thus raises adoption by making genuine claims credible, which mitigates the underinvestment caused by short-term price pressure (Stein, 1989) and unobservable quality (Akerlof, 1970).

4.3 Measured versus True Adoption

Proposition 3 (Measurement). In equilibrium, (I) s = 0 and R = 0, and vague claims are uninformative; (II) s = 1, $\omega = 1 - \lambda$, and $R = 1/\lambda$; (III) the quantities are given by Eq. (7); and (IV) $s = \lambda$, ω = 0, and R = 1.

$$s = \lambda\alpha\Delta/e, \quad \omega = 1 - e/(\alpha\Delta), \quad R = \alpha\Delta/e. \quad (7)$$

In Regime III, $\partial\lambda/\partial e = 1/\bar{K} > 0$, whereas $\partial s/\partial e = -\alpha\Delta[(1 - \alpha)\Delta - c]/(\bar{K}e^2) < 0$: stronger enforcement raises true adoption and lowers measured substantive claims.

Three implications follow. A count of substantive AI claims overstates adoption whenever enforcement is incomplete, by the factor αΔ/e in Regime III. Because the measure and true adoption move in opposite directions within Regime III, comparisons across settings with different enforcement can produce trends with the wrong sign. Vague claims carry no information, so pooling them with substantive claims adds only noise.

The model also suggests a correction. In Regimes III and IV, where vague claims are observed, $\mu(N) = 0$, so the price premium of a substantive claim is $P(S) - P(N) = \Delta\mu(S)$, and true adoption can be recovered as

$$\lambda = s \cdot \mu(S) = s \cdot [P(S) - P(N)]/\Delta, \quad (8)$$

where Δ can be estimated from the premium of independently verified claims, so that the claim rate is weighted by the credibility that announcement returns assign to claims.

4.4 Residual Mispricing

Let $M = E[(P - V)^2]$ denote expected squared mispricing. Firms sending N in Regimes III and IV, and firms sending S in Regime IV, are correctly priced, while firms pooled at a common belief μ have mispricing variance $\Delta^2\mu(1 - \mu)$. Hence

$$M = \Delta^2\lambda(1 - \lambda) \text{ in Regimes I and II}, \quad M = \Delta^2\lambda\omega \text{ in Regime III}, \quad M = 0 \text{ in Regime IV}. \quad (9)$$

In Regime III, a higher sanction lowers the washing share but raises adoption, which increases the number of genuine claims pooled with false ones, so its net effect on M is ambiguous.

5. The Regulator's Problem

5.1 Welfare

The regulator maximizes welfare, defined as the net value created by adoption minus a mispricing loss and the cost of enforcement:

$$W(q) = \lambda\Delta - \bar{K}\lambda^2/2 - \theta M - \kappa q^2/2. \quad (10)$$

The first two terms equal the integral of $(\Delta - K)/\bar{K}$ over $K \le \lambda\bar{K}$. The weight $\theta \ge 0$ reflects the misallocation of capital and the losses of investors who trade at incorrect prices, and $\kappa q^2/2$ is a convex audit cost. Sanctions and proprietary costs are transfers, to the public purse and to rivals. Without enforcement, Regime I obtains with $\lambda = (1 - \alpha)\Delta/\bar{K}$; the first best $\lambda^{FB} = \Delta/\bar{K}$ cannot be reached through enforcement because of c and α.

5.2 Optimal Audit Intensity

In Regime III, $\lambda = (a + e)/\bar{K}$ with $a = (1 - \alpha)\Delta - c$, $M = \Delta^2\lambda[1 - e/(\alpha\Delta)]$, and $q = e/F$, so W is quadratic in e.

Proposition 4 (Optimal enforcement). Within Regime III, welfare is strictly concave in e if $D \equiv 1 - 2\theta\Delta/\alpha + \kappa\bar{K}/F^2 > 0$, and its unconstrained maximizer is

$$e^* = [\Delta - a - \theta\Delta^2 + \theta\Delta a/\alpha]/D. \quad (11)$$

If $\hat{e} \le e^* < \alpha\Delta$, the best policy with enforcement is $q = e^*/F$; if $e^* \ge \alpha\Delta$, it is full credibility, $q = \alpha\Delta/F$; and if $e^* < \hat{e}$, it lies at the lower boundary of Regime III. The optimal audit intensity q^* is this policy if it yields more welfare than no enforcement, and $q^* = 0$ otherwise.

