The Transparency Tax in Strategic Classification: Quantifying the Cost of Fully Targetable Decision Rules

Liz Lemma Future Detective

January 14, 2026

Abstract

Transparency is increasingly mandated for algorithmic decision systems, yet strategic adaptation (Goodhart's law) can cause fully disclosed models to be gamed. Building on strategic classification (Hardt et al.) and subsequent work on opacity/randomness/noise (e.g., Braverman–Garg; Ghalme et al.; Cohen et al., as summarized in Podimata's 2025 survey), we formalize a 2026-relevant policy question: what is the welfare/accuracy cost of requiring decision rules to be fully targetable by agents? We introduce a clean Stackelberg model in which the principal chooses a linear classifier and a graded-disclosure/noise parameter that is publicly auditable but reduces agents' ability to precisely target the boundary. We define the transparency tax as the gap between the optimal equilibrium misclassification error under mandated full transparency (deterministic, fully targetable boundary) and the optimal error when limited graded disclosure is allowed. In a tractable Gaussian latent-skill model with proxy manipulation costs, we derive closed-form best responses, prove a lower bound showing the tax scales with (i) proxy-manipulation opportunity and (ii) low-cost heterogeneity, and provide an explicit graded-disclosure policy achieving a matching upper bound up to constants. Extensions discuss discrete manipulation graphs and alternative disclosure channels (coarse explanations, weight suppression). The results deliver a quantitative object for regulators and a constructive design principle for practitioners: not whether to disclose, but how to disclose in an auditable way that preserves predictive validity in equilibrium.

Table of Contents

2. 2. Related work: strategic classification (Hardt et al.), strategic ERM/PAC, partial information/opacity (Ghalme et al., Cohen et al.), randomness/noise (Braverman–Garg), fairness externalities; position as an information-design/commitment problem with an auditable constraint.

3. 3. Model: latent qualification, observable proxies, manipulation technology/cost heterogeneity, disclosure/noise channel; define equilibrium and loss; define proxy-opportunity index.

4. 4. Equilibrium under mandated full transparency ($\sigma = 0$): characterize best responses, induced false positives/false negatives, and the principal's optimal proxy weight β_T; closed forms in the Gaussian/two-point-cost benchmark.

5. 5. Graded disclosure via auditable noise ($\sigma > 0$): characterize best responses via a smooth acceptance probability; derive manipulation cutoffs and comparative statics; show how graded disclosure restores proxy usefulness.

6. 6. Main theorems: Transparency Tax lower bound (unavoidable under full transparency) and matching upper bound (explicit (β^*, σ^*)). Discuss sharpness and when bounds are tight.

7. 7. Extensions (flagged): (a) multi-dimensional proxies with a low-cost subspace; (b) alternative disclosure channels (coarse thresholds, subset of weights); (c) manipulation graphs with degree/expansion indices. Note where numerical methods are required.

8. 8. Policy and implementation: auditability vs public disclosure; legally acceptable randomness/rounding; guidance for regulators and model builders; discussion of fairness implications (qualitative).

9. 9. Simulations (optional): verify scaling laws under synthetic data; illustrate effect of cost heterogeneity and proxy correlation; brief demonstration on a benchmark dataset (if available).

10. 10. Conclusion and open problems: dynamic learning with repeated interactions; equilibrium selection and strategic experimentation; connections to persuasion and mechanism design.

1 Introduction

Across domains as diverse as credit underwriting, hiring, university admissions, welfare eligibility, and online content moderation, regulators and institutions are converging on a common governance intuition: if automated decisions affect people's lives, then the rules should be *transparent*. By 2026 this intuition is no longer aspirational. Major regimes now require some combination of (i) meaningful explanations to affected individuals, (ii) auditable documentation of model inputs and decision logic, and (iii) advance notice of how scores are constructed and used. The motivating goal is straightforward: transparency can discipline arbitrary discretion, enable legal contestation, and facilitate external oversight.

At the same time, transparency changes the strategic environment. When decision boundaries are legible and stable, agents adapt. They learn which features "count," invest in those features, and, when possible, manipulate the measurement process itself. This is the basic mechanism behind Goodhart's law: when a proxy becomes a target, it ceases to be a good proxy. The modern form of Goodhart's law is not merely behavioral drift; it is *strategic optimization* against a disclosed scoring rule. It shows up in test prep industries that teach to the exam, in search engine optimization that targets ranking signals, in résumé keyword stuffing tailored to automated screeners, in "credit repair" services that exploit the quirks of scoring formulas, and in vendor ecosystems that sell compliance artifacts engineered to satisfy procurement rubrics. In each case, more information about the rule can expand the set of profitable manipulations, potentially degrading the very accuracy and fairness properties that transparency aims to protect.

This paper studies that tension as an economic design problem: a principal must commit to an auditable decision policy, knowing that agents will respond strategically to whatever aspects of the policy are made targetable. Our focal question is not whether transparency is normatively valuable in general, but rather how to quantify one concrete cost of transparency mandates in settings with manipulable proxies. In many high-stakes uses of machine learning, the principal does not directly observe the latent construct that matters (true qualification, ability to repay, risk), but instead relies on measured features that are only imperfectly related to that construct. Some features are "hard-to-manipulate" (e.g., long-run performance history, third-party verified records), while others are easier to manipulate (self-reports, short-term behaviors, presentation, timing, or the feature extraction pipeline itself). Governance debates often treat transparency as orthogonal to statistical performance: reveal the model, and then separately police discrimination or accuracy. Our claim is that, once strategic adaptation is incorporated, transparency can directly and mechanically shift the feasible accuracy frontier.

To make this idea operational, we introduce a metric we call the *trans-*

parency tax. Informally, it is the gap between the best misclassification performance the principal can achieve under a fully transparent, fully targetable policy and the best performance under an alternative regime that is still auditable but less targetable. The economic logic is simple. Under full transparency, the principal may be forced to "play it safe" by downweighting manipulable but informative proxies, or by choosing conservative thresholds, because any strong reliance on a manipulable feature invites cheap gaming near the decision boundary. This self-censorship is a real performance cost: the principal discards predictive information not because it is irrelevant, but because it is strategically unsafe to use when the rule is perfectly legible. Under an appropriately designed opacity mechanism, by contrast, the principal can sometimes retain reliance on informative proxies while reducing the marginal returns to manipulation. The difference in achievable loss is the transparency tax.

Crucially, the opacity mechanism we study is not secrecy about the rule. In governance practice, "black-boxing" is increasingly unacceptable, and often illegal. Instead, we analyze *graded disclosure*: the policy is publicly committed and auditable, but includes a controlled amount of randomness (or equivalently, coarsening/rounding/measurement noise) that is disclosed as a distribution rather than as a realized draw. In our baseline model, the principal uses a linear score built from two proxies and then adds mean-zero noise before thresholding. Agents observe the policy parameters and optimize manipulation of the manipulable proxy. This design captures an increasingly realistic institutional posture: a regulator may demand that the decision process be explainable and externally verifiable, while still allowing documented tie-breaking, randomized audits, stochastic reviews, or privacy-preserving perturbations that prevent perfect "boundary targeting." The key point is that audited randomness can be a *commitment device*: it changes the slope of the acceptance probability with respect to manipulable inputs, thereby reducing the marginal benefit of small manipulations.

Our analysis formalizes two features of the strategic environment that matter for policy. First, manipulation incentives concentrate near the acceptance boundary. When acceptance is deterministic and the boundary is known, an agent who is just below the cutoff has a sharply defined, minimal manipulation that flips the decision. This creates a "manipulability region" whose mass can be large in continuous-feature environments, and whose composition can be systematically adverse (e.g., disproportionately unqualified agents may be clustered near the cutoff in score space). Second, adding a small amount of disclosed noise smooths the acceptance probability. Smoothing imposes an upper bound on the marginal gain from manipulating the proxy: beyond a point, moving the manipulable feature by an extra unit yields only a limited improvement in acceptance probability. When manipulation costs are heterogeneous, this upper bound induces a cutoff: only sufficiently low-cost agents manipulate, and above a threshold cost no one

does. In this sense, graded disclosure can act like a "rate limiter" on gaming: it reduces the extent to which a perfectly known rule can be exploited with infinitesimal adjustments.

These ingredients allow us to articulate a governance-relevant tradeoff. Transparency mandates are often motivated by *procedural* values (contestability, legitimacy, due process). Our model isolates an *outcome* cost that can arise even when the principal's objective is socially aligned with accurate classification. Under full transparency, the principal may rationally choose a rule that is less informative (e.g., it downweights a useful proxy) solely to reduce gaming. This reduction in informativeness is not a failure of optimization; it is the equilibrium response to a more manipulable information environment. Conversely, permitting a limited, auditable form of graded disclosure can reduce gaming and allow the principal to use more information, improving accuracy. The transparency tax measures precisely this wedge.

We emphasize that graded disclosure is not a free lunch. Randomness also introduces intrinsic classification error: even a perfectly qualified individual may be rejected due to the noise realization, and vice versa. Thus the effect of noise on loss is generally non-monotone. Too little noise leaves the boundary targetable; too much noise turns the decision into a lottery. The policy design problem is therefore to choose a level of noise that balances (i) deterrence of strategic manipulation and (ii) the baseline cost of randomization. This balancing is central for governance: regulators may cap permissible randomness (e.g., forbidding "arbitrary" decisions), while principals may be constrained by legal standards that require consistency across similar cases. Our framework accommodates these constraints by restricting the noise parameter to an admissible range and evaluating the best achievable performance within that cap.

Our main conceptual result is that the transparency tax is generically positive whenever two conditions hold: there exists a proxy that is both informative and manipulable, and there exists a nontrivial mass of agents with sufficiently low manipulation costs. In such environments, a fully transparent rule creates a profitable manipulation region that the principal must defensively manage, which in turn forces a sacrifice in predictive power. Allowing graded disclosure—while still committing to an auditable policy—can shrink the manipulation region enough that the principal can reintroduce reliance on the informative proxy. In benchmark Gaussian environments with a simple two-point distribution of manipulation costs, we can make this logic quantitative: the tax scales with the mass of low-cost agents and with an explicit *proxy-opportunity index* capturing how much probability mass lies within manipulable distance of the decision boundary under the best transparent rule. This index can be interpreted as a measure of "how much gaming surface area" transparency exposes.

Why does this matter for 2026 governance? Because the regulatory conversation is shifting from whether to regulate algorithms to how to opera-

tionalize accountability without inadvertently encouraging strategic evasion. In labor markets, disclosure of screening criteria may advantage sophisticated applicants and intermediaries, potentially worsening inequality even as it improves procedural fairness. In lending, revealing feature weights can seed an ecosystem of targeted "score inflation" services that increase default risk and undermine safety-and-soundness goals. In public benefits, deterministic and legible eligibility rules can be exploited in ways that divert resources from intended recipients, while also inviting political backlash when the system appears "gameable." More broadly, as AI systems become embedded in adversarial environments (spam, fraud, cyber abuse), transparency can change the effective threat model. A policy toolkit that treats *auditable randomization* as legitimate—when properly bounded and disclosed—may therefore be an essential complement to transparency mandates, not an exception that undermines them.

We do not claim that adding noise is always desirable, nor that transparency should be weakened categorically. Rather, the model clarifies a specific mechanism by which "more transparency" can reduce accuracy *in equilibrium*, and it characterizes when a limited, rule-bound form of opacity can improve outcomes. The broader message is that transparency and performance are linked through strategic response, so governance must reason about them jointly. The transparency tax offers a compact way to measure what is at stake: it converts an abstract concern ("Goodhart's law") into a comparative-statics object that depends on manipulability, proxy informativeness, and the feasible scope of graded disclosure.

The remainder of the paper builds this argument in a sequence that mirrors the intuition above: we first formalize the model of manipulable and hard-to-manipulate proxies under an auditable linear scoring rule with optional graded disclosure; we then characterize agent best responses under both deterministic and noisy acceptance; and finally we analyze the principal's optimal policy and derive lower and upper bounds on the transparency tax under benchmark distributions. We close by discussing how the same logic extends beyond continuous proxies to richer "manipulation graphs," where gaming corresponds to moving along feasible edges in feature space, and where the structure of reachability shapes the magnitude of the tax.

2 Related Work

Our analysis sits at the intersection of strategic classification, robust learning in the presence of agents who respond to deployed models, and information design under institutional constraints. The common thread is that prediction rules are not evaluated on a fixed data-generating process: once a rule is deployed and understood, the distribution of observed features becomes an equilibrium object. In this section we situate our contribution relative to

(i) the strategic classification and "gaming" literature, (ii) learning-theoretic treatments such as strategic ERM/PAC, (iii) work on partial information and opacity as a policy instrument, (iv) the role of randomness/noise as a deterrence mechanism, and (v) the emerging literature on fairness externalities and heterogeneous manipulation capacity. We close by clarifying how we conceptualize graded disclosure as an *auditable commitment* problem, which is the organizing perspective of the paper.

