Causal Recourse Contracts: Implementing Genuine Improvement and Blocking Proxy Gaming

Liz Lemma Future Detective

January 14, 2026

Abstract

Strategic adaptation to algorithmic decisions can be either harmful gaming (changing non-causal proxies) or beneficial investment (changing causal features). Building on the strategic classification framework (Stackelberg principal–agent), the causal perspective that 'strategic classification is causal modeling in disguise,' and recent work distinguishing gaming from improvement, we propose a clean mechanism-design formulation: the principal chooses both a decision rule and a recourse policy (a verifiable counterfactual "contract") that guides agents toward actions that genuinely change the ground-truth outcome. In a tractable causal model where qualification depends only on causal features and agents can invest in those features at separable single-crossing costs, we characterize an implementable mechanism that (i) induces all improvable agents to become qualified, (ii) guarantees zero false positives from proxy manipulation, and (iii) is optimal among all safe mechanisms. We also outline extensions to limited recourse menus (administrative simplicity) and to learning causal thresholds from strategic responses, clarifying when closed-form solutions suffice and when approximation or numerical methods are needed.

Table of Contents

1. 1. Introduction: why 'recourse as contract' is the 2026 mechanism-design lens; link to Goodhart's law, causal vs proxy features, and policy demands for actionable explanations.

2. 2. Related work and positioning: strategic classification (Hardt et al.), online/partial information, fairness externalities, causal strategic learning (Miller et al., Shavit et al., Bechavod et al., Ahmadi et al.), and algorithmic recourse; what is not yet unified.

3. 3. Model: causal/proxy split, verifiable causal features, investment technology, proxy manipulation, decision rule + recourse menu, and equilibrium concept.

4. 4. Baseline impossibility/limits: why standard score-based rules using proxies invite proxy gaming; why 'no false positives' is restrictive but policy-relevant; define the class of safe mechanisms.

5. 5. Structure theorem: reduction to a causal sufficient statistic (threshold on causal score) for safe mechanisms; closed-form characterization of minimal-cost improvement under single-crossing costs.

6. 6. Main implementability theorem: construct (f,r) and prove (i) zero proxy false positives, (ii) all improvable agents invest, (iii) ex post optimality among safe mechanisms.

7. 7. Extensions (menu complexity): if relative costs vary across agents, show when a finite menu is needed; characterize the K=1 and K=d_C cases (closed form), and flag that general K becomes a coverage/submodular optimization problem requiring approximation.

8. 8. Extensions (unknown parameters): sketch identification and learning of (β,τ) from observed strategic improvements under mild richness assumptions; relate to strategic IV/causal strategic regression.

9. 9. Welfare and policy interpretation: what 'no false positives' buys, when to relax it to ε-FP, and how to interpret recourse menus as regulated improvement pathways (credit-building, training, health compliance).

10. 10. Discussion and open problems: verifiability, measurement error in causal features, dynamic investment, group constraints, and practical deployment considerations.

1 Introduction: recourse as contract under strategic responses

Automated decisions increasingly sit at the intersection of prediction, policy, and strategy. A lender, employer, or platform rarely wants to merely *predict* an outcome; it wants to *select* agents so as to maximize some notion of benefit (e.g., true repayment, job performance, safety) while respecting constraints (e.g., legal limits on disparate impact, procedural rights, capacity, and—crucially for operational risk—avoiding "bad accepts"). Meanwhile, the people subject to these systems are not passive data points. They study the rules, adapt their behavior, and invest in the attributes that appear to matter. In 2026, this is no longer an edge case: the default environment for deployed scoring systems is one in which agents respond strategically, sometimes in ways that improve true outcomes and sometimes in ways that merely improve *appearances*.

This strategic feedback loop is a modern restatement of Goodhart's law: when a measure becomes a target, it ceases to be a good measure. In our setting the "measure" is a reported feature vector $\hat{x}$ and the "target" is acceptance. Goodhart's law is often invoked as an informal warning about over-optimization. Our starting point is that, once agents can act, Goodhart's law is not just a cautionary tale but a mechanism-design problem. The principal commits to a rule; agents best respond; the resulting distribution of features and outcomes is endogenous. The central question becomes: how should a principal design a decision rule and accompanying guidance so that the induced strategic behavior improves *true qualification* rather than merely gaming the interface?

A key distinction organizing the problem is between *causal* features and *proxy* features. Many deployed models are trained on whatever correlates with the desired outcome, but correlation does not tell us what can be safely incentivized. Some attributes are outcome-relevant and actionable: paying down revolving debt, completing a certification, adding safety equipment, adopting better security hygiene. Others are proxies: behavioral traces, platform engagement, stylistic markers, or network position—variables that may predict well in historical data yet have limited causal relationship to the outcome of interest, or whose manipulation does not improve the underlying target. The policy and governance conversation has converged on this point from multiple directions. "Actionability" is now a common demand in algorithmic accountability: an adverse action notice or an explanation is expected to tell an individual what they could do to obtain a different decision. But actionability without causal discipline can be worse than useless: it can redirect effort toward superficial signals, amplify arms races in manipulation, and erode both welfare and trust.

This paper frames *recourse*—the practice of offering individuals a path to

change a decision—as a *contractual object* in a strategic environment. The usual presentation of recourse is informational and ex post: once the model rejects an applicant, the system provides a counterfactual explanation ("if your income were higher by $X, you would be approved") or a set of feature changes. Such tools are valuable, but they often treat the decision rule as fixed and treat the recommendation as a separate, downstream communication problem. Our lens is different. We treat the principal as committing not only to an acceptance rule $f(\cdot)$ but also to a recourse policy $r(\cdot)$ that maps a rejection into a recommended action or menu of actions. In other words, recourse is not merely an explanation; it is part of the mechanism. Once we take strategic response seriously, this commitment changes behavior ex ante. The promise "if you do e, then you will be accepted" is economically closer to a contract than to advice, and it must be designed with the same attention to incentives and feasibility.

Thinking of recourse as contract clarifies two practical desiderata that are often in tension. The first is *safety*: avoiding false positives in the operational sense of accepting someone who is not truly qualified. The second is *opportunity*: ensuring that individuals who could become qualified through reasonable effort are not locked out by opaque rules, spurious correlations, or costly trial-and-error. Both desiderata are explicitly present in regulation and institutional practice. Financial regulators emphasize sound underwriting and model risk management alongside requirements for consumer explanations. Employers face pressure to justify selection criteria and to provide pathways for advancement, while also being accountable for quality and safety. Platforms want to deter fraud and low-quality participation without discouraging productive investment. In all these domains, the principal is effectively choosing which investments to reward.

The causal/proxy split provides a clean way to discipline this choice. We imagine that baseline features decompose as $x = (x^C, x^P)$, where x^C are causal features that directly determine true qualification and can be improved through real investment, while x^P are proxies that may correlate with outcomes but do not themselves affect qualification. Agents can take actions that increase causal features (an investment vector $e \geq 0$) and, separately, can manipulate proxies (an action m producing $\hat{x}^P$). This captures a stylized but ubiquitous reality: improving real skills or financial health is costly but beneficial, while "presentation" changes can be cheaper and sometimes purely cosmetic. Importantly, we assume causal changes are verifiable once made—an assumption that matches many institutional settings (degrees can be checked, debt balances can be observed, safety equipment can be inspected), and that makes recourse meaningful as a commitment device rather than a cheap-talk suggestion.

Under this view, a decision rule that depends on proxies invites exactly the kind of Goodhart behavior that practitioners fear. If acceptance increases the agent's payoff by $\gamma > 0$, then any feature that influences f becomes a

target. When f loads on a proxy, an agent can invest in manipulating $\hat{x}^P$ without improving y (true qualification). Even if such manipulation is individually costly, it can be socially wasteful and can degrade the principal's objective by producing "bad accepts." This is not a subtle failure mode; it is structural. The principal may start with a highly predictive model, but once the proxy becomes a target, the mapping from proxy to outcome changes. In equilibrium, the principal can end up selecting agents who are good at gaming the proxy rather than good at the underlying task.