Proof sketch. Using $\bar{K}\lambda = a + e$, the first-order condition in Regime III is $(\Delta - a - e - \theta\Delta^2 + 2\theta\Delta e/\alpha + \theta\Delta a/\alpha)/\bar{K} - \kappa e/F^2 = 0$, which is linear in e and yields Eq. (11); the second derivative is $-D/\bar{K}$.

The optimal audit intensity falls with κ. A higher sanction makes expected sanctions cheaper, and when $\kappa\bar{K}/F^2 > 1 - 2\theta\Delta/\alpha$, as at the baseline, $q^* = e^*/F$ rises with F rather than falling as in the standard deterrence model. Above a critical κ, q^* jumps to zero, because small expected sanctions only move the market from silence to pooling washing.

5.3 Commitment and Discretion

Proposition 5 (Value of commitment). If the regulator chooses q after firms have adopted and disclosed, it sets $q = 0$, and the unique equilibrium is silence. Commitment to q^* raises welfare by $W(q^*) - W(0) \ge 0$.

Proof sketch. Once adoption and prices are determined, audits no longer affect λ or M, and sanctions are transfers, so the only remaining effect of q is its cost. Anticipating q = 0, firms face e = 0 < c, and Proposition 1 gives silence. The credibility of AI disclosures thus depends on that of the enforcement agency.

5.4 Extensions: Audit Errors and Sanction Caps

First, suppose that an audited genuine claim is wrongly sanctioned with probability $\varepsilon \in [0, 1)$. The expected cost of S for type H becomes $c + \varepsilon e$, and Proposition 1 holds with c replaced by $c + \varepsilon e$. The gain from adoption becomes $(1 - \alpha)\Delta - c + (1 - \varepsilon)e$ in Regime III and $\Delta - c - \varepsilon e$ in Regime IV, which decreases in e, so errors weaken the adoption effect of enforcement (Polinsky & Shavell, 2000). Second, the sanction may be capped, for example by limited liability. Because full credibility requires $q = \alpha\Delta/F$, a low cap makes credible disclosure expensive in audits.

6. Numerical Experiments

6.1 Parameter Values

The model is solved for the baseline values in Table 2. The values are stylized rather than calibrated: they satisfy A1 and A2, place the optimum in the interior of a regime, and make each mechanism visible. With $\Delta = 10$ as the unit of value, $\alpha = 0.5$ gives equal weight to price and fundamental value, and $c = 1$ equals 10% of the value of adoption. With $\bar{K} = 20$, one quarter of firms adopt without enforcement, while first-best adoption is one half and first-best welfare is 2.5. All quantities were computed in Python from the closed forms, and a grid search over 200,001 values of q reproduces Eq. (11).

Table 2. Baseline Parameter Values

Parameter	Symbol	Value	Role in the experiments
Value created by adoption	Δ	10	Scale normalization
Weight on the current price	α	0.5	Short-term price pressure; varied in Figure 4(b)
Proprietary cost of a substantive claim	c	1	10% of Δ; satisfies A2 ($c < 1.25$)
Upper bound of adoption cost	$\bar{K}$	20	$\lambda = 0.25$ without enforcement; satisfies A1
Sanction for a false claim	F	20	Varied in the extension (F = 10, 40)
Welfare weight of mispricing	θ	0.03	Varied in Figure 4(a) ($\theta = 0$, 0.06)
Enforcement cost parameter	κ	40	Varied in Figure 4 ($\kappa = 5$ to 120)
Audit error rate	ε	0	Varied in the extension ($\varepsilon = 0.05$ to 0.20)