Strategic classification and gaming responses. A large body of work formalizes Goodhart-style failures by explicitly modeling agents who alter features in response to a classifier. Early and influential formulations treat the interaction as a Stackelberg game: a decision-maker commits to a rule, agents best respond by modifying observed covariates subject to a cost, and the decision-maker's performance is evaluated at the resulting equilibrium distribution. This perspective appears in several strands under the umbrella of *strategic classification* (e.g., **?** and follow-ups), which study how a classifier's choice changes when individuals can manipulate features. Closely related is the literature on *algorithmic recourse*, which asks what actions individuals can take to change an adverse decision and how to design rules that admit feasible and meaningful recourse (e.g., **?**; see also survey work such as **?**). While recourse and strategic classification share an action-cost structure, their normative focus differs: recourse often treats actions as welfare-improving pathways, whereas strategic classification emphasizes that actions may be *purely cosmetic* (proxy manipulation) and can degrade predictive validity.

Within strategic classification, two modeling choices are particularly relevant for our setup. First, many papers posit a *deterministic* decision boundary and study how the principal should choose a classifier anticipating best responses. In such models, manipulability is naturally concentrated near the boundary: small changes in features can flip acceptance, producing discontinuous incentives. Second, several papers distinguish between features that are costly to manipulate (or not manipulable) and those that are easy to change, motivating designs that rely more heavily on "stable" attributes. Our baseline model builds on these foundations but emphasizes a governance-relevant constraint: policies must remain *auditable*, so secrecy about the rule is not an available instrument. This constraint is central to why we focus on a particular form of opacity that remains consistent with transparency mandates.

Strategic ERM/PAC and learning in games. A complementary line of work asks how to *learn* predictors when agents respond strategically, often in worst-case or sample-complexity terms. This includes strategic extensions of empirical risk minimization and PAC learning, where the training objec-

tive accounts for a manipulation model or response function, and guarantees are derived for generalization under strategic behavior (e.g., the "strategic PAC" and "strategic ERM" literatures). These papers are valuable for understanding what can be learned reliably when manipulation is present, and they often highlight identification challenges: the principal observes manipulated covariates, not latent intent, and the mapping from unmanipulated to manipulated distributions depends on equilibrium behavior.

Our approach is more mechanism-design oriented than learning-theoretic. We take the statistical environment as given (a latent qualification and noisy proxies) and focus on how a principal should *commit* to a policy under an explicit institutional design space. That said, the learning-theoretic viewpoint motivates our emphasis on simple, interpretable scoring rules (linear in proxies) and on policy parameters that can be documented and audited. It also motivates the comparative-statics lens of a "tax": we ask how a constraint (full transparency) shifts the feasible frontier, rather than how quickly one can learn the optimal unconstrained rule.

Partial information, opacity, and commitment. Several recent papers study how limiting agents' information about the rule can reduce gaming. This includes models where the principal discloses only coarse score categories, provides noisy feedback, withholds feature weights, or otherwise restricts what is revealed to agents about how actions translate into decisions (e.g., work by **?** and **?**, among others). The key economic logic is familiar: reducing information reduces the precision with which agents can target the boundary, thereby reducing manipulation incentives. In many algorithmic contexts, however, opacity as *secrecy* is either infeasible (rules are reverse engineered), normatively contested (due process and contestability), or legally restricted (documentation and explanation requirements). This paper is motivated precisely by this tension: the interesting design space is not "transparent vs. secret," but "transparent and auditable vs. transparent and auditable *with graded disclosure*."

In this sense, our mechanism is closer in spirit to *information design* (Bayesian persuasion and its descendants): the principal chooses an information structure that shapes agents' posterior beliefs and thus their actions. The twist is that our information structure must itself be auditable and policy-legible. We therefore model graded disclosure as a publicly committed distribution over decision-relevant noise, rather than as hidden randomness. This commitment interpretation aligns with governance practice (documented tie-breaking, random audits, stochastic review) and distinguishes our contribution from work that relies on the principal's ability to keep the mapping from features to decisions secret.

Randomness and noise as deterrence mechanisms. Randomization has a long history as a tool in adversarial settings: it can reduce exploitability by preventing exact targeting. In algorithmic decision-making, randomized classifiers and noisy thresholds have been studied both as robustness devices and as ways to trade off incentives against accuracy. Recent theoretical work (e.g., **?** and related papers) formalizes how adding noise to decision rules can limit an adversary's or agent's ability to reliably achieve a desired outcome with small perturbations, often producing bounds on the marginal gain from manipulation. Our model leverages a similar idea but places it in a policy-commitment frame: the principal chooses a noise level σ subject to an admissible cap $\bar{\sigma}$, and agents know the distribution of the noise but cannot condition on its realization. This yields an explicit and governance-interpretable "rate limit" on manipulation: the acceptance probability becomes smooth in the manipulable proxy, which can generate a cutoff in who manipulates when costs are heterogeneous.

A key point of departure from some robustness-oriented work is that we treat randomization as *costly* in baseline classification terms: it introduces intrinsic error even absent manipulation. The principal's problem is therefore not to maximize deterrence, but to balance deterrence against the loss from making decisions partly stochastic. This tradeoff is central to our notion of a transparency tax, because it is precisely what makes graded disclosure a meaningful alternative to full transparency rather than a trivial domination.

Fairness externalities and heterogeneous ability to manipulate. A growing literature emphasizes that strategic responses can interact with fairness in subtle ways. If manipulation costs differ systematically across groups (due to access to coaching, resources, documentation, or intermediaries), then a rule that is nominally group-blind can generate disparate impacts through differential adaptation. Related work studies equilibrium effects of deployed prediction systems on downstream outcomes, including feedback loops and externalities (e.g., **?**; **?**; and subsequent work on equilibrium fairness). Our benchmark cost heterogeneity (a distribution G with a low-cost mass) is deliberately parsimonious, but it is intended to capture exactly the channel highlighted in this literature: the presence of a population that can manipulate cheaply can force the principal to change the rule in ways that affect everyone, not only manipulators. In our model, this shows up as the principal defensively downweighting a manipulable but informative proxy under full transparency, which can increase errors on non-manipulators as well.

We view this channel as a form of *fairness externality*: even if only a subset of agents manipulates, the equilibrium response can alter acceptance for others. While our main objective is accuracy (misclassification), the same externality logic can be applied to group-conditional error rates and to welfare

measures, and we discuss these extensions qualitatively when interpreting comparative statics.

Our contribution and positioning. Relative to the strategic classification literature, our main contribution is not a new manipulation technology per se, but a new *governance-relevant constraint* and an associated welfare-relevant metric. We formalize a setting in which transparency is mandated in the sense that the policy must be publicly committed and auditable, and we ask what is lost when the principal is further constrained to a fully deterministic, perfectly targetable boundary ($\sigma = 0$). The transparency tax quantifies this loss as the gap between the best achievable equilibrium misclassification under full transparency and the best achievable equilibrium misclassification when the principal is allowed a bounded amount of graded disclosure ($\sigma \in [0, \bar{\sigma}]$). This framing is meant to translate a qualitative concern ("Goodhart's law under transparency") into a comparative-statics object that can be tied to primitives: proxy informativeness, the mass of low-cost manipulators, and the institutional admissibility of randomization.

Relative to work on opacity and partial information, our graded-disclosure mechanism is intentionally designed to be compatible with accountability regimes. The principal does not hide the scoring rule; instead, the principal commits to a transparent distribution over randomization. This is why we emphasize *auditable noise*: the policy can be inspected ex ante and validated ex post statistically, even though individuals cannot precisely target the acceptance boundary in any one instance. In this respect, our model is closer to an *information-design-with-commitment* problem than to a secrecy-based security model.

Finally, relative to robustness work on randomized decision rules, we provide an explicitly economic characterization of the equilibrium incentives created by smoothing, and we connect those incentives to the principal's choice of reliance on manipulable proxies. The upshot is a clean mechanism: deterministic transparency creates a sharp manipulability region near the boundary, whereas graded disclosure bounds the marginal return to manipulation and can eliminate manipulation by all but the lowest-cost types. This mechanism is what underlies the positive transparency tax result and the scaling with an explicit proxy-opportunity index in our benchmark calculations.

The next section introduces the model formally: we specify the latent qualification and proxy structure, the manipulation technology and cost heterogeneity, the graded-disclosure channel, and the equilibrium notion we use to evaluate the principal's loss.

3 Model

We study a single-shot classification problem in which a principal (the decision-maker) deploys a rule that is understood by agents and can therefore change their observed behavior. The key tension is that the principal would like to use an informative proxy that agents can manipulate, but making the decision boundary perfectly targetable can induce precisely the kind of proxy gaming that undermines predictive validity.

Latent qualification and labels. Each agent is characterized by a latent qualification $t \in \mathbb{R}$. The principal's normative/ground-truth label is

$$y \;=\; \mathbf{1}\{t \geq 0\} \in \{0, 1\},$$

so that the principal would ideally accept exactly the qualified types. The binary-threshold structure is not essential for our mechanism, but it yields a transparent notion of misclassification and highlights the boundary incentives created by a deployed decision rule.

Observable proxies: a stable feature and a manipulable feature. The principal does not observe t directly. Instead she observes two proxies,

$$z \;=\; t + \eta_z, \qquad p \;=\; t + \eta_p,$$

where η_z and η_p are mean-zero noises that are independent of t and independent of each other. Throughout we allow (t, η_z, η_p) to be sub-Gaussian with variances $(1, \sigma_z^2, \sigma_p^2)$, which is sufficient for our comparative-statics arguments. For closed-form benchmark calculations we will specialize to the Gaussian case. Economically, z represents a *hard-to-manipulate* signal (e.g., a verified credential or a third-party record), while p represents a *manipulable* signal (e.g., a test score that can be coached, a self-reported feature, or a proxy that can be inflated through cosmetic actions).

Two modeling assumptions matter for interpretation. First, both proxies are informative about t absent manipulation, so there is genuine predictive value at stake. Second, only p is directly manipulable in our baseline; this stark asymmetry is a stylized way to capture that some features are institutionally or technologically "sticky," while others are much easier to move.

Principal's policy: a linear score and graded disclosure. The principal commits to a publicly auditable policy consisting of a linear score

$$s \;=\; z + \beta p$$

and a graded-disclosure parameter $\sigma \geq 0$. Operationally, β controls the principal's reliance on the manipulable proxy p. The parameter σ controls the

amount of decision noise disclosed and committed to by the institution. After observing reported features, the principal draws $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of all other variables and accepts iff

$$a \;=\; \mathbf{1}\{s + \varepsilon \geq 0\}.$$

When $\sigma = 0$, the rule is deterministic and therefore perfectly targetable: acceptance occurs exactly when $s \geq 0$. When $\sigma > 0$, the acceptance probability is smooth in the score; crucially, agents know the distribution of ε but do not observe its realization before acting. We interpret this as *auditable randomization*: the institution can document and statistically validate the randomization scheme (e.g., rounding rules, stochastic review, tie-breaking lotteries), even though any given agent cannot condition her action on the realized coin flip.

We impose a regulatory or institutional cap $\bar{\sigma} > 0$ and restrict attention to $\sigma \in [0, \bar{\sigma}]$. This cap encodes that excessive randomness may be legally impermissible, normatively undesirable, or operationally infeasible.

Manipulation technology and heterogeneous costs. After observing (z, p) and the policy (β, σ), the agent may manipulate only the proxy p by choosing an action $\Delta \in \mathbb{R}$, reporting

$$\hat{p} \;=\; p + \Delta.$$

Manipulation is costly: each agent draws a marginal cost parameter $\kappa > 0$ from a distribution G, independent of (t, η_z, η_p), and pays $\kappa|\Delta|$. The absolute value captures that increasing or decreasing p is costly in magnitude, and linearity delivers a sharp characterization of "move just enough" incentives under deterministic thresholds. Cost heterogeneity is central: it captures that some agents have access to coaching, documentation, intermediaries, or slack resources that make proxy movements cheaper. Our benchmark will often be the two-point specification $\kappa \in \{\kappa_L, \infty\}$ with $\mathbb{P}(\kappa = \kappa_L) = \alpha$, which cleanly separates a low-cost strategic mass from a non-manipulating mass.

Timing and information. The interaction is a Stackelberg game:

1. Nature draws $(t, \eta_z, \eta_p, \kappa)$. The agent observes (z, p, κ).
2. The principal commits to (β, σ) (policy is publicly known and auditable).
3. The agent chooses Δ, generating $\hat{p} = p + \Delta$.
4. The principal observes $(z, \hat{p})$, draws $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and sets $a = \mathbf{1}\{z + \beta\hat{p} + \varepsilon \geq 0\}$.

Agents understand the mapping from $(z, \hat{p})$ to acceptance probabilities, but cannot condition on ε. This is the minimal commitment structure needed for graded disclosure to matter: if ε were observable to the agent prior to manipulation, the agent would again be able to target the realized boundary.