Recourse-as-contract provides a constructive alternative: design the mechanism so that the only profitable route to acceptance is to become qualified. Intuitively, this means two things. First, the acceptance rule should be "safe" in the sense that it accepts only when the underlying causal criterion is met. Second, the principal should *tell rejected agents how to meet the causal criterion at minimum cost*, so that those who can profitably invest will do so, and those who cannot will rationally stop rather than chase proxy mirages. The novelty is not the idea that "you should reward true qualification," which is obvious. The novelty is that, in a strategic setting with actionable recourse, we can treat this as an implementability problem: can we write down a simple mechanism that (i) eliminates proxy-only gaming, (ii) induces all privately worthwhile causal investment, and (iii) does so with minimal informational and computational burden?

To build that mechanism-design logic, we will later impose a specific cost structure that is both analytically transparent and economically interpretable: causal investment has separable marginal costs scaled by a private type θ, so that higher-θ agents find investment uniformly more expensive. This "single-crossing" structure captures heterogeneity in time constraints, liquidity, opportunity cost, or access to resources, and it implies a natural cutoff behavior: for a given baseline x^C, some agents will invest to qualify and others will not. The principal need not observe θ to design good recourse; indeed, one motivation for recourse is to overcome precisely such private-information frictions by making the path to qualification explicit. When the cheapest way to increase true qualification is common across agents (up to scale), the principal can offer a particularly simple recourse contract: a single recommended action direction that raises the causal score at the lowest marginal cost.

This lens also reframes what it means to provide an "actionable explanation." In many policy discussions, actionability is treated as a normative add-on: after building the best predictor, we owe individuals a list of changes they can make. Our claim is more structural: actionability is a design primitive that can *improve the objective the principal cares about.* A recourse policy is a way of shaping equilibrium behavior. If the principal can commit to accept when a verifiable causal threshold is reached, and can credibly specify the minimal-cost steps to reach that threshold, then recourse transforms strategic behavior from adversarial manipulation into productive

investment. In that sense, recourse is not merely a compliance artifact; it is a tool for safely increasing true positives—accepting more truly qualified agents—without relaxing safety.

At the same time, we emphasize limitations and scope. The causal/proxy split is a modeling choice that abstracts away from ambiguous or partially causal features, from measurement error, and from settings where "verification" is imperfect. In practice, many attributes are neither purely causal nor purely proxy; they may be partially manipulable, partially informative, and entangled with fairness concerns. Our goal is not to claim that real systems can always cleanly separate x^C from x^P. Rather, the model illuminates the tradeoff that emerges once we insist on two operational requirements that organizations routinely face: do not accept unqualified agents (a robust "no false positives" constraint), and provide recourse that is genuinely actionable. Under these requirements, reliance on manipulable proxies is a recipe for either gaming or excessive conservatism. The mechanism-design perspective identifies a path between these extremes: restrict eligibility to verifiable causal features, and use recourse to guide investment toward the cheapest causal improvements.

Finally, "recourse as contract" helps connect theory to practice in a way that purely predictive framing often misses. Institutions already behave as if they are writing contracts: lenders publish underwriting criteria; employers post job requirements and training ladders; platforms specify reputation thresholds and remediation steps. What is new is the degree to which these contracts are mediated by algorithms trained on proxy-rich data. Our framework provides a disciplined way to decide which parts of an algorithmic score should be treated as eligible for contractual commitment, and which should be quarantined as potentially gameable signals. It also suggests a practical principle: if a feature cannot be defended as outcome-relevant and verifiable once changed, then building recourse around it is likely to induce wasteful behavior, and building acceptance around it is likely to violate safety once agents respond.

The remainder of the paper formalizes this logic. We define a principal-agent game in which the principal commits to an acceptance rule and a recourse policy, agents choose both causal investments and proxy manipulations, and outcomes depend only on the causal state. We then show that, under transparent causal qualification and single-crossing investment costs, there exists a simple, computationally efficient mechanism that is simultaneously safe (no proxy false positives) and maximally opportunistic among safe mechanisms (it induces all privately beneficial causal improvements). The economic message is that explanations and incentives cannot be separated: once agents act, the "right" explanation is the one that implements the desired equilibrium.

2 Related work and positioning

Our framework sits at the intersection of strategic classification, causal perspectives on strategic behavior, and the rapidly growing literature on algorithmic recourse. Each of these areas has developed powerful tools, but they are often studied in isolation: strategic classification typically takes the decision rule as the object to optimize given manipulation; recourse work typically takes the rule as fixed and asks for post hoc recommended changes; and causal strategic learning asks what can be learned or implemented when agents can change features, but often without an explicit *no-false-positives* safety constraint that is central in many operational settings. Our goal is to unify these strands around a mechanism-design question: what decision rules and recourse policies can a principal *commit* to so that strategic responses are redirected toward outcome-improving investments rather than proxy gaming, while maintaining a robust "no bad accepts" guarantee.

Strategic classification and manipulation. A foundational thread begins with the strategic classification model of Hardt et al. and related work on agents who modify observable features to obtain a favorable classification. In this line, the classifier chooses a rule, agents respond by changing their features subject to costs, and the induced distribution of observed features differs from the training or baseline distribution. Subsequent work has explored equilibrium structure, the welfare and efficiency of strategic adaptation, and the vulnerability of linear and threshold classifiers to manipulation, as well as the difficulty of learning when the data-generating process is endogenously shaped by the deployed model.

Our paper shares this basic premise—classification rules induce behavioral change—but departs from the canonical strategic classification framing in two ways that matter for policy and for mechanism design. First, we explicitly separate features into *causal* attributes that determine true qualification and *proxy* attributes that may be predictive but outcome-irrelevant. This moves the analysis from "manipulation is costly" to "some manipulations are socially and operationally meaningless," which is precisely the setting where Goodhart effects are most damaging. Second, we treat the principal's operational concern as a *safety constraint*: the principal should avoid accepting agents who are not truly qualified, not merely in expectation but in a pointwise (equilibrium) sense. Much of strategic classification optimizes an accuracy or utility objective under strategic response, sometimes allowing an equilibrium tradeoff between false positives and false negatives. In lending, hiring, safety-critical access control, and many regulated domains, the principal often faces a hard constraint against false positives ("bad accepts"), making the problem closer to robust mechanism design than to standard risk minimization.

Endogenous data, repeated interaction, and partial information. A separate but related literature studies learning and decision-making under feedback loops: the deployed rule changes the population that is observed, which in turn affects future learning and performance. This includes work on selective labels, censorship, and dynamics of decision policies, and connects to performative prediction and policy-induced distribution shift. There are also online and bandit-style models where a principal updates decisions under partial feedback and where agents may respond strategically over time.

Our contribution is not an online learning algorithm, but the "recourse as contract" perspective is naturally compatible with these settings. The key connection is that, when agents invest in verifiable causal features, the principal observes *action-induced variation* that can be informative about the causal boundary (a point we sketch in our learning extension). In contrast, when decision rules load on manipulable proxies, the induced variation is often adversarially selected and can degrade identifiability. Thus, while we analyze a Stackelberg equilibrium in a one-shot environment for clarity, the mechanism we construct is designed to be stable under repeated use: it eliminates incentives for proxy manipulation, thereby reducing one major source of endogenous nonstationarity.

Fairness, externalities, and the normative role of incentives. A growing body of work emphasizes that algorithmic decisions have equilibrium and externality effects: selection policies can change incentives to invest in education or skills, can shift resources across groups, and can generate disparate impacts even if the deployed predictor is "accurate" on historical data. Some papers formalize how fairness constraints interact with strategic response, while others study long-run dynamics under different selection rules, including the possibility of "self-fulfilling" disadvantage when groups face different opportunity costs or different access to investments.