6.2 Disclosure Regimes and Measurement Bias

Figure 2 shows how adoption, claims, and measurement bias vary with the expected sanction. The thresholds of Proposition 1 are $e = 1$, $\hat{e} = 1.33$, and $\alpha\Delta = 5$ ($q = 0.05$, 0.067, and 0.25). Below $q = 0.05$, no firm makes a substantive claim and 25% of firms adopt. Once the expected sanction exceeds the proprietary cost, measured adoption jumps to 100% while true adoption remains near 25%. Through Regime III, true adoption rises linearly to 45%, measured claims fall from 100% to 45%, and the washing share falls linearly from 73% to zero; panel (b) shows the measurement ratio $\alpha\Delta/e$ falling from 3.75 to one. An analyst who compared markets

with expected sanctions of 2 and 3 would observe claim rates of 75% and 58% and might infer lower adoption in the second market, although true adoption is 30% and 35%, respectively.

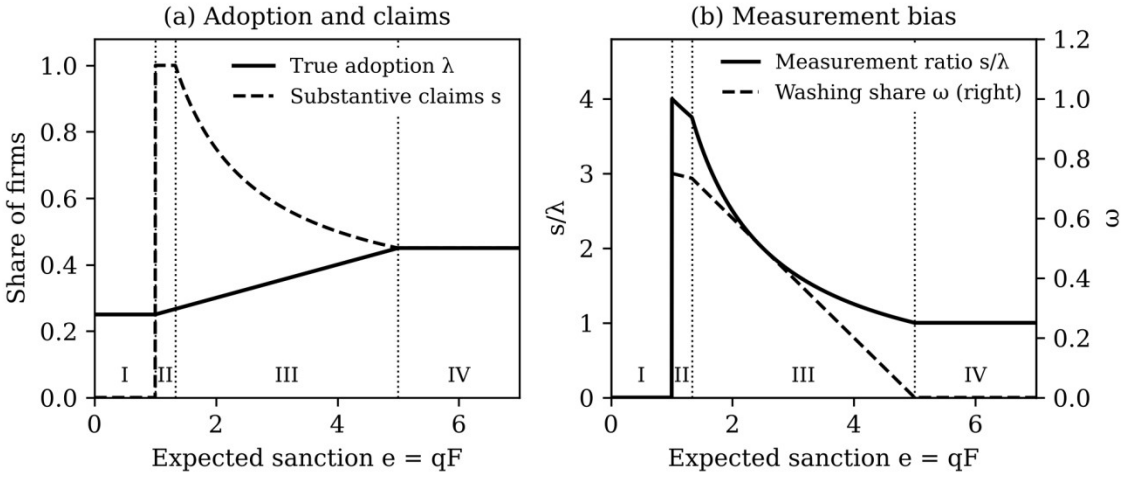


Figure 2. Adoption, Substantive Claims, and Measurement Bias across Disclosure Regimes

6.3 Welfare and Optimal Enforcement

Figure 3 decomposes welfare as a function of audit intensity, and Table 3 reports five enforcement scenarios, where e = qF = 20q. Welfare is not monotone in q. Weak enforcement (q = 0.06) lowers welfare from 1.313 to 1.275 by moving the market from silence to pooling washing, which adds one percentage point of adoption but raises mispricing and enforcement costs. Welfare peaks at q* = 0.15 (e* = 3, as in Eq. (11)), where 35% of firms adopt, 58% make substantive claims, 40% of those claims are false, and welfare is 1.405, 7.0% more than without enforcement. Full credibility (q = 0.25) eliminates washing and raises adoption to 45%, but its enforcement cost of 1.25 lowers welfare to 1.225, below the no-enforcement level. Under discretion the regulator would not audit (Proposition 5), so the value of commitment is 0.0925.