Agent utility and best response. Each agent values acceptance at 1. Under $\sigma > 0$, her acceptance probability given (z, p, κ) and action Δ is

$$\mathbb{P}(a = 1 \mid z, p, \Delta; \beta, \sigma) \;=\; \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right),$$

so utility is

$$U(z, p, \kappa; \beta, \sigma, \Delta) \;=\; \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) - \kappa|\Delta|.$$

When $\sigma = 0$, acceptance is deterministic and utility becomes

$$U(z, p, \kappa; \beta, 0, \Delta) \;=\; \mathbf{1}\{z + \beta(p + \Delta) \geq 0\} - \kappa|\Delta|.$$

We denote an optimal manipulation choice by $\Delta^*(z, p, \kappa; \beta, \sigma)$. Under $\sigma > 0$, the smoothness of $\Phi(\cdot)$ implies a standard marginal condition: the benefit of increasing Δ is proportional to the slope of the acceptance probability, while the cost is κ. Under $\sigma = 0$, the benefit is discontinuous at the boundary, yielding "jump" incentives: if manipulation is worthwhile, it is optimal (under a natural tie-break) to move just enough to cross.

Principal loss and the transparency tax. The principal cares about misclassification relative to $y = \mathbf{1}\{t \geq 0\}$. Given the equilibrium manipulation response Δ^*, define the principal's loss as

$$\mathcal{L}(\beta, \sigma) \;=\; \mathbb{P}(a \neq y),$$

where $a = \mathbf{1}\{z + \beta(p + \Delta^*) + \varepsilon \geq 0\}$. The principal chooses (β, σ) to minimize $\mathcal{L}$, subject to $\sigma \in [0, \bar{\sigma}]$. We interpret $\sigma = 0$ as *mandated full transparency* in the strong sense of a fully deterministic and thus perfectly targetable boundary. Graded disclosure corresponds to allowing $\sigma > 0$ while keeping the policy auditable.

To quantify the value of graded disclosure, we define the *transparency tax* as the equilibrium performance gap between the best deterministic transparent policy and the best auditable graded-disclosure policy:

$$\text{Tax} \;=\; \inf_{\beta \in \mathbb{R}} \mathcal{L}(\beta, 0) \;-\; \inf_{\beta \in \mathbb{R},\ \sigma \in [0, \bar{\sigma}]} \mathcal{L}(\beta, \sigma).$$

This object is not about secrecy: both sides of the comparison allow the policy to be fully known. The only difference is whether the institution may commit to a bounded amount of transparent randomization in the acceptance step.

A proxy-opportunity index. A recurring quantity in our analysis is the mass of agents who lie *within manipulable distance* of the acceptance boundary under a deterministic rule. Intuitively, these are precisely the agents for whom perfect targetability creates a profitable "buy acceptance" opportunity, and hence where Goodhart-type distortions concentrate.

Formally, fix a proxy weight $\beta > 0$ and consider the pre-manipulation score $s = z + \beta p$. Under $\sigma = 0$, an agent with $s < 0$ can cross by choosing the minimal action $\Delta_{\min} = -s/\beta$, which costs $\kappa(-s/\beta)$. Since acceptance is worth 1, a low-cost agent manipulates whenever $\kappa(-s/\beta) < 1$, i.e., whenever

$$s \in \Big[-\frac{\beta}{\kappa}, 0 \Big).$$

This motivates the *proxy-opportunity index* for a given cost level κ:

$$\Delta_{\text{proxy}}(\beta; \kappa) \;:=\; \mathbb{P}\left(-\frac{\beta}{\kappa} \le z + \beta p < 0 \right),$$

the probability mass in the manipulability window just below the cutoff. In the two-point benchmark, the relevant index is $\Delta_{\text{proxy}}(\beta; \kappa_L)$, and when we evaluate equilibrium distortions under the best transparent rule we will often write $\Delta_{\text{proxy}} := \Delta_{\text{proxy}}(\beta_T; \kappa_L)$, where β_T is the principal's optimal proxy weight under $\sigma = 0$. This index is "opportunity" rather than "action": it measures how many agents could profitably game if they are in the low-cost group, which is why it naturally scales with the low-cost mass α in our benchmark bounds.

Discussion and scope. Two features of the model are doing the conceptual work. First, reliance on a manipulable proxy (β) creates an incentive gradient in Δ, but the nature of that gradient depends sharply on whether acceptance is deterministic ($\sigma = 0$) or smoothed ($\sigma > 0$). Second, heterogeneity in κ ensures that policies cannot be evaluated only at a representative agent: a small mass of low-cost manipulators can force the principal to defensively reduce β, affecting classification performance for the broader population.

In the next section we take the deterministic transparency benchmark seriously by setting $\sigma = 0$ and characterizing agents' best responses, the induced false-positive and false-negative rates, and the principal's optimal choice of β_T, including closed forms under the Gaussian/two-point-cost benchmark.

4 Equilibrium under mandated full transparency ($\sigma = 0$)

When $\sigma = 0$, acceptance is a deterministic threshold rule: the principal accepts iff the (post-manipulation) score is nonnegative. This is the polar

case of *full targetability*: conditional on $(\beta, 0)$, an agent can compute exactly how far she must move the manipulable proxy p to flip the decision. The equilibrium therefore has a stark "move-just-enough" structure that will be the source of both tractability and distortion.

4.1 Agent best responses: crossing the deterministic boundary

Fix a proxy weight β. (In our baseline where p is positively informative about t, it is without loss to focus on $\beta \geq 0$; negative β both worsens prediction and makes manipulation *decrease* the score.) Let the pre-manipulation score be

$$s \;=\; z + \beta p.$$

Under $\sigma = 0$, an agent who chooses Δ is accepted iff

$$z + \beta(p + \Delta) \;\geq\; 0 \quad\Longleftrightarrow\quad \Delta \;\geq\; -\frac{s}{\beta} \qquad (\beta > 0).$$

If $s \geq 0$, the agent is already accepted and any nonzero Δ only adds cost, so $\Delta^* = 0$. If $s < 0$, the agent can secure acceptance by choosing the *minimal* boundary-crossing action

$$\Delta_{\min}(s) \;=\; -\frac{s}{\beta} \;>\; 0.$$

Because acceptance is worth exactly 1 and costs are linear, any $\Delta > \Delta_{\min}$ is strictly dominated by $\Delta_{\min}$. Thus the problem reduces to a discrete choice: do nothing and be rejected, or pay $\kappa\Delta_{\min}$ and be accepted. With the natural tie-break toward minimal movement, a best response is

$$\Delta^*(z, p, \kappa; \beta, 0) = \begin{cases} 0, & s \geq 0, \\ -\dfrac{s}{\beta}, & s < 0 \text{ and } \kappa\left(-\dfrac{s}{\beta}\right) < 1, \\ 0, & s < 0 \text{ and } \kappa\left(-\dfrac{s}{\beta}\right) \geq 1. \end{cases}$$

Equivalently, for $\beta > 0$ the set of types who manipulate is exactly those with scores in the *manipulability window*

$$s \in \Big[-\frac{\beta}{\kappa}, 0 \Big).$$

This is the core mechanical implication of mandated transparency: the decision boundary becomes a purchasable good for agents sufficiently close to it, and the width of the purchasable region scales linearly with β and inversely with κ.

Two immediate equilibrium implications are worth highlighting.

Bunching at the cutoff. All manipulators choose $\Delta = -s/\beta$, which implies their post-manipulation score satisfies $z + \beta(p + \Delta) = 0$ exactly. Thus the reported proxy $\hat{p}$ exhibits a point mass at the acceptance boundary: the classic "bunching" pattern familiar from tax notch models and strategic test-taking.

Manipulation is driven by geometry, not beliefs. Because the acceptance mapping is deterministic, best responses do not depend on any subtle inference: the only objects that matter are the realized distance to the boundary $-s/\beta$ and the cost κ. This will change sharply once $\sigma > 0$, where marginal incentives depend on the slope of an acceptance probability rather than a jump.

4.2 Acceptance regions and induced error types

Under $\sigma = 0$, the principal's realized decision depends on whether the agent is able and willing to move into acceptance. Let a denote the acceptance outcome in equilibrium.

For an agent with cost κ, the equilibrium acceptance rule can be written directly as a threshold in the *pre*-manipulation score:

$$a \;=\; \mathbf{1}\{s \geq 0\} \;\vee\; \mathbf{1}\left\{s \in \Big[-\frac{\beta}{\kappa}, 0\Big)\right\} \;=\; \mathbf{1}\left\{s \geq -\frac{\beta}{\kappa}\right\},$$

where the last equality uses that if $s \geq 0$ the condition $s \geq -\beta/\kappa$ already holds. Intuitively, low-cost agents *effectively face a shifted cutoff*: they are accepted whenever their unmanipulated score exceeds $-\beta/\kappa$, because they can "buy" the remaining distance to zero.

This representation makes the error decomposition transparent. Misclassification is

$$\mathcal{L}(\beta, 0) \;=\; \mathbb{P}(t \geq 0,\, a = 0) \;+\; \mathbb{P}(t < 0,\, a = 1).$$

Under heterogeneous κ, we can view $\mathcal{L}$ as a mixture over cost types: low-cost agents contribute more false positives (because their cutoff is shifted left), and also fewer false negatives (because some qualified agents with slightly negative scores can manipulate into acceptance). The principal's problem is precisely to choose β to balance (i) the predictive value of loading on p against (ii) the extra false positives induced by making p an effective lever.

4.3 Two-point costs: a clean equilibrium characterization

The benchmark G we will repeatedly use is

$$\kappa \in \{\kappa_L, \infty\}, \qquad \mathbb{P}(\kappa = \kappa_L) = \alpha.$$

Agents with $\kappa = \infty$ never manipulate, so they face cutoff $s \geq 0$. Agents with $\kappa = \kappa_L$ manipulate whenever beneficial and thus face cutoff $s \geq -\beta/\kappa_L$. Therefore equilibrium acceptance is

$$a \;=\; \begin{cases} \mathbf{1}\{s \geq 0\}, & \kappa = \infty, \\ \mathbf{1}\{s \geq -\beta/\kappa_L\}, & \kappa = \kappa_L, \end{cases} \qquad s = z + \beta p.$$

The equilibrium mass of manipulators is also immediate:

$$\mathbb{P}(\Delta^* \neq 0) \;=\; \alpha \cdot \mathbb{P}\Big(s \in \Big[-\frac{\beta}{\kappa_L}, 0\Big)\Big) \;=\; \alpha \cdot \Delta_{\text{proxy}}(\beta; \kappa_L).$$

This expression emphasizes that manipulation is concentrated among agents "just below" the transparent cutoff, and that increasing β enlarges this strategic mass mechanically.

Correspondingly, false positive and false negative *rates* can be written as

$$\mathrm{FP}(\beta) = (1-\alpha)\,\mathbb{P}(s \geq 0 \mid t < 0) + \alpha\,\mathbb{P}\Big(s \geq -\frac{\beta}{\kappa_L} \,\Big|\, t < 0\Big),$$

$$\mathrm{FN}(\beta) = (1-\alpha)\,\mathbb{P}(s < 0 \mid t \geq 0) + \alpha\,\mathbb{P}\Big(s < -\frac{\beta}{\kappa_L} \,\Big|\, t \geq 0\Big),$$

and $\mathcal{L}(\beta, 0) = \frac{1}{2}\mathrm{FP}(\beta) + \frac{1}{2}\mathrm{FN}(\beta)$ when t is symmetric around 0. These formulae make the qualitative trade-off stark: larger β tends to reduce $\mathbb{P}(s < 0 \mid t \geq 0)$ by improving prediction, but it also increases the gap between the non-manipulating cutoff 0 and the manipulating cutoff $-\beta/\kappa_L$, thereby raising acceptance among unqualified low-cost agents.

4.4 Gaussian benchmark: explicit distributions and closed-form components

To obtain closed forms, suppose $t \sim \mathcal{N}(0,1)$, $\eta_z \sim \mathcal{N}(0, \sigma_z^2)$, $\eta_p \sim \mathcal{N}(0, \sigma_p^2)$, independent. Then

$$s \;=\; z + \beta p \;=\; (1+\beta)t + \underbrace{\eta_z + \beta\eta_p}_{=:u_\beta}, \qquad u_\beta \sim \mathcal{N}(0, v_\beta), \;\; v_\beta = \sigma_z^2 + \beta^2\sigma_p^2,$$

with $u_\beta \perp t$. Hence (t, s) is jointly normal with correlation

$$\rho(\beta) \;=\; \frac{\mathrm{Cov}(t,s)}{\sqrt{\mathrm{Var}(t)\mathrm{Var}(s)}} \;=\; \frac{1+\beta}{\sqrt{(1+\beta)^2 + v_\beta}}.$$

For the non-manipulating mass ($\kappa = \infty$), misclassification under the rule $a = \mathbf{1}\{s \geq 0\}$ is exactly the probability that t and s have opposite signs. A standard bivariate-normal identity yields the closed form

$$\mathbb{P}\big(\mathbf{1}\{s \geq 0\} \neq \mathbf{1}\{t \geq 0\}\big) \;=\; \frac{1}{2} - \frac{1}{\pi}\arcsin(\rho(\beta)).$$

This expression makes precise the usual "signal quality" logic: increasing β can increase $\rho(\beta)$ when p is informative, improving accuracy absent gaming.