We view our model as complementary to this agenda. By making the recourse policy an explicit part of the mechanism, we highlight a channel through which a principal can affect not only who is accepted, but also *how* agents invest. In particular, restricting eligibility to verifiable causal features can be read as an institutional commitment to reward investments that truly improve qualification, rather than investments in presentation or identity-correlated proxies. At the same time, we do not claim this resolves fairness concerns. Heterogeneous cost types θ can encode real structural constraints, and "verifiable causal features" may themselves reflect unequal access. Our results should therefore be interpreted as a positive implementability statement—what can be guaranteed under a safety constraint—rather than as a normative fairness guarantee. Indeed, one motivation for our menu-complexity extension is that heterogeneous relative costs may require richer recourse menus to avoid systematically advantaging those whose cheapest

improvement directions align with the principal's single recommended path.

Causal strategic learning and causal incentives. Recent work has begun to incorporate causal structure into strategic environments, asking which aspects of a decision rule can be safely incentivized when agents can manipulate inputs. Papers by Miller et al., Shavit et al., Bechavod et al., Ahmadi et al. (and others in this emerging area) explore settings where the principal has a causal model of how actions or interventions affect outcomes, and where strategic agents may manipulate observed variables that are not causally relevant. This literature is motivated by the same core issue we emphasize: correlation-based predictors can create incentives to manipulate variables that do not improve the outcome of interest.

Our approach is aligned with the causal strategic learning perspective but emphasizes a different design lever. Rather than optimizing over arbitrary predictors subject to strategic response, we build around a *ground-truth causal qualification rule* and ask how to implement it safely and opportunistically through a mechanism that includes recourse. The causal structure we assume is deliberately stark—true qualification is a deterministic threshold in x^C—because it lets us isolate the mechanism-design logic: if qualification is verifiable in causal features, then dependence on proxies is not merely unnecessary but potentially harmful under a no-false-positives requirement. This yields a "safe sufficiency" principle: the causal score is the minimal statistic that can be used for eligibility without inviting proxy-only gaming. We see this as a useful benchmark even when the causal model is noisy or partially observed: it clarifies what extra assumptions or auditing capacity are required to safely use richer signals.

Algorithmic recourse, counterfactual explanations, and actionability. A large literature on algorithmic recourse studies how to provide individuals with suggested changes that would flip a model's decision. This includes counterfactual explanations (e.g., Wachter et al.), actionable recourse methods that incorporate feasibility and costs (e.g., Ustun et al.), and surveys and extensions that address constraints, uncertainty, causal dependencies among features, and robustness. A recurring distinction is between *recourse as explanation* (what changes would alter the model output) and *recourse as intervention* (what changes are feasible and meaningful in the real world).

We build directly on the actionability emphasis but shift the object of analysis: we treat recourse not as a post hoc explanation layered on top of a fixed classifier, but as part of a commitment device chosen jointly with the decision rule. This difference matters in strategic environments. If the principal issues recourse recommendations that are inconsistent with the true drivers of qualification, agents may rationally invest in changes that improve

the reported score but not the outcome, inducing waste and operational risk. Conversely, if the principal can commit to accept upon verifiable causal improvement, recourse becomes a contract-like promise that coordinates agent investment on socially productive actions. In this sense, recourse is not only an individual right or transparency tool; it is an instrument for equilibrium selection.

A second distinction concerns guarantees. Much of the recourse literature evaluates the existence or cost of a path from rejection to acceptance under the deployed rule, often focusing on individual-level feasibility. In high-stakes domains, however, principals often want a system-level safety guarantee: no individual should be accepted unless truly qualified. Our mechanism is designed to satisfy such a guarantee in equilibrium, even when proxies are freely manipulable. This emphasis aligns more closely with model risk management and operational governance than with purely explanatory notions of recourse.

What is not yet unified—and our contribution. Across these literatures, three components are rarely brought together in a single, tractable model.

First, strategic classification highlights manipulation but often treats all features symmetrically as "inputs to a classifier." Recourse work highlights actionability but often abstracts from the principal's need to remain safe under strategic adaptation. Causal strategic learning highlights causal irrelevance of proxies but often focuses on learning or prediction rather than on a constructive recourse mechanism. We unify these by (i) separating causal and proxy features, (ii) imposing an explicit no-false-positives constraint as a mechanism-design requirement, and (iii) treating recourse as a contractual commitment that shapes best responses.

Second, there is a gap between *computing* a counterfactual and *implementing* it under private costs. Even if the principal knows the direction of minimal cost improvement in feature space, an agent's willingness to invest depends on private type θ and the value of acceptance γ. Our single-crossing cost structure makes this implementability problem transparent: it yields a cutoff rule for take-up and allows the principal to design a simple recourse contract that induces all privately worthwhile causal investment without needing to observe θ.

Third, the complexity of recourse menus is underexplored from a mechanism-design viewpoint. When agents differ in relative costs across causal dimensions (violating single-crossing), a one-size-fits-all recommendation can be inefficient and inequitable. We therefore flag a computational boundary: designing an optimal small menu becomes a coverage problem over types, with NP-hardness and submodular approximation structure. This connects recourse design to classical algorithmic mechanism design questions (menu

complexity and approximate optimization), and helps explain why "simple recourse" can fail precisely in heterogeneous populations.

The next section formalizes the model that supports these connections. We specify the causal/proxy split, the action technologies for causal investment and proxy manipulation, the principal's commitment to a decision rule and recourse policy, and the equilibrium concept that ties them together. This sets up our main implementability result: a mechanism that is simultaneously safe (no proxy false positives) and maximally opportunistic among safe mechanisms (it induces all privately beneficial causal improvements).

3 Model

We study a principal who must make a binary acceptance decision (e.g., approve a loan, admit an applicant, grant access to a resource) for strategic agents. The central modeling choice is to distinguish between *causal* features—attributes that genuinely determine whether an agent is qualified—and *proxy* features—attributes that may be predictive in historical data but are outcome-irrelevant and potentially manipulable. The principal commits not only to an acceptance rule but also to a *recourse policy* that specifies what improvements would guarantee acceptance. This commitment turns "recourse" from an after-the-fact explanation into a contract-like instrument that shapes incentives.

Agents, features, and true qualification. Each agent has a baseline feature vector

$$x = (x^C, x^P) \in \mathbb{R}^{d_C} \times \mathbb{R}^{d_P},$$

where x^C denotes causal (actionable and outcome-relevant) features and x^P denotes proxy (outcome-irrelevant) features. We assume a deterministic, verifiable causal notion of qualification:

$$y \;=\; h^\star(x^C) \;=\; \mathbf{1}\{\beta^\top x^C \geq \tau\}, \tag{1}$$

with known $\beta \in \mathbb{R}^{d_C}_+$ and threshold $\tau \in \mathbb{R}$. We write the causal score as $s(x^C) = \beta^\top x^C$. The maintained interpretation is that, conditional on the underlying causal attributes x^C, proxies x^P do not affect whether the agent is truly qualified. This is a stark assumption, but it is precisely the benchmark in which proxy-based decision rules are most vulnerable to Goodhart-style manipulation: any weight placed on x^P can only be justified by correlation, not by causal relevance.

We also assume that causal features are *verifiable once changed*. Concretely, if an agent invests in a credential, skill, safety training, or measurable performance, then the updated x^C can be audited or validated. This verifiability is what allows the principal to safely condition acceptance on causal

improvements; it is also what separates our mechanism-design problem from settings where all features are cheap talk.

Actions: causal investment and proxy manipulation. After observing the mechanism, the agent chooses two kinds of actions.

Causal investment. The agent selects an investment vector $e \in \mathbb{R}_+^{d_C}$, producing post-investment causal features

$$\hat{x}^C \;=\; x^C + e.$$

We treat causal investment as monotone and additive for tractability: effort can improve causal features but cannot directly decrease them.

Proxy manipulation. The agent may also take a proxy manipulation action m, which affects the reported proxy features $\hat{x}^P$. We leave the mapping from m to $\hat{x}^P$ domain-specific: m could be continuous (e.g., spending on marketing or presentation), discrete (e.g., selecting into a reporting category), or even combinatorial (e.g., editing an online profile or network). The key restriction is that proxy manipulation does *not* change true qualification y, which depends only on $\hat{x}^C$ via (1). Thus,

$$\hat{x} = (\hat{x}^C, \hat{x}^P) = (x^C + e,\ \hat{x}^P(m)).$$

This explicit separation lets us represent "gaming" as costly changes in proxies that may influence a proxy-sensitive rule but do not improve the outcome-relevant state.