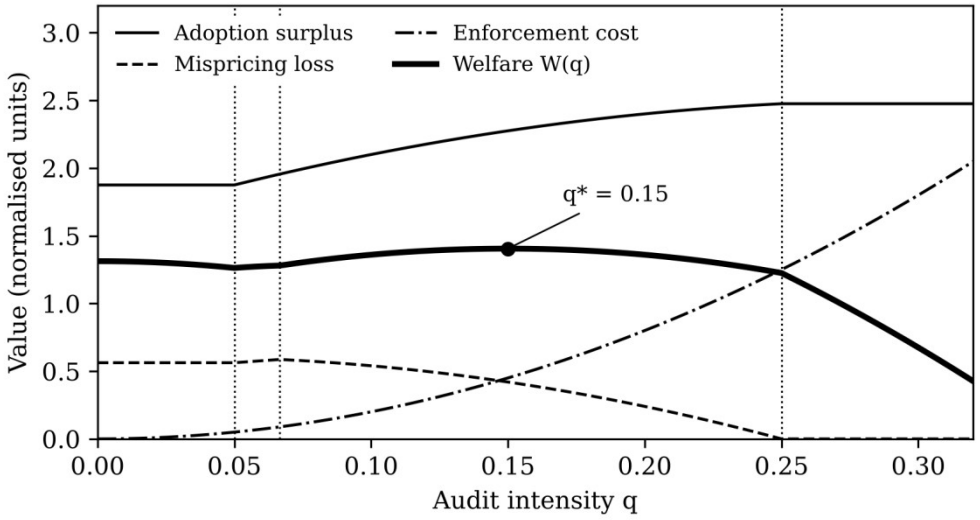


Figure 3. Welfare Components as Functions of Audit Intensity

Table 3. Equilibrium Outcomes under Alternative Enforcement Scenarios

Outcome	No enforcement	Weak	Moderate	Optimal	Full credibility
Audit intensity q	0	0.06	0.10	0.15	0.25
Regime	I Silence	II Pooling	III Partial	III Partial	IV Truthful
Washing probability σ_L	0	1	0.643	0.359	0
True adoption λ	0.250	0.260	0.300	0.350	0.450
Substantive claims s	0.000	1.000	0.750	0.583	0.450
Washing share ω	–	0.740	0.600	0.400	0.000
Measurement ratio s/λ	0.00	3.85	2.50	1.67	1.00
Adoption surplus	1.875	1.924	2.100	2.275	2.475
Mispricing loss θM	0.563	0.577	0.540	0.420	0.000
Enforcement cost $\kappa q^2/2$	0.000	0.072	0.200	0.450	1.250
Welfare W	1.313	1.275	1.360	1.405	1.225

6.4 Validation of the Analytical Results

Table 4 reports two numerical checks. First, an exhaustive search evaluated all profiles (σ_H, σ_L) on a 1001 × 1001 grid at one expected sanction per regime, solving adoption as a fixed point, forming beliefs by Bayes' rule and D1, and recording the largest gain from switching messages as the regret. In every case exactly one profile had a regret below 10^{-4}, and it coincided with the closed form. Second, a Monte Carlo simulation of one million firms in each of 20 replications reproduced the closed forms to within 0.0002 (washing shares of 0.7401 and 0.3999 at e = 1.2 and e = 3), with a standard deviation of about 0.0005, or about 0.008 for 4,000 listed firms. The main threat to disclosure-based measures is strategic, not statistical.

Table 4. Validation of the Closed-Form Equilibrium

e	Regime	σ_L		Regret	Adoption λ		Claims s	
		Closed form	Grid		Closed form	Monte Carlo	Closed form	Monte Carlo
0.5	I Silence	0	0.000	0	0.2500	0.2499	0.0000	0.0000
1.2	II Pooling	1	1.000	0	0.2600	0.2599	1.0000	1.0000
3.0	III Partial	0.3590	0.359	3.7×10^{-5}	0.3500	0.3500	0.5833	0.5832
6.0	IV Truthful	0	0.000	0	0.4500	0.4500	0.4500	0.4500

7. Sensitivity Analysis and Implications

7.1 Sensitivity Analysis

Figure 4(a) plots q* against the enforcement cost κ. For θ = 0.03, full credibility is optimal for κ ≤ 25.6; beyond that, q* falls within Regime III until κ ≈ 50.7, where it jumps to zero. With θ = 0.06 and θ = 0, enforcement stops at κ ≈ 60.9 and 45.5. Figure 4(b) shows that when α is low, the optimum is full credibility and measured claims are unbiased; as α rises, full credibility becomes more expensive, the regulator settles for partial washing, and the measurement ratio reaches 2.56 at α = 0.7 when κ = 40. Measured claims are most misleading where the market rewards AI claims most.