For the manipulating mass ($\kappa = \kappa_L$), the cutoff is shifted to $s \geq -\beta/\kappa_L$. The relevant error terms are of the form

$$\mathbb{P}(t < 0,\, s \geq c), \qquad \mathbb{P}(t \geq 0,\, s < c), \quad \text{where } c = -\beta/\kappa_L.$$

These are bivariate normal probabilities and can be written compactly using the bivariate normal CDF $\Phi_2(\cdot, \cdot; \rho)$ (or, equivalently, Owen's T function). Concretely, letting $\sigma_s(\beta) = \sqrt{(1+\beta)^2 + v_\beta}$ so that s/σ_s is standard normal, we can express

$$\mathbb{P}(t < 0,\, s \geq c) \;=\; \Phi(0) \;-\; \Phi_2\!\left(0,\; \frac{c}{\sigma_s(\beta)};\; \rho(\beta)\right),$$

and similarly for $\mathbb{P}(t \geq 0,\, s < c)$. While less elementary than the arcsin formula at $c = 0$, these expressions remain one-line and numerically stable, and they let us compute $\mathcal{L}(\beta, 0)$ and β_T exactly in the Gaussian benchmark.

4.5 The principal's optimal transparent weight β_T

Under mandated transparency, the principal solves

$$\beta_T \in \arg\min_{\beta \geq 0}\; \mathcal{L}(\beta, 0),$$

with $\mathcal{L}$ computed under the induced manipulation behavior above. We do not generally obtain a simple closed form for β_T because β affects loss through two distinct channels: (i) statistical fit (via $\rho(\beta)$ and the distribution of s), and (ii) strategic distortion (via the window $[-\beta/\kappa_L, 0)$ and the shifted cutoff for low-cost agents).

What we can characterize sharply, however, is the direction of the distortion relative to the non-strategic benchmark. If agents could not manipulate, the Bayes-optimal classifier (under the Gaussian prior/noise structure) accepts iff the posterior mean $\mathbb{E}[t \mid z, p] \geq 0$. Because that posterior mean is linear in (z, p), this decision rule is equivalent to a linear threshold $z + \beta_{\text{Bayes}} p \geq 0$ with

$$\beta_{\text{Bayes}} \;=\; \frac{\sigma_z^2}{\sigma_p^2}.$$

Under transparency with $\alpha > 0$ and finite κ_L, the principal internalizes that raising β increases the mass of unqualified agents who can cheaply cross the cutoff. This induces a *strategic shrinkage* of the proxy weight:

$\beta_T \;<\; \beta_{\text{Bayes}}$ whenever manipulation is sufficiently prevalent (large α) or sufficiently cheap (small κ

In the extreme, as $\alpha \to 1$ and $\kappa_L \to 0$, any reliance on p renders acceptance almost fully purchasable near the boundary, and the principal optimally retreats toward $\beta_T \approx 0$, effectively screening on the hard-to-manipulate proxy

z alone. Conversely, as $\alpha \to 0$ or $\kappa_L \to \infty$, the strategic constraint vanishes and $\beta_T \to \beta_{\text{Bayes}}$.

This is the sense in which mandated full transparency imposes a real opportunity cost: it does not merely add manipulators on top of an otherwise optimal statistical rule; it pushes the principal away from the best statistical use of an informative proxy. In the next section we show that allowing auditable graded disclosure ($\sigma > 0$) changes the geometry of incentives—replacing a discrete jump with a bounded marginal gain—and can therefore relax this strategic shrinkage while remaining fully policy-transparent.

5 Graded disclosure via auditable noise ($\sigma > 0$)

When $\sigma > 0$, acceptance is no longer a knife-edge event that can be guaranteed by crossing a deterministic boundary. Instead, the principal commits to a *distribution* over effective cutoffs through the additive noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, which is publicly known but not observed at the time of manipulation. This graded-disclosure regime preserves auditability (the policy is still fully specified by (β, σ)) while making the decision *less targetable*: moving p upward shifts acceptance probabilities smoothly rather than flipping acceptance with certainty.

5.1 Smooth incentives: from "buying acceptance" to "buying probability"

Fix (β, σ) with $\beta > 0$ and $\sigma > 0$. Given observed (z, p, κ), an agent who chooses Δ obtains acceptance probability

$$\mathbb{P}(a = 1 \mid z, p, \Delta; \beta, \sigma) = \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right),$$

and therefore solves

$$\max_{\Delta \in \mathbb{R}} \ \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) - \kappa|\Delta|.$$

Relative to $\sigma = 0$, the crucial change is that the marginal benefit of manipulation is governed by the *slope* of the probit link. Differentiating on each side of $\Delta = 0$ yields, for $\Delta > 0$,

$$\frac{\partial}{\partial \Delta} \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) = \frac{\beta}{\sigma} \phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right),$$

so the agent trades off a smooth marginal gain against the constant marginal cost κ. Since $\phi(\cdot) \le 1/\sqrt{2\pi}$, the marginal gain is *uniformly bounded*:

$$\sup_{\Delta} \frac{\partial}{\partial \Delta} \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) \ = \ \frac{\beta}{\sigma\sqrt{2\pi}}.$$

This bound is the mechanical force behind graded disclosure: unlike the deterministic jump of size 1, the largest possible marginal return to moving p is finite, and it falls as σ rises.

5.2 Best responses and the manipulation cutoff in κ

The bounded marginal return immediately implies a clean deterrence condition. If an agent's marginal cost exceeds the maximum marginal gain, then manipulation is never optimal:

$$\kappa > \frac{\beta}{\sigma\sqrt{2\pi}} \quad \Longrightarrow \quad \Delta^*(z, p, \kappa; \beta, \sigma) = 0 \text{ for all } (z, p).$$

Intuitively, even if the agent could move the score to the steepest point of the acceptance curve (where the slope is maximized), the probability increase per unit Δ would still not justify the cost.

When $\kappa \leq \beta/(\sigma\sqrt{2\pi})$, some agents do manipulate, but the structure differs sharply from "move just enough." For $\Delta > 0$, any interior optimum must satisfy the first-order condition

$$\frac{\beta}{\sigma}\phi\left(\frac{z + \beta(p + \Delta^*)}{\sigma}\right) = \kappa.$$

A useful way to read this condition is that the agent chooses a *post-manipulation standardized score*

$$x^*(\kappa; \beta, \sigma) \text{ such that } \phi(x^*) = \frac{\kappa\sigma}{\beta},$$

and then moves just enough to reach that target, provided reaching it requires $\Delta > 0$. Concretely, if x^* is selected on the relevant (typically nonnegative) branch,[1] then

$$\Delta^*(z, p, \kappa; \beta, \sigma) = \max\left\{0, \frac{\sigma x^*(\kappa; \beta, \sigma) - \big(z + \beta p\big)}{\beta}\right\}.$$

This expression makes two points transparent. First, the *identity* of manipulators depends jointly on costs and baseline score: low κ agents manipulate more often, and among a given κ, manipulation is concentrated among those whose initial score $s = z + \beta p$ lies sufficiently below the cost-dependent target σx^*. Second, conditional on manipulating, agents typically do *not* all land

[1] Because ϕ is symmetric and unimodal, $\phi(x) = c$ has two solutions $\pm x$ when $c < 1/\sqrt{2\pi}$. The global optimum selects the branch consistent with the direction of manipulation and the fact that additional movement into the far tails has sharply diminishing marginal returns. For our baseline with $\beta > 0$, profitable manipulation is upward ($\Delta \geq 0$) and the relevant solution corresponds to pushing the score toward (and possibly beyond) the region where acceptance probability is already high.

exactly at the same cutoff. In particular, there is no analogue of the deterministic bunching at $s = 0$; instead, manipulation tends to push agents into a neighborhood where marginal returns are commensurate with κ, producing a smoother distortion of the score distribution.

5.3 Comparative statics: how σ disciplines strategic behavior

The cutoff $\beta/(\sigma\sqrt{2\pi})$ delivers immediate comparative statics that match the policy intuition.

Noise reduces manipulation. Holding β fixed, increasing σ lowers the maximum slope of the acceptance curve and therefore weakly shrinks the set of cost types who ever manipulate. In particular, for any cost distribution G,

$$\mathbb{P}(\Delta^* \neq 0) \;\leq\; G\left(\frac{\beta}{\sigma\sqrt{2\pi}}\right),$$

up to the additional requirement that the realized score is low enough to make a positive Δ valuable. In the two-point benchmark $\kappa \in \{\kappa_L, \infty\}$, this becomes stark: either σ is large enough to deter *all* low-cost manipulation (if $\kappa_L > \beta/(\sigma\sqrt{2\pi})$), or else the entire low-cost mass is potentially active.

Proxy weight increases both usefulness and manipulability. Holding σ fixed, increasing β increases the marginal gain from improving p and therefore (i) expands the set of costs for which manipulation is privately worthwhile and (ii) increases the optimal Δ^* for those who manipulate. This is the smooth analogue of the transparency case: loading more on a manipulable proxy makes the proxy more "valuable to move." The difference is that graded disclosure turns this into a *bounded* incentive effect that can be offset by increasing σ.

Cheaper manipulation increases distortion, but in a continuous way. Lower κ raises the target acceptance slope the agent is willing to pay for, which corresponds to pushing the post-manipulation score further into regions where acceptance probability is high. Thus, conditional on being below the relevant target, low-cost agents choose larger Δ^*. Unlike the deterministic setting, however, these effects scale smoothly with κ; there is no discrete "buy acceptance for sure" region.

5.4 How graded disclosure restores proxy usefulness

From the principal's perspective, σ creates a new instrument that relaxes the strategic shrinkage motive. Under full transparency, increasing β mechanically widens a purchasable window near the cutoff. Under graded disclo-

sure, increasing β does raise incentives, but the principal can simultaneously choose σ to keep the marginal return to manipulation below the relevant cost levels.

This logic is especially clean in the two-point cost benchmark. Suppose a fraction α of agents have finite cost κ_L, and the remainder never manipulate. If the principal selects (β, σ) satisfying

$$\kappa_L > \frac{\beta}{\sigma\sqrt{2\pi}},$$

then even the lowest-cost strategic agents optimally choose $\Delta^* = 0$. In that region of the policy space, equilibrium play coincides with the *no-gaming* environment: the only effect of σ is the principal's own randomized tie-breaking around the cutoff, not endogenous feature distortion. This is precisely the channel through which graded disclosure can "restore" the value of p: once manipulation is deterred, the principal can safely raise β toward the statistically optimal weight (the Bayes coefficient in the Gaussian benchmark) without inducing a corresponding surge in strategic false positives.

Of course, σ is not a free lunch. Even if it deters manipulation, it introduces intrinsic classification randomness by occasionally flipping decisions near the boundary. Thus, for fixed β, the loss $\mathcal{L}(\beta, \sigma)$ typically reflects a balance between two forces: larger σ reduces strategic distortion (a first-order gain when manipulation is prevalent) but increases baseline noise in the accept/reject rule (a second-order cost when the score is already well-separated). The principal's optimum σ^* is therefore generally interior when manipulation is possible, and it is pinned at $\sigma^* = 0$ only when either manipulation is absent (e.g., $\alpha = 0$ or κ_L very large) or the regulatory cap $\bar{\sigma}$ is too tight to meaningfully affect incentives.

The key takeaway for what follows is that σ changes the geometry of best responses: it replaces a discontinuous, fully targetable threshold with a smooth acceptance probability whose slope is bounded. This single technical fact yields a policy-relevant conclusion: even a *small*, auditable amount of graded disclosure can eliminate manipulation by a nontrivial set of cost types, allowing the principal to increase reliance on an informative but manipulable proxy. In the next section, we formalize this trade-off in the main transparency-tax bounds: a lower bound showing that full transparency induces an unavoidable accuracy loss when proxies are both useful and gameable, and a matching upper bound showing that an explicit graded-disclosure policy (β^*, σ^*) recovers (up to constants) the lost accuracy.

6 Main results: an unavoidable transparency tax and a matching graded-disclosure remedy

We now formalize the central claim of the paper: when a proxy is both (i) *predictively useful* for y and (ii) *cheaply manipulable* for a nontrivial mass of agents, then mandated full transparency ($\sigma = 0$) imposes an accuracy loss that cannot be eliminated by tuning the proxy weight β. Moreover, the loss is not merely existential: in a benchmark Gaussian environment we can lower bound it by an explicit "proxy-opportunity" index, and we can match that lower bound (up to constants) with a constructive graded-disclosure policy (β^*, σ^*) that respects the regulatory cap $\sigma \le \bar{\sigma}$.