Costs and private types (single-crossing). Agents differ in their cost of improving causal features. Each agent has a private type $\theta > 0$ that scales the marginal cost of investment. We assume a separable linear cost:

$$c_C(e;\theta) \;=\; \theta \sum_{j=1}^{d_C} w_j e_j, \tag{2}$$

where $w \in \mathbb{R}_+^{d_C}$ are known per-feature marginal cost weights. This "single-crossing" structure implies that agents agree on the relative attractiveness of different causal improvements: types differ only in overall difficulty, not in which direction is cheapest. The assumption is strong, but it yields a clean implementability logic and a transparent cutoff characterization of who takes up recourse; later, one can relax it to heterogeneous relative costs at the price of richer menus and computational complexity.

Proxy manipulation carries an arbitrary cost $c_P(m) \geq 0$, which may be heterogeneous across domains but is not scaled by θ in our baseline. Allowing arbitrary c_P underscores that the principal should not rely on proxies being "hard to manipulate": even if proxy manipulation is cheap, the mechanism we build will make it irrelevant for acceptance.

Principal's mechanism: decision rule and recourse policy. The principal commits to a mechanism

$$M = (f, r).$$

The decision rule f maps reported features to an acceptance decision:

$$f : \mathbb{R}^{d_C} \times \mathbb{R}^{d_P} \to \{0, 1\}, \qquad a = f(\hat{x}).$$

The recourse policy r maps a rejected report $\hat{x}$ to a set (or menu) of recommended causal actions:

$$r : \mathbb{R}^{d_C} \times \mathbb{R}^{d_P} \to 2^{\mathbb{R}_+^{d_C}}.$$

We interpret $r(\hat{x})$ as a *guarantee menu*: if the agent were to take any $e \in r(\hat{x})$, then the principal commits that the resulting causal features would satisfy the qualification threshold, and thus (under a suitable eligibility rule) the agent would be accepted. Operationally, this resembles policy statements like "if you complete credential Z, you will be eligible," or "if you raise verified income by X, you will meet the underwriting standard."

In the one-shot environment, recourse is informational: the interaction ends after the principal's decision, but the agent's knowledge of r can affect the initial investment choice because it clarifies the path to acceptance. In a two-stage variant (which we view as a natural interpretation in many institutions), rejected agents may implement a recommended e and reapply; our equilibrium characterization is compatible with either interpretation because what matters is that the principal can credibly commit to accept after verifiable causal improvement.

Timing and information. The interaction proceeds as follows:

1. The principal commits to $M = (f, r)$.
2. The agent observes M and her baseline features x, and privately knows θ.
3. The agent chooses actions (e, m), generating the report $\hat{x} = (x^C + e, \hat{x}^P(m))$.
4. The principal observes $\hat{x}$ and outputs $a = f(\hat{x})$.
5. If $a = 0$, the principal provides the recourse menu $r(\hat{x})$. (In the one-shot model this is advisory; in a two-stage implementation it can be executed and verified.)

The principal observes reported features $\hat{x}$, including $\hat{x}^C$. Crucially, causal changes are assumed verifiable: the principal can condition acceptance on $\hat{x}^C$ without being vulnerable to misreporting of x^C. Proxies may be manipulable in arbitrary ways; the principal does not observe the action m directly, only the induced $\hat{x}^P$.

Preferences. An accepted agent obtains private value $\gamma > 0$ (e.g., the benefit of receiving a loan, job, or access). Agent utility is

$$U_A(e, m; M, x, \theta) \;=\; \gamma\, f(\hat{x}) \;-\; c_C(e; \theta) \;-\; c_P(m). \tag{3}$$

True qualification after investment is

$$y \;=\; \mathbf{1}\{\beta^\top (x^C + e) \geq \tau\} \;=\; \mathbf{1}\{\beta^\top \hat{x}^C \geq \tau\}.$$

We focus on a principal who wishes to maximize true positives—accepting qualified agents—subject to a stringent operational safety constraint against false positives. Formally, the principal's objective can be written as

$$\max_{M=(f,r)} \; \mathbb{E}\big[\, \mathbf{1}\{f(\hat{x}) = 1\} \cdot y \,\big], \tag{4}$$

subject to a no-false-positives condition described below. We view (4) as a reduced form for settings where the principal derives value from serving qualified agents but faces large penalties (regulatory, reputational, or safety-related) for accepting unqualified ones. Extensions can add a capacity constraint $\mathbb{E}[f(\hat{x})] \leq q$ or place weight on agent welfare; our baseline isolates the incentive and safety logic.

Equilibrium concept (Stackelberg with best responses). Given a committed mechanism M and baseline x, an agent with type θ chooses (e, m) to maximize (3). We denote the best-response correspondence by $BR(M, x, \theta)$, with equilibrium actions $(e^\star, m^\star) \in BR(M, x, \theta)$. A Stackelberg equilibrium consists of a mechanism M and best responses $(e^\star, m^\star)$ for all agents such that: (i) agents best respond to M; and (ii) M satisfies the principal's constraints (notably, safety) given induced behavior.

Because the principal is committing ex ante, this is not a Bayesian persuasion problem: we are not designing information structures. Instead, we are designing rules and credible promises about what changes will be rewarded.

Safety and correctness requirements. The institutional commitment we model is that acceptance should be safe: no agent should be accepted unless truly qualified. Since behavior is endogenous, we impose this constraint at the level that matters operationally—on realized reports in equilibrium. Specifically, the *no-false-positives* (No-FP) requirement is:

$$f(\hat{x}) = 1 \;\Rightarrow\; \beta^\top \hat{x}^C \geq \tau, \tag{5}$$

for all $\hat{x}$ that can arise as an equilibrium report under M. This is intentionally stronger than an "in expectation" constraint: a principal in a safety-critical domain often cannot justify knowingly allowing some fraction of unqualified accepts, even if the average outcome is good.

We also require that recourse recommendations be *correct* in the same causal sense. For any rejected $\hat{x}$ and any recommended action $e \in r(\hat{x})$, the recommendation must indeed lead to qualification:

$$\forall \hat{x} \text{ with } f(\hat{x}) = 0, \;\; \forall e \in r(\hat{x}) : \quad \beta^\top (\hat{x}^C + e) \geq \tau. \tag{6}$$

This captures a "no empty promises" principle: the menu cannot include actions that would still leave the agent unqualified. In applications, this corresponds to offering only those improvement plans that would satisfy underwriting, licensing, or safety thresholds if completed.

Notice that (6) speaks only to the causal features $\hat{x}^C$. We do not require recommendations about proxies, both because proxies do not affect qualification and because, in our target use-cases, recommending proxy manipulation would be normatively and operationally dubious.

Discussion of modeling choices. Three modeling choices are worth highlighting because they delimit what the mechanism can and cannot do.

First, qualification is a deterministic threshold in x^C. This is a benchmark for environments with hard eligibility criteria (e.g., meeting a verified income requirement, holding a required credential, passing a safety test). It is not meant to deny that many real outcomes are noisy; rather, it clarifies what is implementable *when* the principal can anchor decisions to verifiable causal criteria.

Second, we assume the principal knows β and τ. This isolates the strategic-design problem from statistical estimation. In many institutions, these parameters correspond to policy-set standards rather than learned coefficients. When they must be estimated, the recourse-induced investments can themselves become a source of identifying variation; we return to this possibility only as an extension.

Third, the single-crossing cost structure (2) is deliberately chosen to make "which improvement should be recommended" a well-defined object. If agents have different relative costs across causal dimensions, then an optimal recourse policy generally requires a menu of options; our baseline can be seen as the simplest environment in which a single recommended action can coordinate investment efficiently.