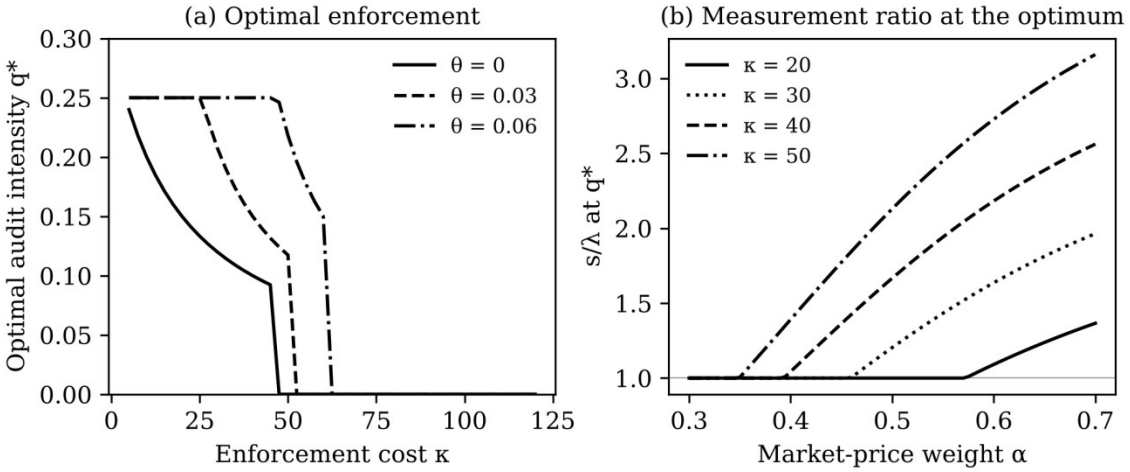


Figure 4. Sensitivity of Optimal Enforcement and Measurement Bias to Enforcement Cost and Price Pressure

Table 5 reports ±20% changes in each parameter. The optimal audit intensity is most sensitive to Δ, $\bar{K}$, and κ: a 20% increase in Δ raises q* from 0.150 to 0.201, and a 20% increase in $\bar{K}$ or κ lowers it to 0.123. The sanction has a threshold effect: a 20% lower F makes enforcement too costly, whereas a 20% higher F raises q* to 0.189 and adoption to 42.7%. The case Δ = 8 violates the first bound in A2, so Regime II disappears, but the optimum remains in Regime III.

Table 5. One-at-a-Time Sensitivity of the Optimal Policy to ±20% Parameter Changes

Parameter (baseline)	−20%				+20%			
	q*	λ*	W*	s/λ	q*	λ*	W*	s/λ
Δ (10)	0.111	0.261	0.937	1.81	0.201	0.451	1.927	1.49
α (0.5)	0.192	0.442	1.676	1.04	0.137	0.287	1.203	2.18
c (1)	0.148	0.358	1.422	1.69	0.152	0.342	1.386	1.64
$\bar{K}$ (20)	0.193	0.491	1.901	1.30	0.123	0.269	1.109	2.04
F (20)	0	0.250	1.313	0.00	0.189	0.427	1.613	1.10
θ (0.03)	0.135	0.335	1.493	1.85	0.169	0.369	1.327	1.48
κ (40)	0.193	0.393	1.521	1.30	0.123	0.323	1.331	2.04

For the extensions of Section 5.4, audit error rates of $\varepsilon = 0.05$, 0.10, and 0.20 lower q^* slightly, to 0.149, 0.147, and 0.143, reduce adoption to 34.1%, 33.3%, and 31.4%, raise the washing share to 42.9% at $\varepsilon = 0.20$, and lower welfare to 1.343, without changing the regime. Sanction caps matter more. With $F = 10$, the optimum is no enforcement; with $F = 40$, full credibility is reached with $q^* = 0.125$, adoption rises to 45%, and welfare rises to 2.16 because the cost of full credibility falls to 0.31.