Throughout this section we focus on the benchmark that makes the trade-off sharp while remaining transparent to compute: $t \sim \mathcal{N}(0,1)$, $\eta_z \sim \mathcal{N}(0, \sigma_z^2)$, $\eta_p \sim \mathcal{N}(0, \sigma_p^2)$ independent, and a two-point manipulation-cost distribution $\kappa \in \{\kappa_L, \infty\}$ with $\mathbb{P}(\kappa = \kappa_L) = \alpha$. This isolates the key force: a fraction α of agents can manipulate at constant marginal cost κ_L, while the rest are non-strategic.

6.1 A proxy-opportunity index

A recurring object in our bounds is the amount of probability mass that lies *within manipulable distance of the acceptance boundary* under a transparent rule. Under $\sigma = 0$ and $\beta > 0$, low-cost agents with baseline score $s = z + \beta p < 0$ can guarantee acceptance by choosing the minimal crossing action $\Delta_{\min} = -s/\beta$, and they do so whenever $\kappa_L \Delta_{\min} < 1$, i.e.,

$$s \in \Big[-\frac{\beta}{\kappa_L}, 0 \Big).$$

This "purchasable strip" is the transparent analogue of a margin: the thicker the strip, the more agents can flip the decision at bounded cost.

Motivated by this, we define a proxy-opportunity index that captures the *density of agents near the cutoff* scaled by the manipulable strip width. One convenient choice (sufficient for our results) is

$$\Delta_{\text{proxy}}(\beta) \;:=\; \mathbb{P}\Big(s \in \Big[-\frac{\beta}{\kappa_L}, 0 \Big) \Big), \qquad s = z + \beta p,$$

evaluated under the *no-manipulation* distribution of (z, p). Under the Gaussian benchmark, s is itself Gaussian, so $\Delta_{\text{proxy}}(\beta)$ is explicit:

$$s \sim \mathcal{N}(0,\ \mathrm{Var}(s)), \qquad \mathrm{Var}(s) = \mathrm{Var}(z) + \beta^2 \mathrm{Var}(p) + 2\beta\, \mathrm{Cov}(z, p),$$

and thus $\Delta_{\text{proxy}}(\beta) = \Phi\big(0/\sqrt{\mathrm{Var}(s)}\big) - \Phi\big(-\beta/(\kappa_L \sqrt{\mathrm{Var}(s)})\big) = \Phi\big(\beta/(\kappa_L \sqrt{\mathrm{Var}(s)})\big) - \frac{1}{2}$. For small widths this is approximately linear:

$$\Delta_{\text{proxy}}(\beta) \;\approx\; \frac{\beta}{\kappa_L} \cdot f_s(0) \;=\; \frac{\beta}{\kappa_L} \cdot \frac{1}{\sqrt{2\pi\, \mathrm{Var}(s)}}.$$

We emphasize two interpretation points. First, Δ_{proxy} is larger when κ_L is smaller (cheaper gaming) and when the score distribution has substantial mass near the cutoff (a common feature of selective decisions). Second, because $\text{Var}(s)$ depends on β, the index automatically accounts for the fact that re-weighting the proxy changes not only incentives but also the geometry of the score distribution.

6.2 Lower bound: full transparency forces strategic shrinkage

Under mandated transparency, the principal's only lever is β, and β faces a tension. If β is large, the score uses the informative proxy p aggressively, which is statistically desirable in the no-gaming world. But the same large β enlarges the purchasable strip $[-\beta/\kappa_L, 0)$, and hence expands strategic acceptance among low-cost agents who were initially just below the boundary. If β is small, manipulation incentives weaken, but the principal leaves predictive power on the table.

The next theorem makes this trade-off quantitative.

Theorem 1 (Transparency tax lower bound). *In the Gaussian/two-point-cost benchmark, suppose $\sigma_p^2 > 0$ (so p contains independent information about t beyond z) and $\alpha > 0$. Let $\beta_T \in \arg\min_\beta \mathcal{L}(\beta, 0)$ be an optimal transparent weight. Then there exist universal constants $c_1, c_0 > 0$ such that*

$$\mathcal{L}(\beta_T, 0) \;\geq\; \mathcal{L}_{\text{no-gaming}}(\beta_{\text{Bayes}}) \;+\; c_1\, \alpha\, \Delta_{\text{proxy}},$$

where β_{Bayes} is the Bayes-optimal linear weight in the no-manipulation environment (equivalently, the coefficient implied by $\mathbb{E}[t \mid z, p]$), and Δ_{proxy} is an explicit proxy-opportunity term (e.g., $\Delta_{proxy} = \Delta_{proxy}(\tilde{\beta})$ for some $\tilde{\beta}$ in a neighborhood of β_{Bayes}). In particular, for any $\bar{\sigma} > 0$, the transparency tax satisfies $\text{Tax} > 0$.

Proof sketch and economic content. The proof has a "minimax" structure: we show that any transparent β must pay *either* a statistical price (from under-using the proxy) *or* a strategic price (from induced manipulation). Concretely, consider two regimes.

(i) Large β: If β is large enough to meaningfully approach β_{Bayes}, then a nontrivial mass of low-cost agents satisfy $s \in [-\beta/\kappa_L, 0)$. These agents flip from rejection to acceptance under manipulation, generating strategic acceptances that are not aligned with y. Because manipulation is concentrated just below the cutoff, it increases the acceptance rate precisely where the classifier is most error-prone; in misclassification terms, it induces an $\Omega(\alpha\, \Delta_{\text{proxy}}(\beta))$ contribution to error (the exact decomposition depends on whether we track the resulting false positives, false negatives, or both, but the core is that a constant fraction of these flips are "wrong" under the benchmark symmetry).

(ii) Small β: If β is reduced to suppress manipulation, the classifier becomes closer to one that uses only z. Under our proxy-usefulness condition (formally, $\mathrm{Var}(\mathbb{E}[t \mid z, p]) > \mathrm{Var}(\mathbb{E}[t \mid z])$), this incurs a strictly positive statistical regret relative to $\mathcal{L}_{\text{no-gaming}}(\beta_{\text{Bayes}})$. In the Gaussian case, this gap can be bounded below using standard comparison arguments for linear separators: the Bayes weight delivers a strictly larger signal-to-noise ratio than any β constrained to be too small.

The lower bound follows by combining these: any β that avoids the strategic term must be small enough to suffer the statistical term, and any β that avoids the statistical term must be large enough to trigger the strategic term. The proxy-opportunity index enters through an anti-concentration bound for s: under Gaussianity, the probability of lying in an interval of width β/κ_L around the cutoff is proportional to that width times the density at the cutoff, yielding a linear-in-β/κ_L component that cannot be "optimized away".

6.3 Upper bound: an explicit graded-disclosure policy that recovers the loss

The second theorem shows that the lower bound is essentially tight: a simple graded-disclosure policy can remove (most of) the strategic term while permitting the principal to use a more statistically efficient β.

Theorem 2 (Matching upper bound via graded disclosure). *In the same benchmark, assume $\bar{\sigma} > 0$. Then there exist (β^*, σ^*) with $\sigma^* \in (0, \bar{\sigma}]$ and $\beta^* > \beta_T$ such that*

$$\mathcal{L}(\beta^*, \sigma^*) \;\le\; \mathcal{L}(\beta_T, 0) \;-\; c_2\, \alpha\, \Delta_{\text{proxy}},$$

for a universal constant $c_2 > 0$ (matching c_1 up to constant factors). One sufficient construction is to choose σ^* and β^* satisfying the deterrence inequality

$$\kappa_L \;>\; \frac{\beta^*}{\sigma^* \sqrt{2\pi}},$$

so that $\Delta^*(z, p, \kappa_L; \beta^*, \sigma^*) = 0$ for all (z, p), and then set β^* close to β_{Bayes} subject to this constraint and $\sigma^* \le \bar{\sigma}$.

Proof sketch and economic content. The key step is to use the bounded-slope property of the probit acceptance function under $\sigma > 0$. When $\kappa_L > \beta/(\sigma\sqrt{2\pi})$, even the cheapest agents never find it profitable to move p, so the equilibrium coincides with the no-manipulation world. In that case, the only cost of σ is the principal's own randomized tie-breaking near the boundary. Because that intrinsic randomness affects only agents whose (post-manipulation) scores lie within $O(\sigma)$ of the cutoff, it can be made small by choosing σ^* small—yet still large enough (relative to β^*) to shut down manipulation incentives.

Operationally, we pick σ^* at (or near) the smallest value allowed by the deterrence inequality given a target β^*, and then pick β^* as large as possible (ideally near β_{Bayes}) given the cap $\bar{\sigma}$. The improvement over transparency comes from two sources that move in opposite directions under σ: (i) strategic error falls discontinuously to zero once manipulation is deterred, while (ii) intrinsic noise error rises smoothly with σ. This creates room for a net gain, and the gain scales with the mass of agents who would otherwise have been in the purchasable strip—exactly $\alpha\,\Delta_{\text{proxy}}$.

6.4 Sharpness, tightness, and when the bounds bind

We view Theorems 1 and 2 as a matched characterization rather than an artifact of the benchmark. The linear dependence on α is tight: if $\alpha \to 0$, strategic behavior vanishes and so does the transparency tax. The dependence on κ_L through the strip width β/κ_L is also tight in the transparent regime: under $\sigma = 0$, the best response is literally characterized by whether an agent can afford the minimal crossing action, so the set of strategic movers expands linearly in $1/\kappa_L$ near the cutoff.

Where the constants (and the extent of tightness) matter is the role of the cap $\bar{\sigma}$. If $\bar{\sigma}$ is large enough that the principal can satisfy $\kappa_L > \beta_{\text{Bayes}}/(\bar{\sigma}\sqrt{2\pi})$, then graded disclosure can both deter manipulation *and* set $\beta^* \approx \beta_{\text{Bayes}}$, making the upper bound especially sharp: the remaining gap to the no-gaming optimum is then driven primarily by the intrinsic randomization, which can be made small. If instead $\bar{\sigma}$ is very tight, the principal may be unable to deter manipulation at Bayes-like β. In that case graded disclosure still helps (by shrinking marginal incentives even when it does not eliminate them), but the simple "full deterrence" construction becomes conservative; the true optimum may involve partial deterrence and an interior trade-off.

Finally, we stress what our theorems do *not* claim. We are not arguing that adding noise is always good: if manipulation costs are high or the proxy is uninformative, the optimal σ^* is indeed zero and the tax disappears. Rather, the message is conditional and policy-relevant: when transparency makes a manipulable proxy *too targetable*, auditable graded disclosure can strictly improve accuracy while remaining compatible with oversight, because it changes the incentive geometry from "buying acceptance" to "buying probability" with bounded marginal returns.

7 Extensions and robustness (flagged)

Our baseline analysis intentionally collapses the strategic channel into a single manipulable scalar proxy p, additive actions $\Delta \in \mathbb{R}$, and a particularly tractable disclosure instrument (additive Gaussian noise ε). In practice, however, (i) models typically use *many* proxies, (ii) institutions often have a menu of disclosure/opacity levers besides explicit randomization,

and (iii) manipulation is frequently *combinatorial* (rewriting a resume, rebundling transactions, creating a portfolio of signals) rather than additive. In this section we sketch three extensions that preserve the same economic logic—transparency makes a boundary targetable, graded disclosure reduces marginal returns to targeting—while clarifying what does and does not continue to admit closed forms. We flag where numerical methods become unavoidable.

7.1 (a) Multi-dimensional proxies with a low-cost manipulable subspace

Let the proxy become $p \in \mathbb{R}^d$ with score

$$s \;=\; z + \beta^\top p, \qquad \beta \in \mathbb{R}^d,$$

and allow the agent to choose a vector manipulation $\Delta \in \mathbb{R}^d$ so that $\hat{p} = p + \Delta$. A natural way to formalize "a low-cost subspace" is to partition coordinates into *hard* and *soft* components, $p = (p_H, p_M)$, and to assume manipulation is feasible only (or cheaply only) on p_M. For instance, take a separable linear cost

$$\text{cost}(\Delta) \;=\; \kappa\|\Delta_M\|_1 \;+\; \infty \cdot \mathbf{1}\{\Delta_H \neq 0\},$$

or more generally $\kappa\|\Delta_M\|$ for a norm $\|\cdot\|$ capturing the "technology" of manipulation. This formulation makes two points transparent.