What the model enables. Within this structure, the principal's design problem becomes sharply articulated. Because proxies can be manipulated without improving y, any rule that rewards proxies risks shifting effort toward wasteful m rather than productive e. Conversely, because causal features are verifiable, the principal can—at least in principle—commit to reward only causal improvements. The recourse policy r is the additional lever that resolves an implementability friction: even if a causal threshold

is safe, agents may not know the cheapest way to reach it, and the principal can use recourse to communicate (and credibly commit to) a minimal path. The next section uses this model to articulate a baseline limitation: why proxy-sensitive score rules naturally invite proxy gaming, and why the no-false-positives requirement—while restrictive—is exactly what makes the safe mechanism-design question policy-relevant.

4 Baseline limits: proxy gaming, and why we restrict to safe mechanisms

Our model is deliberately constructed to isolate a tension that is often blurred in purely predictive treatments of classification: once a decision rule is deployed, it becomes a *target*. Features that were merely correlated with outcomes in historical data need not remain informative when agents can strategically move them. In our setting, the sharpest version of this tension arises because proxy features are, by assumption, outcome-irrelevant: manipulating x^P does not change true qualification y. Any acceptance rule that places weight on x^P therefore creates an immediate wedge between what the principal rewards and what the principal ultimately cares about.

Why standard proxy-sensitive score rules invite proxy gaming. A useful foil is the familiar score-threshold rule

$$f(\hat{x}) = \mathbf{1}\{\tilde{s}(\hat{x}^C, \hat{x}^P) \geq t\},$$

where $\tilde{s}$ could be a learned predictor (logit score, random forest margin, etc.) that loads on both causal and proxy features. Historically, such a rule may be accurate because x^P is correlated with x^C (or with y) in past data. But under strategic response, the agent does not "follow the correlation"; she follows the *gradient of acceptance probability* with respect to the manipulable coordinates.

This has two immediate implications.

First, if $\tilde{s}$ is responsive to proxies and there exists any feasible manipulation m that changes $\hat{x}^P$ in the direction that improves $\tilde{s}$, then proxy manipulation is a privately rational substitute for causal improvement whenever it is cheaper. Formally, holding fixed e, any action m that increases $f(\hat{x})$ weakly increases utility by γ and costs only $c_P(m)$. Thus, unless proxy manipulation is prohibitively expensive *for all* agents, a proxy-sensitive rule generically induces positive proxy spending in equilibrium. This is a pure deadweight loss in our benchmark because it does not improve y.

Second, and more importantly for institutional safety, proxy sensitivity threatens false positives. The core logic is simple: if the acceptance region contains *any* reports with causal score below threshold, then an agent can be

accepted without becoming qualified by steering the proxies into that region. Concretely, suppose there exist two reports $\hat{x} = (\hat{x}^C, \hat{x}^P)$ and $\hat{x}' = (\hat{x}^C, \hat{x}^{P'})$ that share the same causal features $\hat{x}^C$, but

$$\beta^\top \hat{x}^C < \tau, \qquad f(\hat{x}) = 0, \qquad f(\hat{x}') = 1.$$

Then acceptance is achievable "purely through proxies" at that $\hat{x}^C$. Any unqualified agent with baseline $x^C = \hat{x}^C$ (or with x^C that can be brought to $\hat{x}^C$ at negligible causal cost) can obtain γ without changing y by selecting a manipulation m that induces $\hat{x}^{P'}$. If such an m exists with $c_P(m) < \gamma$, it is a best response. In short: once x^P helps, it can be *engineered*.

This observation is the mechanism-design version of Goodhart's law. It is also a Lucas-critique point: the conditional distribution of y given (x^C, x^P) under the old regime does not equal the distribution induced after agents optimize against a proxy-sensitive policy. In applications, this is precisely what we see when (i) "search engine optimization" replaces substantive quality, (ii) "resume keyword stuffing" replaces skill, (iii) temporary balance transfers replace long-run creditworthiness, or (iv) superficial compliance replaces safety culture. In each case, the proxy remains easy to move, and the outcome-relevant state remains unchanged.

Why relying on proxy manipulation costs is fragile. One might object that in many domains proxy manipulation is not free: $c_P(m)$ may be substantial, and perhaps large enough to deter gaming. Our baseline does not deny this; rather, it treats it as an unreliable foundation for safety.

The reason is that the principal typically does not observe m or $c_P(\cdot)$, and cannot rule out low-cost manipulations for some agents or some new manipulation technology tomorrow. If acceptance depends on proxies, safety hinges on an empirical claim of the form "no agent can cheaply move x^P into an accepted region while keeping $\beta^\top \hat{x}^C < \tau$." This is a worst-case statement over a strategic population and an evolving manipulation set. In many regulated settings, that is exactly the kind of claim an institution cannot credibly make.

Moreover, even if manipulation is costly on average, it can still be *selectively* cheap: specialized intermediaries, informational arbitrage, or outright fraud can create a thin but consequential tail of low c_P. Because our safety constraint is pointwise (no false positives), such tails matter. A mechanism that is "mostly safe" is often not operationally acceptable when the harm of a single false positive is large.

Why the no-false-positives constraint is restrictive—but policy-relevant. We now explain why we impose an explicit no-false-positives requirement rather than the more common statistical objective that trades off false positives and false negatives.

The restriction is clear: with noisy outcomes or uncertain measurement, a principal typically cannot guarantee zero false positives. Our benchmark is instead aimed at environments where qualification can be anchored to verifiable causal criteria—exactly the situations in which institutions *do* write hard eligibility rules. Examples include: meeting a certified training requirement, passing a safety inspection, holding a credential, satisfying a verified income threshold, meeting a capital or reserve requirement, or complying with a legal standard. In such cases, "accept only if the standard is met" is not just a preference; it is a compliance obligation.

Formally, our no-false-positives condition can be read as a robust operational constraint:

$$\text{(No-FP)} \qquad f(\hat{x}) = 1 \;\Rightarrow\; \beta^\top \hat{x}^C \geq \tau$$

for reports $\hat{x}$ that occur in equilibrium (and, in the strongest version, for all reports). This is restrictive in the same way that many real constraints are restrictive: it forces the principal to *separate* the decision criterion (causal qualification) from whatever proxies happen to be predictive in a historical dataset.

Two additional remarks clarify the role of this restriction.

First, it is not a moralized "never make mistakes" assumption; it is a modeling device to capture domains where the principal faces effectively infinite (or very large) penalties from false positives. A lender may face binding regulatory constraints; a platform may face catastrophic safety risk; an employer may face licensing rules; a public agency may face statutory eligibility requirements. In these settings, the principal's optimization problem is naturally written as "maximize access subject to meeting the standard."

Second, the restriction is precisely what makes recourse a meaningful policy instrument. If false positives were allowed, the principal could always increase measured performance by relaxing thresholds or by admitting borderline agents probabilistically. Under no-FP, the only way to expand acceptance is to increase the mass of agents who *actually become qualified*. That is, safety converts the problem from one of "better prediction" to one of "better incentives and improvement."

Defining the class of safe mechanisms. We therefore focus attention on mechanisms $M = (f, r)$ that are safe in the strategic sense: they remain free of false positives once agents best respond.

There are (at least) two natural safety notions. The weaker, equilibrium-based notion aligns with how policies are experienced: only realized reports matter. The stronger notion is a robustness requirement: the acceptance rule must never accept an unqualified report, regardless of how it is reached. Because our agents can manipulate proxies in rich ways, the stronger notion is often the relevant one in practice—yet it turns out not to be materially more demanding in our deterministic qualification benchmark.

We package the constraints we impose as follows.

Definition 4.1 (Safe mechanism)**.** A mechanism $M = (f, r)$ is *safe* if, for every baseline x and type θ, and for every best response $(e^\star, m^\star) \in BR(M, x, \theta)$ generating report $\hat{x} = (x^C + e^\star, \hat{x}^P(m^\star))$, the following hold:

1. **No proxy false positives (equilibrium safety):** $f(\hat{x}) = 1 \Rightarrow \beta^\top \hat{x}^C \geq \tau$.
2. **Recourse correctness:** for every rejected report $\hat{x}$ with $f(\hat{x}) = 0$, every recommended action $e \in r(\hat{x})$ satisfies $\beta^\top(\hat{x}^C + e) \geq \tau$.