7.2 Implications for Measuring AI Adoption

Analysts who build AI adoption measures from disclosures should separate substantive claims from vague mentions, interpret claim rates jointly with enforcement, weight claims by their credibility as in Eq. (8), and validate text-based measures against sources that firms do not choose, such as job postings (Acemoglu et al., 2022; Babina et al., 2024) or surveys (McElheran et al., 2024).

7.3 Managerial and Policy Implications

For genuine adopters, the value of disclosure depends on how easily it can be imitated. When enforcement is weak, adopters benefit from disclosures that are costly to imitate, such as quantified results, third-party assurance, or auditable details, which raise the effective cost of a false claim in the same way as e. Regulators should judge enforcement by its effect on adoption and price informativeness rather than by the number of false claims deterred, since 40% of substantive claims remain false at the optimum. Weak enforcement can be worse than none, commitment to an announced enforcement priority creates value that a reactive agency cannot, and higher sanction limits make credible disclosure cheaper to achieve.

7.4 Limitations and Future Research

Types are binary, whereas AI use varies in intensity; investors are assumed to be rational, whereas credulous investors would enlarge the price stake of washing; audits are untargeted; the parameters are stylized; and the model is static. Empirical work could test the central prediction, that claim rates and adoption diverge when enforcement changes, by combining text-based measures with independent adoption data around enforcement actions.

8. Conclusion

This study has developed a game-theoretic model of corporate AI disclosure with endogenous adoption, substantive claims that are verifiable through costly audits, and a regulator that chooses audit intensity. The unique D1 equilibrium moves from silence through pooling and partial washing to truthful disclosure as the expected sanction rises. Measured substantive claims overstate adoption by the ratio of the price stake of a claim to the expected sanction and can fall while adoption rises. The optimal audit intensity has a closed form, tolerates a substantial share of false claims, and requires commitment; in the baseline, it raises adoption from 25% to 35% and welfare by 7.0%. AI disclosures should therefore be read in light of the enforcement environment that produces them, and text-based measures of AI adoption should be adjusted for the credibility of claims.

Acknowledgement

We take responsibility for any remaining errors and omissions.

Author Contributions

Conceptualization, C.C.W. and Z.Q.W.; Methodology, C.C.W. and Z.Q.W.; Formal Analysis, C.C.W. and Z.Q.W.; Investigation, C.C.W.; Data Curation, C.C.W.; Software, C.C.W.; Validation, C.C.W. and Z.Q.W.; Resources, C.C.W. and Z.Q.W.; Visualization, C.C.W.; Writing – Original Draft, C.C.W.; Writing – Review & Editing, C.C.W. and Z.Q.W.; Supervision, Z.Q.W.; Project Administration, C.C.W. and Z.Q.W. We have read and approved the final manuscript and take responsibility for the work.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval and Consent

Not applicable. This study is theoretical and computational and did not involve human participants, personal data, or animals.

Data Availability Statement

No empirical data were used. The parameter values used in the numerical experiments are reported in Table 2. The Python code used to solve the model, to perform the grid-search and Monte Carlo validation, and to generate the figures is available from the corresponding author on reasonable request.

Declaration of Generative AI Use

During the preparation of this manuscript, the authors used Claude (Anthropic) to assist with drafting the text, checking the derivations, and writing the Python code for the numerical experiments and figures. The authors reviewed, verified, and edited all content and take full responsibility for the content of the publication.

References

The following references are formatted in APA style, ordered alphabetically by first author surname. DOI links are provided for all journal articles, and URLs are provided for regulatory documents.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. Journal of Labor Economics, 40(S1), S293–S340. DOI: 10.1086/718327

Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. The Quarterly Journal of Economics, 84(3), 488–500. DOI: 10.2307/1879431

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics, 151, 103745. DOI: 10.1016/j.jfineco.2023.103745

Becker, G. S. (1968). Crime and punishment: An economic approach. Journal of Political Economy, 76(2), 169–217. DOI: 10.1086/259394

Bertomeu, J., & Marinovic, I. (2016). A theory of hard and soft information. The Accounting Review, 91(1), 1–20. DOI: 10.2308/accr-51102

Beyer, A., Cohen, D. A., Lys, T. Z., & Walther, B. R. (2010). The financial reporting environment: Review of the recent literature. Journal of Accounting and Economics, 50(2–3), 296–343. DOI: 10.1016/j.jacceco.2010.10.003