First, only the projection of the classifier onto the manipulable subspace matters for incentives. Writing $\beta = (\beta_H, \beta_M)$, an agent's acceptance probability under $\sigma > 0$ is $\Phi((z + \beta_H^\top p_H + \beta_M^\top(p_M + \Delta_M))/\sigma)$. The marginal gain from moving along any direction u in the manipulable subspace is bounded by the maximal slope of the probit link times $|\beta_M^\top u|$. Consequently, deterrence conditions generalize from a scalar bound to a bound on β_M. Under an ℓ_2 cost $\kappa\|\Delta_M\|_2$, a sufficient no-manipulation condition is

$$\kappa \;>\; \frac{\|\beta_M\|_2}{\sigma\sqrt{2\pi}},$$

since the directional derivative of $\Phi(\cdot)$ is at most $1/(\sigma\sqrt{2\pi})$ and the best direction aligns with β_M. Under ℓ_1 cost, the relevant quantity is $\|\beta_M\|_\infty$ instead.

Second, the transparent "purchasable strip" becomes a *purchasable slab* whose thickness depends on distance to the hyperplane measured in the manipulation norm. Under $\sigma = 0$, a low-cost agent who is rejected at baseline ($s < 0$) can cross by solving

$$\min_{\Delta_M} \; \|\Delta_M\| \quad \text{s.t.} \quad z + \beta^\top p + \beta_M^\top \Delta_M \geq 0.$$

The minimal manipulation cost is the dual norm distance:

$$\|\Delta_M\|_{\min} \;=\; \frac{(-s)_+}{\|\beta_M\|_*},$$

where $\|\cdot\|_*$ is the dual norm (e.g., $\|\beta_M\|_2$ dual to ℓ_2, $\|\beta_M\|_\infty$ dual to ℓ_1). The analogue of the strip condition $\kappa_L \Delta_{\min} < 1$ becomes

$$s \in \Big[-\frac{\|\beta_M\|_*}{\kappa_L},\, 0\Big),$$

so the same geometry reappears with β replaced by the "effective manipulable weight" $\|\beta_M\|_*$. This yields a natural multi-dimensional proxy-opportunity index

$$\Delta^{(d)}_{\text{proxy}} \;:=\; \mathbb{P}\Big(s \in \Big[-\frac{\|\beta_M\|_*}{\kappa_L},\, 0\Big)\Big) \quad \text{(under no manipulation)}.$$

When (z, p) are jointly Gaussian, s is still Gaussian for any fixed β, so this particular index remains explicit. What ceases to be explicit in general is the principal's *optimal* choice of β when β now trades off (i) statistical value across all coordinates and (ii) strategic exposure through $\|\beta_M\|_*$. Even in the Gaussian benchmark, optimizing $\beta \mapsto \mathcal{L}(\beta, \sigma)$ is typically a nonconvex problem because the equilibrium manipulation region depends on β through both the score distribution and the manipulable thickness. For this reason, multi-dimensional calibration is a natural point where numerical methods (grid search over σ plus gradient-based methods over β, or bilevel optimization with simulated equilibrium responses) become practically necessary.

7.2 (b) Alternative disclosure channels beyond additive noise

Our graded-disclosure instrument $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a stylized stand-in for "auditable randomness." Institutions often implement similar incentive effects using other disclosure channels that may be more legally or operationally acceptable than "injecting noise." Three examples are particularly close in economic content.

Coarse thresholds and rounding. Suppose the principal computes a real-valued score s but commits to accept based on a coarsened version $\tilde{s} = \text{round}_h(s)$ (round to a grid of step $h > 0$), accepting iff $\tilde{s} \geq 0$. From the agent's perspective, this behaves like a threshold with an implicit tie region of width h: if $s \in [-h/2, h/2)$, small manipulations can flip acceptance only by jumping a discretization bin. This sharply reduces the marginal value of infinitesimal manipulation (indeed, it makes the acceptance probability locally flat away from bin edges), at the cost of introducing deterministic "quantization error" near the boundary. Unlike Gaussian noise, rounding

generates non-smooth best responses and can create bunching at bin edges; characterizing equilibrium manipulation in closed form is generally difficult even with one-dimensional p, and simulations are typically needed once we allow heterogeneous κ and correlated features.

Partial disclosure of weights (or feature subsets). Another common channel is to disclose only a subset of coefficients or only a subset of features. Formally, let the principal's true score be $s = z + \beta^\top p$, but the publicly disclosed object is $\tilde{s} = z + \tilde{\beta}^\top p$ where $\tilde{\beta}$ is a coarse summary (e.g., some coordinates omitted, or coefficients binned). Agents then best respond to their perceived mapping from Δ to acceptance probability, which is a function of the disclosure policy. This can be modeled as a commitment to a *set* of feasible weights $\mathcal{B}$ (disclose $\mathcal{B}$ but not $\beta \in \mathcal{B}$), with the actual β drawn from a distribution supported on $\mathcal{B}$ after the manipulation stage. The acceptance probability becomes a mixture $\mathbb{E}_{\beta \sim D}[\mathbf{1}\{z + \beta^\top (p + \Delta) \geq 0\}]$ (or the probit analogue), which again converts "buying acceptance" into "buying probability" with bounded marginal returns. The mixture instrument can be fully auditable (the distribution D is public; the draw can be verifiable ex post), but closed-form equilibrium characterizations are rare because the mixture generally destroys the single-index structure that makes the probit case tractable.

Randomized cutoffs (threshold lotteries). Instead of adding noise to scores, the principal can randomize the cutoff: accept iff $s \geq \tau$ where τ is drawn from a known distribution after the agent acts. This is equivalent to additive noise with $\varepsilon = -\tau$, but it can be easier to communicate institutionally ("the acceptance bar is drawn from a narrow band"). The economic object remains the slope of the acceptance probability as a function of Δ. In particular, for any absolutely continuous cutoff distribution with density bounded by M, the marginal benefit of manipulation is uniformly bounded by $M \cdot \|\beta_M\|_*$, yielding deterrence conditions analogous to $\kappa > M\|\beta_M\|_*$. This observation is useful because it decouples the incentive effect from Gaussianity: what matters is not normality per se, but the existence of a bounded density (equivalently, bounded "responsiveness") around the decision boundary.

Across these channels, our main qualitative claim is robust: instruments that bound the marginal responsiveness of acceptance to manipulable features allow the principal to rely more heavily on those features without inducing as much gaming. What changes is the analytic tractability of equilibrium; outside the probit-with-linear-cost environment, numerical characterization of best responses is often required.

7.3 (c) Manipulation graphs and expansion-style opportunity indices

Finally, many real manipulations are not well captured by additive shifts. Consider a discrete (or discretized) proxy state space V, where each node $v \in V$ represents an observable "profile" (a credit file, a transcript, a transaction history), and where feasible manipulations correspond to edges in a directed graph $E \subseteq V \times V$. An action is then a path π from the initial state v to some v', with total cost equal to the sum of edge costs; acceptance depends on the terminal node via a rule $a = \mathbf{1}\{v' \in A\}$ for an acceptance set $A \subseteq V$ induced by the principal's scoring policy.

Under full transparency (a deterministic A), a low-cost agent chooses the cheapest path into A, so the set of manipulators is characterized by a reachability neighborhood:

$$\mathcal{N}_r(A) \;:=\; \{v \in V : \text{ there exists a directed path from } v \text{ to some } a \in A \text{ of cost } \le r\}.$$

With value 1 for acceptance and linear marginal cost κ, the relevant radius is $r = 1/\kappa$. The transparent analogue of the purchasable strip is therefore $\mathcal{N}_{1/\kappa_L}(A) \setminus A$: agents initially outside A but within cost-distance $1/\kappa_L$ can "buy" acceptance. This suggests an opportunity index of the form

$$\Delta_{\text{graph}}(A) \;:=\; \mathbb{P}\big(v \in \mathcal{N}_{1/\kappa_L}(A) \setminus A\big),$$

with the probability taken under the no-manipulation distribution over initial nodes. Lower bounds on the transparency tax can then be expressed in terms of $\alpha\,\Delta_{\text{graph}}(A_T)$, where A_T is the best transparent acceptance set induced by the principal's optimal transparent rule. The dependence on graph structure enters through bounds on $\Delta_{\text{graph}}(A)$ in terms of combinatorial quantities—maximum out-degree Δ, growth rates of neighborhoods, or expansion/conductance-like measures that control how quickly $\mathcal{N}_r(A)$ enlarges as r increases.

Graded disclosure in graphs can be implemented by randomizing over acceptance sets $\{A_1, \dots, A_m\}$ after the agent acts (analogous to our post-manipulation noise). The agent then chooses a path to maximize $\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{v' \in A_j\} - \kappa \cdot \text{cost}(\pi)$, so a move must be "robustly" good across many possible A_j's to be worthwhile. This converts reachability of a *single* set into something closer to reachability of an *average* or *intersection* structure, thereby shrinking the effective neighborhood that is profitable to enter. In bounded-degree graphs, one can often upper bound the agent's best achievable acceptance probability gain by a function scaling like $\log \Delta$ or Δ depending on whether the randomization family has a "hashing" property (many boundaries with limited overlap) or is more redundant.

Analytically, however, graphs are the point where closed forms largely disappear. Even computing $\Delta_{\text{graph}}(A)$ for a given A requires shortest-path

computations and integration against the node distribution; optimizing over policies is a bilevel problem over a combinatorial action space. For realistic graphs (large V, heterogeneous edge costs, and acceptance sets induced by a learned score), numerical methods—shortest-path or min-cost reachability for agent responses, plus heuristic or approximate optimization for the principal—are typically required. We view this not as a weakness of the economic mechanism, but as an honest reflection of the complexity of manipulation technologies in operational settings.

8 Policy and implementation: auditability, disclosure, and the case for *bounded responsiveness*

Our model draws a sharp distinction that often gets blurred in policy debates: *auditability* is not the same as *full public targetability*. In the formalism, the principal's policy is auditable because it is a committed mapping from observed inputs to acceptance, parameterized by (β, σ) and a publicly specified distribution for the noise ε. Mandated transparency corresponds to setting $\sigma = 0$, which makes the acceptance boundary deterministic and therefore perfectly targetable by agents who can manipulate p. The key implementation lesson is that regulators can insist on strong ex post verifiability (auditing) while still permitting institutions to avoid fully targetable decision rules by allowing $\sigma > 0$ (graded disclosure), rounding, or randomized cutoffs. Put differently: what generates the transparency tax is not oversight, but *determinism at the boundary* when manipulable channels exist.

Auditability versus public disclosure. A common worry is that any nondeterminism is "unaccountable" or "nontransparent." Our framework suggests a more precise compliance target: require that (i) the distribution of randomness be disclosed and (ii) the realized randomness be *verifiable ex post* to authorized auditors, even if it is not predictable ex ante to applicants. This is exactly the informational structure in which graded disclosure reduces the marginal value of manipulation without allowing the principal to hide arbitrary discrimination behind noise. Concretely, an auditable graded-disclosure policy can be implemented by logging, for each decision, the realized draw ε (or equivalently the randomized cutoff), together with an integrity proof that the draw came from the stated distribution and was generated after the agent's action.

This distinction matters for legal and governance purposes. Many legal regimes are primarily concerned with whether similarly situated individuals are treated differently *for forbidden reasons*, and whether the decision process is reviewable. A properly designed randomized procedure can satisfy those goals if it (a) is independent of protected attributes conditional on the inputs, (b) is drawn from a pre-committed mechanism, and (c) can be

reconstructed in an audit. In our notation, the compliance object is not the realization of ε, but the public commitment to $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ (or another specified distribution) and the guarantee that it is independent of (t, z, p, κ).

Legally acceptable randomness: tie-breaking, threshold lotteries, and rounding. Institutions already use randomness in settings where deterministic tie-breaking is viewed as unfair or gameable: lotteries for oversubscribed schools, random audits in tax enforcement, and randomized inspection in safety regulation. The mechanism we analyze is closest to two operational tools.

First, *randomized cutoffs* (threshold lotteries): accept iff $s \geq \tau$ where τ is drawn from a known distribution after the applicant's submission. This is often easier to explain than "we add noise," and it is behaviorally equivalent to additive noise with $\varepsilon = -\tau$. Second, *rounding/banding* of scores: compute a continuous score but map it into bins (e.g., deciles, letter grades, coarse risk categories), then apply deterministic rules to bins. Both tools share the same incentive property: they bound the *responsiveness* of acceptance to small changes in manipulable proxies. In the Gaussian benchmark, the bound appears transparently as the maximum slope of $\Phi(\cdot)$, yielding the deterrence condition

$$\kappa \; > \; \frac{\beta}{\sigma\sqrt{2\pi}} \quad \Rightarrow \quad \Delta^* = 0.$$

More generally, what matters is not normality, but that the acceptance probability as a function of the score has a uniformly bounded derivative around the boundary. Rounding achieves this by creating flat regions; threshold lotteries achieve it by smoothing the step function.

From a legal-design perspective, these alternatives are valuable because they can be characterized as (i) procedural tie-breaking rules applied near indifference, or (ii) uncertainty bands that reflect measurement error, rather than as "arbitrary randomness." Importantly, our mechanism does not require large σ. The policy case is precisely that small, auditable uncertainty can reduce strategic distortions enough to justify itself, even when it slightly increases baseline classification noise.