We write $\mathcal{M}_{\text{safe}}$ for the set of mechanisms satisfying these two properties.

This definition makes explicit what "safe recourse" must mean in a strategic environment: acceptance cannot be earned by proxy-only changes, and recommendations must truly bridge the causal gap to qualification.

A baseline limitation: proxies cannot safely expand acceptance. Once safety is imposed, it is useful to see what is *not* available to the principal. In our benchmark where $y = \mathbf{1}\{\beta^\top x^C \geq \tau\}$ is deterministic in verifiable causal features, any mechanism that ever accepts an agent with $\beta^\top \hat{x}^C < \tau$ is unsafe by definition. Thus, proxies cannot be used to "screen in" additional agents who have not met the causal standard. At best, proxies could be used as a tie-breaker among already-qualified agents, or as a prediction device when the principal is willing to tolerate some false positives. But under our maintained constraint, proxy-based expansions are precisely what is ruled out.

This is the sense in which no-FP is not merely a technical convenience: it sharply distinguishes two worlds. In the permissive world, the principal may treat x^P as helpful signals and accept probabilistically. In the safe world, the principal must treat x^P as potential attack surfaces; any dependence on x^P must be justified by something other than correlation—namely, by causal relevance or by verifiable constraints. Since proxies are outcome-irrelevant by assumption, that justification is unavailable here.

Where recourse enters: incentives rather than signals. The preceding discussion might sound pessimistic: if proxies are off-limits and qualification is a hard threshold, is there anything left to design? The answer is yes, and it is the central point of the paper: even when eligibility must be anchored to causal score, the principal still controls the *path* by which agents reach that score. By committing to a recourse policy that specifies verifiable improvements that guarantee crossing the threshold, the principal can redirect agent effort away from proxy manipulation and toward causal investment. In other words, the lever is not "use more features" but "shape strategic behavior."

The next section makes this precise. We show that, within $\mathcal{M}_{\text{safe}}$, the acceptance rule can be reduced without loss to a threshold on the causal sufficient statistic $\beta^\top \hat{x}^C$, and that under single-crossing costs the principal can provide a particularly simple (indeed, single-action) recourse guarantee that implements maximal safe take-up among improvable agents.

5 Structure: safe sufficiency and the geometry of minimal causal improvement

Having narrowed attention to $\mathcal{M}_{\text{safe}}$, we can now characterize what a safe mechanism can *look like* in our deterministic causal benchmark. Two facts do essentially all of the work going forward. First, if we insist on safety, then the principal cannot gain anything from conditioning acceptance on proxies: the only relevant statistic for eligibility is the causal score $s(x^C) = \beta^\top x^C$. Second, under single-crossing linear costs, the agent's cheapest way to raise this score has an extremely simple structure: push along one "best bang-per-buck" causal coordinate. These two observations jointly explain why a very simple recourse policy will be implementable and (ex post) optimal in the next section.

5.1 Safe sufficiency: eligibility can depend only on the causal score

In our environment, the principal's safety requirement is not merely a constraint on outcomes; it is a constraint on the *shape* of the acceptance region. Intuitively, once an agent can freely choose proxies (or choose them at heterogeneous costs unknown to the principal), any acceptance set that is not "cylindrical" in the proxy coordinates creates an attack surface. Safety forces eligibility to be anchored to what ultimately determines qualification.

We formalize this as a reduction: among safe mechanisms, it is without loss to restrict to rules that are thresholds on $s(\hat{x}^C)$ alone.

Proposition 5.1 (Safe sufficiency / reduction to a causal statistic)**.** *Maintain $y = \mathbf{1}\{\beta^\top x^C \geq \tau\}$ with verifiable $\hat{x}^C$. Consider any mechanism $M = (f, r) \in \mathcal{M}_{\text{safe}}$. Define the induced causal-only rule*

$$\bar{f}(\hat{x}^C) \;=\; \mathbf{1}\Big\{\exists \hat{x}^P \text{ s.t. } f(\hat{x}^C, \hat{x}^P) = 1\Big\}.$$

Then:

1. ***(Safety implies causal feasibility)*** *For all $\hat{x}^C$, $\bar{f}(\hat{x}^C) = 1 \Rightarrow \beta^\top \hat{x}^C \geq \tau$.*

2. ***(W.l.o.g. causal-only eligibility)*** *There exists a safe mechanism $\tilde{M} = (\tilde{f}, r)$ with the same recourse policy and with $\tilde{f}(\hat{x}) = \tilde{f}(\hat{x}^C) =$*

$\bar{f}(\hat{x}^C)$ *such that, for every* (x, θ)*, the set of achievable acceptance decisions under best responses weakly expands relative to* M*, while preserving no-false-positives.*

3. ***(Maximal safe acceptance)*** *Among all safe decision rules, the pointwise maximal acceptance rule is*

$$f^{\text{safe}}(\hat{x}) \;=\; \mathbf{1}\{\beta^\top \hat{x}^C \geq \tau\},$$

i.e., accept all and only reports that are causally qualified.

Interpretation. Part (i) says: if a fixed causal report $\hat{x}^C$ can ever be accepted for *some* proxy realization, then $\hat{x}^C$ must already be on the qualified side of the true threshold. Otherwise the mechanism has created a proxy-contingent "backdoor" into acceptance.

Part (ii) is the key simplification: once we know which causal states are ever admissible, conditioning acceptance on the proxy realization is pure slack from the principal's perspective under safety. Replacing f by $\bar{f}$ deletes arbitrary proxy dependence while never creating an unsafe acceptance.

Part (iii) identifies the "frontier" decision rule within $\mathcal{M}_{\text{safe}}$: if the principal's objective is to maximize true positives subject to no false positives, it is optimal to accept whenever the agent is in fact qualified. Any stricter rule (rejecting some qualified $\hat{x}^C$) can only reduce true positives and does not improve safety.

Proof sketch (economic logic). The proof uses the deterministic structure of qualification and the fact that proxies are outcome-irrelevant.

For (i), fix $\hat{x}^C$ and suppose $\bar{f}(\hat{x}^C) = 1$ while $\beta^\top \hat{x}^C < \tau$. By definition of $\bar{f}$, there exists a proxy report $\hat{x}^P$ with $f(\hat{x}^C, \hat{x}^P) = 1$. But then the mechanism accepts an unqualified report $(\hat{x}^C, \hat{x}^P)$, contradicting safety (in the strong form) and, under equilibrium safety, contradicting the existence of any agent who can reach $(\hat{x}^C, \hat{x}^P)$ at finite proxy cost while keeping causal investment fixed. The point is that proxy dependence can only *move* acceptance across proxies at fixed $\hat{x}^C$; if any such movement crosses the true boundary, an unqualified acceptance becomes feasible.

For (ii), define $\tilde{f}(\hat{x}^C, \hat{x}^P) = \bar{f}(\hat{x}^C)$. By (i), $\tilde{f} = 1$ implies $\beta^\top \hat{x}^C \geq \tau$, hence $\tilde{f}$ is safe regardless of proxy reports. Moreover, if under M the agent can induce acceptance at some report with causal part $\hat{x}^C$, then $\bar{f}(\hat{x}^C) = 1$ and thus $\tilde{f}$ also accepts at that causal report. This removes the need for the agent to thread a particular proxy needle, weakly improving incentives for causal investment (or at least never worsening them) while preserving the same safety guarantee.

For (iii), observe that any safe rule must reject all $\hat{x}^C$ with $\beta^\top \hat{x}^C < \tau$. Hence the pointwise maximal safe acceptance set is exactly $\{\hat{x} : \beta^\top \hat{x}^C \geq \tau\}$.

Accepting all such reports is feasible and yields $f(\hat{x}) = 1 \Rightarrow y = 1$ by construction.

A practical reading. The proposition is a formal version of a compliance intuition: if an institution is obligated to admit only those who meet a verifiable standard, then any reliance on non-standard, manipulable covariates is not "extra information"—it is a vulnerability. In this benchmark, the causal score $s(\hat{x}^C)$ is a sufficient statistic for safe eligibility because it is exactly what the institution ultimately must certify.