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333–372. DOI: 10.1257/mac.20180386

Cho, I.-K., & Kreps, D. M. (1987). Signaling games and stable equilibria. The Quarterly Journal of Economics, 102(2), 179–221. DOI: 10.2307/1885060

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. Econometrica, 50(6), 1431–1451. DOI: 10.2307/1913390

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. California Management Review, 54(1), 64–87. DOI: 10.1525/cmr.2011.54.1.64

Dranove, D., & Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. Journal of Economic Literature, 48(4), 935–963. DOI: 10.1257/jel.48.4.935

Dye, R. A. (1985). Disclosure of nonproprietary information. Journal of Accounting Research, 23(1), 123–145. DOI: 10.2307/2490910

Fischer, P. E., & Verrecchia, R. E. (2000). Reporting bias. The Accounting Review, 75(2), 229–245. DOI: 10.2308/accr.2000.75.2.229

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574. DOI: 10.1257/jel.20181020

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. The Journal of Law and Economics, 24(3), 461–483. DOI: 10.1086/466995

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. The Quarterly Journal of Economics, 134(4), 2135–2202. DOI: 10.1093/qje/qjz021

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of Accounting and Economics, 31(1–3), 405–440. DOI: 10.1016/S0165-4101(01)00018-0

Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. American Economic Review, 101(6), 2590–2615. DOI: 10.1257/aer.101.6.2590

Kartik, N. (2009). Strategic communication with lying costs. The Review of Economic Studies, 76(4), 1359–1395. DOI: 10.1111/j.1467-937X.2009.00559.x

Kartik, N., Ottaviani, M., & Squintani, F. (2007). Credulity, lies, and costly talk. Journal of Economic Theory, 134(1), 93–116. DOI: 10.1016/j.jet.2006.04.003

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x

Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. Journal of Economics & Management Strategy, 20(1), 3–41. DOI: 10.1111/j.1530-9134.2010.00282.x

Marquis, C., Toffel, M. W., & Zhou, Y. (2016). Scrutiny, norms, and selective disclosure: A global study of greenwashing. Organization Science, 27(2), 483–504. DOI: 10.1287/orsc.2015.1039

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., & Zolas, N. (2024). AI adoption in America: Who, what, and where. Journal of Economics & Management Strategy, 33(2), 375–415. DOI: 10.1111/jems.12576

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. The Bell Journal of Economics, 12(2), 380–391. DOI: 10.2307/3003562

Polinsky, A. M., & Shavell, S. (2000). The economic theory of public enforcement of law. Journal of Economic Literature, 38(1), 45–76. DOI: 10.1257/jel.38.1.45

Seele, P., & Gatti, L. (2017). Greenwashing revisited: In search of a typology and accusation-based definition incorporating legitimacy strategies. Business Strategy and the Environment, 26(2), 239–252. DOI: 10.1002/bse.1912

Spence, M. (1973). Job market signaling. The Quarterly Journal of Economics, 87(3), 355–374. DOI: 10.2307/1882010

Stein, J. C. (1989). Efficient capital markets, inefficient firms: A model of myopic corporate behavior. The Quarterly Journal of Economics, 104(4), 655–669. DOI: 10.2307/2937861

Stocken, P. C. (2000). Credibility of voluntary disclosure. The RAND Journal of Economics, 31(2), 359–374. DOI: 10.2307/2601045

U.S. Securities and Exchange Commission. (2024, March 18). SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence (Press Release No. 2024-36). URL: https://www.sec.gov/newsroom/press-releases/2024-36

U.S. Securities and Exchange Commission. (2025, January 14). SEC charges restaurant-technology company Presto Automation for misleading statements about AI product (Securities Act Release No. 33-11352; Administrative Proceeding File No. 3-22413). URL: https://www.sec.gov/enforcement-litigation/administrative-proceedings/33-11352-s

Verrecchia, R. E. (1983). Discretionary disclosure. Journal of Accounting and Economics, 5, 179–194. DOI: 10.1016/0165-4101(83)90011-3