Guidance for regulators: what to mandate (and what *not* to mandate). If regulators mandate $\sigma = 0$ in the name of transparency, the model predicts they can inadvertently force principals to *downweight manipulable but informative proxies*, i.e.,

$$\beta_T \; < \; \beta_{\text{Bayes}} \quad \text{when manipulation is present,}$$

creating a predictable loss in accuracy and potentially shifting selection toward less informative (or more historically biased) signals. A more robust

regulatory stance is to mandate auditability and measurement discipline, while allowing limited graded disclosure up to a cap $\bar{\sigma}$. In practice, we can translate this into implementable requirements:

- **Mechanism commitment.** The decision rule (including any randomization) must be fixed ex ante and documented: functional form, parameter values, and the randomness distribution.
- **Verifiable randomness.** The institution must generate randomness using an auditable procedure (e.g., a publicly committed seed, a verifiable random function, or a trusted randomness beacon), with per-decision logs enabling reconstruction.
- **Independence and timing.** The random draw must be generated *after* applicant information is locked, and must be statistically independent of protected attributes and operator discretion.
- **Bounded randomization.** Permit randomness only within a prescribed band (a cap $\bar{\sigma}$ or an equivalent bound on the density of the cutoff distribution), ensuring predictability and limiting welfare loss from excessive noise.
- **Monitoring for gaming.** Require periodic tests for strategic response (e.g., excess mass of applicants just above decision thresholds, sudden changes in proxy distributions, or evidence of coaching markets), with the ability to adjust (β, σ) within the pre-approved policy space.

The model's comparative statics give an interpretable justification for such rules: the value of allowing $\sigma > 0$ is increasing in the mass of low-cost manipulators α and in proxy informativeness. Hence, rigid "no randomness" mandates are most costly precisely in domains where manipulation markets are active and proxies are predictive (credit, hiring assessments, admissions test prep).

Guidance for model builders: designing for strategic robustness. For practitioners, the message is not simply "add noise." It is "control targetability while preserving auditability." Three concrete design steps follow naturally.

(i) *Estimate strategic exposure.* Before deployment, quantify how much of the applicant mass lies within a manipulable distance of the boundary under the candidate model—an empirical analogue of Δ_{proxy}. This can be approximated by measuring how acceptance changes under plausible perturbations to manipulable features and by auditing the availability/cost of manipulation services.

(ii) *Separate measurement from incentives.* If possible, invest in features that are genuinely hard to manipulate (our z-like channels) and treat highly

manipulable features (our p-like channels) as useful but incentive-sensitive. In terms of our score, this often means resisting the temptation to "explain" the model by publishing a fully precise linear threshold that can be reverse-engineered into a step-by-step manipulation guide.

(iii) *Operationalize graded disclosure carefully.* If adopting a $\sigma > 0$ policy (or rounding), ensure the randomness is centralized, logged, and insulated from human discretion. Discretionary noise—where operators can selectively "wiggle" decisions—is precisely what creates accountability problems. Mechanistic noise with audit trails is different: it is a policy instrument akin to randomized audits.

Fairness implications (qualitative): who bears the cost of manipulability and of noise? Fairness interacts with graded disclosure through two opposing channels. On one hand, manipulation opportunities are rarely equally distributed. If low-cost manipulation κ is systematically lower for advantaged groups (because they have better coaching, more time, better documentation, or easier access to credentialing), then deterministic transparency can amplify inequity: advantaged agents are more able to "buy" acceptance by moving p. In our language, group differences in the distribution $G(\kappa)$ imply different manipulation rates, and therefore different false-positive/false-negative distortions across groups under $\sigma = 0$. Introducing graded disclosure that deters manipulation can therefore *improve* fairness by reducing the returns to unequal access to gaming.

On the other hand, adding noise (or coarse banding) can create differential harm if groups are differently concentrated near the boundary. Even if the mechanism is group-blind, randomness is most consequential for those near the cutoff; if one group is disproportionately represented in the marginal region (for structural reasons), they will experience more stochastic outcomes. This is not a decisive argument against graded disclosure, but it does imply a governance obligation: when adopting $\sigma > 0$, institutions should report how outcome variance and marginal acceptance probabilities vary by group, and should consider complementary policies (appeals, second-stage reviews, or alternative pathways) that reduce the burden of randomness on those persistently near the threshold.

A practical synthesis is to view graded disclosure as a *first-stage anti-gaming device*, not as the final word on fairness. If randomness is used, it should be paired with (i) clear, non-gameable guidance on how to become genuinely qualified (raising t, not merely p), and (ii) procedures that preserve substantive due process (e.g., an appeal that verifies hard-to-manipulate evidence). In this way, graded disclosure reduces Goodhart pressure while keeping the system contestable and accountable.

Limitations and compliance risks. Two caveats are worth stating explicitly. First, our mechanism assumes the principal can commit and that the randomness is trusted. If stakeholders suspect manipulation of the randomization (selective reruns, biased seeds), the legitimacy benefits of auditability are lost. This pushes strongly toward verifiable randomness and third-party audits. Second, graded disclosure can be misunderstood as a license to be opaque about *inputs* or *objectives*. Our analysis does not justify hiding which features are used, nor does it justify withholding protected-class audits; it only cautions against making the exact decision boundary perfectly targetable when some inputs are manipulable.

Taken together, the policy implication is narrow but consequential: regulators should aim to cap and audit *responsiveness* to manipulable proxies, rather than mandating deterministic transparency. Doing so preserves the accountability benefits of oversight while mitigating the strategic distortions that otherwise force principals into inferior, less informative models—precisely the transparency tax our framework highlights.

9 Simulations (optional): scaling laws, heterogeneity, and a benchmark illustration

While our main results are analytic, it is useful to verify that the comparative statics and "scaling" implications are visible in finite samples and under modest deviations from the benchmark assumptions. In this section we outline a simulation protocol that (i) recovers the qualitative shape of the transparency tax, (ii) illustrates how heterogeneity in manipulation costs shapes equilibrium distortions, and (iii) (optionally) demonstrates the mechanism on a standard classification dataset by designating a subset of features as manipulable. The goal is not to "estimate" the model, but to sanity-check the economic logic in a controlled environment.

Synthetic data-generating process. We begin from the Gaussian benchmark used in the propositions: draw latent qualification $t \sim \mathcal{N}(0,1)$ and noises (η_z, η_p) jointly Gaussian with mean zero and covariance

$$\mathrm{Var}(\eta_z) = \sigma_z^2, \qquad \mathrm{Var}(\eta_p) = \sigma_p^2, \qquad \mathrm{Corr}(\eta_z, \eta_p) = \rho.$$

Set $z = t + \eta_z$, $p = t + \eta_p$, and $y = \mathbf{1}\{t \geq 0\}$. We treat ρ as a convenient knob for "proxy redundancy": when ρ is high and σ_z^2 is small, z already captures most information in p, so the opportunity cost of downweighting p under transparency is smaller.

For manipulation costs, we consider three families:

1. *Two-point costs* (the benchmark): $\kappa \in \{\kappa_L, \infty\}$ with $\mathbb{P}(\kappa = \kappa_L) = \alpha$.

2. *Mixtures*: κ equals κ_L with probability α and otherwise is drawn from a continuous distribution (e.g., lognormal), capturing a thin "professional coaching" market plus a broader population with heterogeneous frictions.

3. *Continuous heterogeneity*: $\kappa \sim \text{Lognormal}(\mu_\kappa, \sigma_\kappa^2)$ or $\kappa \sim \text{Gamma}(k, \theta)$, which allows us to vary the thickness of the lower tail (the mass of low-cost manipulators) independently from the mean.

Computing best responses and equilibrium outcomes. Fix a policy (β, σ). Each agent observes (z, p, κ) and chooses Δ to maximize

$$U(\Delta) = \Phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) - \kappa|\Delta|, \qquad (\sigma > 0),$$

with the deterministic specialization when $\sigma = 0$. In the transparent case $\sigma = 0$ and $\beta > 0$, the best response is explicit and fast: if $s = z + \beta p \geq 0$, then $\Delta^* = 0$; if $s < 0$, then the minimal crossing action is $\Delta_{\min} = -s/\beta$, and the agent manipulates iff $\kappa\Delta_{\min} < 1$ (using the convention that ties are broken toward minimal movement).

For $\sigma > 0$, we compute Δ^* numerically using the fact that, for $\Delta > 0$, the first-order condition is

$$\frac{\beta}{\sigma}\,\phi\left(\frac{z + \beta(p + \Delta)}{\sigma}\right) = \kappa,$$

and similarly on $\Delta < 0$ with the appropriate sign. Practically, we solve for Δ by one-dimensional root finding on the active side and compare the resulting candidate utility to $U(0)$; this is robust because $U(\Delta)$ is smooth and the marginal gain is bounded above by $\beta/(\sigma\sqrt{2\pi})$. In Monte Carlo, this procedure yields an estimated manipulation rate $\widehat{\mathbb{P}}(\Delta^* \neq 0)$, an estimated misclassification loss $\widehat{\mathcal{L}}(\beta, \sigma)$, and decomposition into false positives/false negatives. We then (approximately) solve the principal's problem by grid search over $(\beta, \sigma) \in [\beta_{\min}, \beta_{\max}] \times [0, \bar{\sigma}]$ or by a coarse-to-fine pattern search; the objective is nonconvex in general because equilibrium behavior changes discretely with (β, σ).

Scaling patterns: where the transparency tax comes from in finite samples. Across a wide range of parameterizations, three patterns are stable.

(1) Tax increases with low-cost mass. Holding $(\sigma_z^2, \sigma_p^2, \rho)$ fixed, the difference

$$\widehat{\text{Tax}} \;=\; \min_{\beta} \widehat{\mathcal{L}}(\beta, 0) \;-\; \min_{\beta, \sigma \in [0, \bar{\sigma}]} \widehat{\mathcal{L}}(\beta, \sigma)$$

is increasing in α in the two-point benchmark and increasing in the lower-tail thickness of G under continuous heterogeneity. Mechanically, under $\sigma = 0$

the set of agents with $s \in [-\beta/\kappa, 0)$ finds it profitable to "just cross," and this region expands as more mass shifts to low κ.

(2) Tax increases with proxy informativeness. Decreasing σ_p^2 (keeping σ_z^2 fixed) increases the benefit of using p absent manipulation. Under transparency, the principal responds by shrinking β to limit gaming, which shows up directly in simulations as $\widehat{\beta}_T < \widehat{\beta}^*$ whenever manipulation is present. The corresponding tax is largest in the regime where p is informative *and* manipulable at low cost.

(3) Optimal noise is "just enough" to flatten the marginal incentive. When we plot $\widehat{\mathcal{L}}(\beta, \sigma)$ along the locus of β values that perform well absent gaming, the loss as a function of σ is typically U-shaped: small σ sharply reduces manipulation (large gain), but large σ eventually adds too much intrinsic randomness (large cost). The σ that minimizes loss often aligns with the deterrence heuristic from the smooth-regime cutoff,

$$\sigma \gtrsim \frac{\beta}{\kappa_L \sqrt{2\pi}},$$

in the sense that the chosen σ is near the smallest value that makes manipulation unattractive for the low-cost tail.

Role of proxy correlation and redundancy. Allowing $\rho \neq 0$ is a useful stress test because it changes the marginal value of p relative to z without changing the manipulability technology. Two qualitative findings are consistent with the model's logic.

First, as ρ increases (proxies become more redundant), the principal relies less on p even absent manipulation; accordingly, the incremental value of graded disclosure tends to fall. Second, when ρ is near zero (signals are complementary) and σ_p^2 is small, p is precisely the kind of feature that a transparent rule struggles to use: it is predictive *and* it creates a wide manipulability region. In this regime the simulations show the sharpest divergence between the transparent optimum $(\widehat{\beta}_T, 0)$ and the graded-disclosure optimum $(\widehat{\beta}^*, \widehat{\sigma}^*)$.

Cost heterogeneity: manipulation is driven by the lower tail, not the mean. Moving beyond two-point costs clarifies a practical point: what matters for strategic distortion is not average cost, but the amount of probability mass with κ below the marginal-incentive bound. For $\sigma > 0$, Proposition 1 implies manipulation is essentially confined to $\{\kappa \leq \beta/(\sigma\sqrt{2\pi})\}$. In simulations with lognormal κ, changing $(\mu_\kappa, \sigma_\kappa)$ to keep $\mathbb{E}[\kappa]$ fixed while thickening the lower tail can increase manipulation rates substantially and raises the value of graded disclosure. This is the computational analogue of the analytic comparative static in α for the two-point model: the "market for gaming" is carried by a relatively small set of low-cost actors.