Limitations of the reduction. The reduction relies on two knife-edge features that we explicitly adopt as a benchmark: (a) the qualification label is deterministic in $\hat{x}^C$, and (b) β and τ are known and the relevant causal features are verifiable once changed. In settings with noise, imperfect verification, or genuinely causal proxy components, proxy-blindness need not be optimal. Our claim here is narrower: *in the environment where proxies are outcome-irrelevant and manipulable, safety alone forces eligibility to ignore them.*

5.2 Minimal-cost improvement under single-crossing linear costs

Once eligibility is anchored to $s(\hat{x}^C) \geq \tau$, the remaining design problem is not about using additional signals; it is about inducing agents to cross the causal threshold by choosing e. This makes the agent's private optimization problem central. In particular, to understand take-up of any recourse policy, we need the minimal causal cost of reaching qualification.

Fix baseline x^C. If $\beta^\top x^C \geq \tau$, the agent is already qualified and needs no investment. Otherwise, she must increase the score by

$$\Delta(x^C) \;=\; \tau - \beta^\top x^C \;>\; 0.$$

Under our separable single-crossing cost specification $c_C(e;\theta) = \theta \sum_j w_j e_j$, the agent's cost-minimization problem conditional on deciding to qualify is the linear program

$$\min_{e \geq 0} \sum_{j=1}^{d_C} w_j e_j \quad \text{s.t.} \quad \sum_{j=1}^{d_C} \beta_j e_j \;\geq\; \Delta(x^C). \tag{7}$$

This is the simplest possible "effort allocation" problem: the agent needs Δ units of score and can purchase score through different causal coordinates, each with a different price-per-score.

Proposition 5.2 (Closed-form minimal improvement cost and the cheapest causal direction)**.** *Assume* $\beta \in \mathbb{R}_+^{d_C}$ *and* $w \in \mathbb{R}_+^{d_C}$*. Let*

$$j^\dagger \in \arg\min_{j:\beta_j>0} \frac{w_j}{\beta_j} \qquad \textit{(a cheapest cost-per-score coordinate).}$$

Then the minimal unscaled *cost in* (7) *is*

$$\min_{e\geq 0:\, \beta^\top e \geq \Delta} \sum_j w_j e_j \;=\; \frac{w_{j^\dagger}}{\beta_{j^\dagger}}\,\Delta,$$

and an optimal action is to invest only in coordinate $j^\dagger$*:*

$$e^\dagger_{j^\dagger} = \Delta/\beta_{j^\dagger}, \qquad e^\dagger_k = 0 \;\; \forall k \neq j^\dagger.$$

Scaling by type, the minimal causal investment cost to qualify is

$$C^\star(x^C,\theta) = \begin{cases} 0, & \beta^\top x^C \geq \tau, \\ \theta\,\dfrac{w_{j^\dagger}}{\beta_{j^\dagger}}\,(\tau - \beta^\top x^C), & \beta^\top x^C < \tau. \end{cases}$$

Proof sketch (geometry of a one-constraint LP). Because there is a single binding linear constraint $\beta^\top e \geq \Delta$ and nonnegativity, an extreme point solution concentrates all mass on one coordinate. The cost of obtaining one unit of score via coordinate j is exactly w_j/β_j. Hence the cheapest way to buy Δ units is to buy them all through a coordinate minimizing this ratio.

Economic meaning of single-crossing here. The crucial feature is that the identity of $j^\dagger$ does *not* depend on θ. All types agree on the ranking of causal investment directions; θ only scales the level of cost. This is the sense in which the environment is single-crossing: cheaper types are willing to invest more in exactly the same "best" direction.

This seemingly technical property is what makes simple, uniform recourse possible. If different types preferred different directions, then recommending a single action could be badly mismatched for a large share of the population; we return to this menu-complexity issue later. In the present benchmark, the principal can point everyone to the same cheapest causal lever.

A cutoff characterization of who will invest. Because acceptance yields value γ, an agent will choose to incur the minimal qualifying cost if and only if it is privately worthwhile:

$$\gamma \;\geq\; C^\star(x^C,\theta).$$

Equivalently, for $\beta^\top x^C < \tau$,

$$\theta \;\leq\; \underbrace{\frac{\gamma}{\tau - \beta^\top x^C} \cdot \frac{\beta_{j^\dagger}}{w_{j^\dagger}}}_{\text{type cutoff given baseline } x^C} \quad .$$

This simple inequality already anticipates the comparative statics: higher γ increases investment; higher τ reduces it; and a higher "price" $w_{j^\dagger}/\beta_{j^\dagger}$ reduces it. Importantly, under safe eligibility (a pure causal threshold), proxy manipulation never relaxes the constraint $\beta^\top(x^C + e) \geq \tau$, so it cannot change whether the inequality holds.

5.3 What these two results buy us

Proposition 5.1 tells us that, under safety, we can treat the principal's eligibility decision as operating on a one-dimensional sufficient statistic $s(\hat{x}^C)$. Proposition 5.2 then tells us that, under single-crossing costs, the agent's best way to increase that statistic is effectively one-dimensional as well: invest along $j^\dagger$ by an amount exactly equal to the score shortfall divided by $\beta_{j^\dagger}$.

These are the structural ingredients behind implementability. Once (i) eligibility is a causal threshold and (ii) the cheapest improvement path has a closed form shared across types, the principal can do something powerful: commit to a recourse policy that explicitly reveals the unique cheapest qualifying move at any rejected $\hat{x}$. The next section uses this to construct a mechanism $M = (f, r)$ that eliminates proxy-driven acceptance, induces all privately improvable agents to become qualified, and is ex post optimal among safe mechanisms.

6 Implementability: a constructive safe mechanism

The previous section reduced the principal's design problem to an unusually clean object: a one-dimensional eligibility boundary in the causal score $s(\hat{x}^C) = \beta^\top \hat{x}^C$, together with a one-dimensional "cheapest direction" for improving that score under single-crossing linear costs. We now turn these structural facts into a fully specified mechanism $M = (f, r)$ and show that it simultaneously delivers (i) safety against proxy gaming, (ii) full take-up among agents for whom causal improvement is privately worthwhile, and (iii) ex post optimality among safe mechanisms.

The economic idea is simple. If the principal commits to (a) an acceptance rule that is *exactly* the true causal standard and (b) a recourse message that tells rejected agents the *cheapest* verifiable way to meet that standard, then the only remaining friction is the agent's own private cost

scale θ. Agents with low enough θ will rationally invest and become qualified; agents with high θ will rationally decline. Proxy manipulation becomes irrelevant because it cannot move the causal score, and the mechanism never conditions acceptance on proxies in the first place.

6.1 Mechanism definition

Fix $\beta \in \mathbb{R}_+^{d_C}$, $\tau \in \mathbb{R}$, and cost weights $w \in \mathbb{R}_+^{d_C}$. Let

$$j^\dagger \in \arg\min_{j:\beta_j>0} \frac{w_j}{\beta_j}$$

be a cheapest cost-per-score coordinate (ties can be broken arbitrarily and fixed once-and-for-all).

We define the mechanism $M^\dagger = (f^\dagger, r^\dagger)$ by:

$$f^\dagger(\hat{x}) = \mathbf{1}\{\beta^\top \hat{x}^C \geq \tau\}, \tag{8}$$

$$r^\dagger(\hat{x}) = \begin{cases} \{0\}, & \beta^\top \hat{x}^C \geq \tau, \\ \{e^\dagger(\hat{x})\}, & \beta^\top \hat{x}^C < \tau, \end{cases} \tag{9}$$

where the singleton recommendation $e^\dagger(\hat{x}) \in \mathbb{R}_+^{d_C}$ is

$$e^\dagger_{j^\dagger}(\hat{x}) = \frac{\tau - \beta^\top \hat{x}^C}{\beta_{j^\dagger}}, \qquad e^\dagger_k(\hat{x}) = 0 \ \ \forall k \neq j^\dagger. \tag{10}$$

Thus, when an agent is rejected at $\hat{x}$, the mechanism tells her to raise the single cheapest causal coordinate $j^\dagger$ by exactly the amount required to clear the causal threshold.