What to plot (recommended diagnostics). To make the strategic mechanism transparent in a figure rather than in equations, we find three diagnostics particularly informative:

- **Manipulation mass near the boundary.** Plot a histogram of the pre-manipulation score $s = z + \beta p$ for those who manipulate under $\sigma = 0$; the mass concentrates in a band just below 0, consistent with $\Delta_{\min} = -s/\beta$.
- **Equilibrium policy comparison.** Plot $(\widehat{\beta}_T, \widehat{\mathcal{L}}(\widehat{\beta}_T, 0))$ versus $(\widehat{\beta}^*, \widehat{\sigma}^*, \widehat{\mathcal{L}}(\widehat{\beta}^*, \widehat{\sigma}^*))$ as α or σ_p^2 varies; the gap visualizes the transparency tax.
- **U-shape in σ.** Fix a proxy weight β (e.g., the non-strategic best) and plot $\widehat{\mathcal{L}}(\beta, \sigma)$ and $\widehat{\mathbb{P}}(\Delta^* \neq 0)$ versus σ to show the deterrence-versus-randomness tradeoff directly.

Optional benchmark dataset illustration (proof of concept). If we want an empirical demonstration, we can implement the equilibrium response on a standard binary classification dataset (e.g., UCI Adult income, COMPAS-like recidivism proxies, or a credit default dataset), with two explicit caveats: (i) the dataset's label is not literally $y = \mathbf{1}\{t \geq 0\}$, and (ii) we must choose (by assumption) which features are manipulable and at what cost.

A simple procedure is:

1. Fit a baseline linear score $s(x) = w^\top x$ (logistic regression or linear SVM) on the training data.
2. Designate a single component x_j as manipulable (our p) and one as hard-to-manipulate (our z), or construct p as a linear combination of "documentable" variables (e.g., reported income, number of tradelines) while treating z as a less manipulable signal.
3. Postulate a cost model $\kappa|\Delta|$ and a cost distribution G across applicants (either calibrated from plausible perturbation magnitudes or explored as sensitivity analysis).
4. Evaluate, on a test set, the strategic equilibrium outcomes under (a) transparency $\sigma = 0$ and (b) graded disclosure $\sigma > 0$ by applying the best-response manipulation to p and then applying the acceptance rule with the corresponding noise.

Even with this stylized mapping, the qualitative patterns typically mirror the synthetic case: under $\sigma = 0$ we see a shift in the manipulable feature's distribution among accepted individuals and a drop in out-of-sample label accuracy relative to the non-strategic classifier; introducing a small σ dampens the return to marginal feature shifts and can recover accuracy by allowing

a higher effective weight on the predictive proxy without inducing as much gaming.

Limitations of the simulation exercise. Simulations do not resolve the normative questions around randomness, nor do they pin down domain-specific costs of manipulation. Their role is narrower: to confirm that the model's equilibrium logic is not an artifact of knife-edge assumptions, and to illustrate how the key objects (manipulation mass near the boundary, optimal β shrinkage under transparency, and the U-shaped role of σ) behave in finite samples. In particular, when we depart from Gaussian noise or introduce mild nonlinearities in $t \mapsto (z, p)$, the same basic force persists as long as (i) acceptance is highly responsive to small changes in a manipulable proxy under determinism, and (ii) graded disclosure bounds that responsiveness in an auditable way.

These computational checks set the stage for the final section, where we discuss what changes—and what becomes harder—once the interaction is repeated, the principal learns over time, and agents can experiment strategically.

10 Conclusion and open problems: dynamics, selection, and design

Our central message is that "transparency" is not a free good in strategic environments. When a decision rule is fully targetable, a manipulable proxy becomes an object of investment rather than measurement, and the principal may rationally retreat from using it—even when it is genuinely informative about latent qualification. Graded disclosure, modeled here as an auditable randomization parameter σ in the acceptance rule, provides a simple countervailing instrument: it bounds the marginal return to micro-manipulations while preserving the option to lean on informative proxies. This produces a wedge—the transparency tax—between the best achievable performance under mandated determinism and the best achievable performance when a limited degree of randomized discretion is permitted.

The natural next step is to ask what happens when the interaction is repeated, the principal learns, and agents can experiment. In practice, few institutions set (β, σ) once and for all. Hiring screens are updated, fraud models are retrained, admissions rubrics drift, and applicants respond strategically to what appears to work. The static equilibrium logic survives in such settings, but the long-run outcomes depend on an additional set of forces: data feedback, commitment problems, and equilibrium selection via experimentation.

Dynamic learning with repeated interactions. Consider a repeated version of the game indexed by periods $\tau = 1, 2, \ldots$, where each period draws new agents with types $(t_\tau, \eta_{z,\tau}, \eta_{p,\tau}, \kappa_\tau)$ and the principal updates a policy $(\beta_\tau, \sigma_\tau)$ based on past observations. The key complication is that the principal rarely observes t directly; instead she observes outcomes that are themselves affected by acceptance decisions and by manipulation. This creates a selective-labels problem layered on top of strategic response: the distribution of observed $(z, \hat{p})$ among accepted agents is endogenously distorted, and any supervised update that treats $\hat{p}$ as "ground truth features" may gradually encode gaming into the model.

Even if the principal does observe some ex post signal of performance (e.g., job performance, loan repayment), it is typically delayed and itself a function of the decision. This leads to a dynamic analog of Goodhart's law: the principal's predictor improves on the manipulated data distribution, which can raise apparent in-sample performance while degrading true out-of-sample classification with respect to $y = \mathbf{1}\{t \geq 0\}$. From the perspective of our framework, the parameter σ plays a dual role in such dynamics. It deters manipulation contemporaneously, but it also stabilizes the learning environment by reducing the incentive for agents to chase small decision-boundary movements. A concrete open question is whether there exists a "stability region" in (β, σ) such that best-response manipulation is not only small (low $\mathbb{P}(\Delta^* \neq 0)$) but also insensitive to small changes in the learned policy, thereby preventing oscillatory arms races between retraining and gaming.

A second dynamic issue is commitment. In our one-shot model the principal commits to an auditable (β, σ). In repeated settings, agents may doubt that the principal will maintain a noisy policy once public scrutiny intensifies or once short-run error is observed. If agents anticipate that the principal will revert to $\sigma = 0$ after a few periods, the deterrence effect of graded disclosure can unravel. Formally, one can model the principal as choosing a policy each period without commitment, and study Markov perfect equilibria where current σ_τ trades off current accuracy against future gaming. Whether the equilibrium features persistent randomization depends on how manipulation today affects the state variable (e.g., the principal's dataset, agents' beliefs about rules, or the distribution of "coaching" services). Understanding when graded disclosure is dynamically time-consistent is an open problem with direct policy relevance: it speaks to whether regulators should permit (and perhaps require) explicit commitment devices for randomization, rather than treating randomness as ad hoc "discretion."

Equilibrium selection and strategic experimentation. Even in the static model, best responses can exhibit corner behavior: with $\sigma = 0$ agents "just cross," while with $\sigma > 0$ manipulation may be interior and governed by

a first-order condition when it occurs. In richer environments—multiple manipulable proxies, non-linear scores, or non-convex costs—agent optimization can admit multiple local optima. Our analysis implicitly selects equilibria by tie-breaking toward minimal movement and by focusing on the economically relevant branch where manipulation is used to increase acceptance. In a repeated environment, such selection cannot be taken for granted: agents can experiment with Δ, observe acceptance outcomes, and update beliefs about the policy and about the efficacy of manipulation.

This raises a distinct equilibrium-selection question: can "gaming cultures" emerge as self-confirming outcomes? Suppose a small measure of agents experiments with manipulation; if this shifts observed accepted populations and prompts the principal to adjust β downward (to counteract gaming), the return to manipulation may increase for those near the boundary, pulling more agents into manipulation. Conversely, if graded disclosure is used to flatten marginal incentives, experimentation may be unprofitable and die out. Modeling this requires combining (i) learning-by-agents about the acceptance technology (including the distribution of ε), with (ii) learning-by-principal about predictive relationships, in a feedback loop.

A promising approach is to study stochastic-approximation dynamics for agent behavior and for the principal's training rule. For example, if agents follow a perturbed best response

$$\Delta_{\tau+1} \approx \Delta_\tau + \gamma_\tau \nabla_\Delta U_\tau(\Delta_\tau),$$

and the principal updates $(\beta_\tau, \sigma_\tau)$ via gradient steps on an empirical loss computed from strategically generated data, one can ask whether the coupled system converges, cycles, or diverges. The static transparency tax suggests that purely deterministic policies may induce dynamics that settle into low-β equilibria (the principal gives up on the manipulable proxy), whereas permitting $\sigma > 0$ could enlarge the basin of attraction of high-β / low-gaming outcomes. Characterizing these basins—and the conditions under which "small auditable noise" is enough to prevent undesirable equilibria—remains open.

Connections to persuasion and information design. Graded disclosure can be interpreted as an information-structure choice: the principal commits not only to a scoring rule but also to how sharply the score maps into decisions. In this sense our model is adjacent to Bayesian persuasion and information design, with a crucial twist: the receiver (the agent) can change the signal by manipulating p. The principal's randomization ε is then a way to commit to a less informative mapping from reported features to acceptance, reducing the private value of precise targeting.

This perspective suggests two generalizations. First, rather than restricting to additive Gaussian ε, one can allow the principal to choose an arbitrary (auditable) acceptance probability function $q(s) \in [0, 1]$ with $a \sim$

Bernoulli($q(s)$). Our smooth-regime cutoff is a special case where $q(s) = \Phi(s/\sigma)$. The optimal $q(\cdot)$ under manipulation costs resembles a "bounded slope" design: since agents respond to $q'(s)\beta$ at the margin, the principal would like to cap q' near the decision boundary while keeping q responsive where strategic pressure is low. Designing the welfare- and accuracy-optimal $q(\cdot)$ under auditability constraints (e.g., monotonicity, Lipschitz bounds, or implementability with simple randomization devices) is an open information-design problem with strategic signal control.

Second, persuasion models emphasize commitment to a signal for the purpose of influencing behavior. Here, influencing behavior (deterring manipulation) is not incidental but central. This highlights a policy tension: some transparency mandates implicitly push institutions toward fully revealing s and using deterministic thresholds, which is equivalent to choosing an extremely steep $q(\cdot)$. Our results suggest that, in strategic settings, such mandates may be counterproductive even when they improve "explainability" in a narrow sense.

Connections to mechanism design (beyond linear scores). From a mechanism-design viewpoint, our environment is a screening problem with costly misreporting: agents can shift a report $\hat{p}$ at linear cost, and the principal chooses an allocation rule (accept/reject) based on observable signals. The linear score restriction is deliberately stark; relaxing it raises both opportunity and complexity. For example, if the principal could condition on $(z, \hat{p})$ nonlinearly, she might carve out regions where p is trusted only when consistent with z, or implement "audit-triggering" rules that make manipulation risky. Introducing audits explicitly (probabilistic verification of p at a penalty) would change the incentive constraint from a pure marginal tradeoff to a jump risk, potentially complementing or substituting for graded disclosure.

A core open problem is to characterize the optimal mechanism under realistic constraints: no transfers (or limited transfers), limited auditing capacity, and legal restrictions on randomization or disparate treatment. In many regulated domains the principal cannot freely choose σ or cannot randomize by individual; instead she may be allowed only coarse randomization (e.g., random audits, randomized additional review). Understanding when such constrained mechanisms can replicate the deterrence effect we study—perhaps via randomized audits rather than randomized acceptance—is both technically and practically important.

Normative and policy questions. Randomization in high-stakes decisions raises legitimate concerns: perceived arbitrariness, procedural fairness, and accountability. Our model does not claim that noise is always desirable; rather, it clarifies when some controlled, auditable non-determinism

can reduce strategic distortion enough to improve accuracy with respect to a socially relevant label. This invites normative extensions: if the social objective weights false positives and false negatives differently across groups, how should σ be chosen? Does graded disclosure amplify or reduce inequities when manipulation costs κ differ systematically across populations? A particularly important open question is whether a uniform σ is optimal under fairness constraints, or whether the constrained optimum involves targeted robustness—effectively choosing policies that equalize marginal manipulability across groups rather than equalizing raw acceptance rates.

Finally, there is a measurement question embedded in policy: regulators and auditors often observe only the implemented rule, not the strategic effort it induces. Developing empirical diagnostics for "proxy gaming" that do not rely on observing Δ directly—e.g., distributional shifts in p conditional on z, bunching near implied decision boundaries, or changes in the correlation between p and downstream outcomes—would make the transparency tax concept operational.

Summary of open problems. To sharpen the agenda, we see five concrete directions: (i) dynamic equilibrium with learning and selective labels; (ii) equilibrium selection under agent experimentation and principal retraining; (iii) optimal information design over acceptance probabilities $q(s)$ under auditability; (iv) mechanism design with audits, limited commitment, and legal constraints; and (v) welfare and fairness analysis when manipulation costs and proxy access are heterogeneous. Each direction preserves the core economic logic illuminated here—targetability creates incentives, and policy can shape those incentives—while moving closer to the institutional realities that motivate transparency debates.