Operationally, $M^\dagger$ is implementable in time polynomial in d_C: compute $\beta^\top \hat{x}^C$, compare to τ, and (if needed) compute a single scalar gap $(\tau - \beta^\top \hat{x}^C)/\beta_{j^\dagger}$. Importantly, neither $f^\dagger$ nor $r^\dagger$ requires any optimization at runtime.

6.2 Main theorem

Theorem 6.1 (Implementability and ex post optimality of $M^\dagger$)**.** *Maintain the deterministic causal qualification rule $y = \mathbf{1}\{\beta^\top x^C \geq \tau\}$ with verifiable causal features, arbitrary manipulable proxies, and single-crossing linear investment costs $c_C(e;\theta) = \theta \sum_j w_j e_j$. Consider the mechanism $M^\dagger = (f^\dagger, r^\dagger)$ defined in (8)–(10). Then in Stackelberg equilibrium:*

1. ***(Zero proxy false positives)** No accepted agent is unqualified: for all equilibrium reports $\hat{x}$, $f^\dagger(\hat{x}) = 1 \Rightarrow y = 1$. Moreover, proxy manipulation cannot increase acceptance probability under $f^\dagger$.*

2. ***(Full take-up among improvable types)*** *For every (x^C, θ), if the minimal causal cost to qualify satisfies*

$$C^\star(x^C, \theta) = \min_{e \geq 0:\, \beta^\top (x^C + e) \geq \tau} \theta \sum_j w_j e_j \;\leq\; \gamma,$$

then the agent chooses a best response investment that makes her qualified (and hence accepted), and may be taken to be $e^\dagger(x^C)$ concentrating on $j^\dagger$. If $C^\star(x^C, \theta) > \gamma$, she does not invest to qualify.

3. ***(Ex post optimality among safe mechanisms)*** *Fix any alternative safe mechanism $\tilde{M} = (\tilde{f}, \tilde{r})$ that bases eligibility only on verifiable causal features and satisfies no-false-positives in equilibrium. For every realization (x^C, θ), the acceptance/qualification outcome under $M^\dagger$ weakly dominates that under $\tilde{M}$ in true positives. Equivalently, $M^\dagger$ maximizes true positives pointwise subject to safety and incentive compatibility.*

6.3 Why the theorem is true

We prove each part by following the agent's best-response problem under $M^\dagger$ and using the geometry of the minimal-cost improvement characterized earlier.

(i) Safety and irrelevance of proxies. Under $f^\dagger$, acceptance is equivalent to $\beta^\top \hat{x}^C \geq \tau$. But in our environment $y = 1$ is *defined* by the same inequality applied to the realized causal state. Therefore, on the equilibrium path (indeed, for any report),

$$f^\dagger(\hat{x}) = 1 \;\Rightarrow\; \beta^\top \hat{x}^C \geq \tau \;\Rightarrow\; y = h^\star(\hat{x}^C) = 1.$$

This is the strongest possible version of "no false positives": the acceptance region is exactly the qualified region in the causal coordinates.

Proxy manipulation is similarly neutralized by construction. The agent's choice of m affects only $\hat{x}^P$, while $f^\dagger$ depends only on $\hat{x}^C$. Hence m cannot change the acceptance decision, and any m with $c_P(m) > 0$ is strictly dominated by the cheapest proxy action (often "do nothing"). In equilibrium, we can therefore take $m^\star$ to minimize c_P, and proxies drop out of the strategic analysis entirely.

(ii) Full take-up among agents for whom improvement is privately worthwhile. Fix baseline x^C and type θ. The agent chooses $e \geq 0$ to maximize

$$\gamma\, \mathbf{1}\{\beta^\top (x^C + e) \geq \tau\} - \theta \sum_j w_j e_j,$$

since proxy actions do not affect the indicator under $f^\dagger$. There are two cases.

If $\beta^\top x^C \geq \tau$, then the agent is already qualified and accepted with $e = 0$; because costs are nonnegative and separable, any $e \neq 0$ weakly reduces utility. Thus $e^\star = 0$.

If $\beta^\top x^C < \tau$, then the agent faces an all-or-nothing tradeoff: she can remain below threshold and receive utility 0 (net of the proxy baseline), or she can cross the threshold and obtain γ minus the cost of doing so. By Proposition 5.2, the minimal (type-scaled) cost of crossing is exactly

$$C^\star(x^C, \theta) = \theta \, \frac{w_{j^\dagger}}{\beta_{j^\dagger}} \, (\tau - \beta^\top x^C),$$

achieved by investing only in coordinate $j^\dagger$ with magnitude $(\tau - \beta^\top x^C)/\beta_{j^\dagger}$. Therefore:

- If $\gamma \geq C^\star(x^C, \theta)$, then choosing $e = e^\dagger(x^C)$ yields acceptance and nonnegative net gain; any accepted action must satisfy $\beta^\top(x^C + e) \geq \tau$ and hence has cost at least $C^\star$, so $e^\dagger$ is (weakly) optimal among all acceptance-inducing deviations.
- If $\gamma < C^\star(x^C, \theta)$, then every acceptance-inducing investment yields negative net utility, so the best response is to choose $e = 0$ (or any non-qualifying e, which is weakly dominated by $e = 0$ given nonnegativity).

This establishes the cutoff form of equilibrium behavior and the "full take-up" claim: every agent who can be profitably induced to become qualified *does* become qualified under $M^\dagger$.

Where does recourse enter? Formally, $r^\dagger$ is correct by construction: if an agent is rejected at $\hat{x}$, then $\beta^\top(\hat{x}^C + e^\dagger(\hat{x})) = \tau$, so following the recommendation would make her qualified. Economically, the recourse map makes the cheapest qualifying move salient and verifiable: it is an explicit certificate of what the principal will recognize as adequate improvement. In a two-stage "reapply" variant, $r^\dagger$ also directly implements the dynamics (reject $\to$ invest $\to$ accept) with no strategic ambiguity about which improvements count.

(iii) Ex post optimality among safe mechanisms. We now argue that $M^\dagger$ is not merely safe and incentive compatible; it is pointwise optimal in true positives given those constraints.

Take any alternative safe mechanism $\tilde{M}$ that bases eligibility on verifiable causal features. By the definition of safety (no accepted unqualified agents), $\tilde{f}$ cannot accept any report with $\beta^\top \hat{x}^C < \tau$. Hence, for a given realized causal state, $\tilde{M}$ can generate acceptance only by (a) accepting agents who are already qualified or (b) inducing agents to invest until they become qualified, at which point acceptance is permitted.

But incentive compatibility imposes a hard upper bound on whom $\tilde{M}$ can induce to invest. If an agent's minimal qualifying cost exceeds γ, then

even the *cheapest* path to becoming qualified yields negative net payoff, so no mechanism that does not subsidize effort can rationally induce that agent to cross the threshold. Thus, for each (x^C, θ), the set of agents who can possibly become accepted-and-qualified under any safe mechanism is exactly:

$$\Big\{(x^C, \theta) : \beta^\top x^C \geq \tau\Big\} \cup \Big\{(x^C, \theta) : \beta^\top x^C < \tau \text{ and } C^\star(x^C, \theta) \leq \gamma\Big\}.$$

Part (ii) showed that $M^\dagger$ attains acceptance (and hence true positives) for *all* agents in this feasible set, and none outside it. Therefore, for every realization (x^C, θ), no safe mechanism can generate more true positives than $M^\dagger$; this is ex post optimality. Taking expectations over the population immediately yields optimality for the principal's expected true-positive objective.

6.4 What the construction means (and what it does not)

The theorem formalizes a strong "policy design" lesson in this deterministic benchmark: if the institution's obligation is to avoid admitting unqualified agents, then the optimal safe policy is to (i) base acceptance entirely on the verifiable causal standard and (ii) communicate a concrete, minimal, verifiable improvement plan to those who fall short. In applied settings, this corresponds to recourse that is not merely explanatory ("you were rejected because your score was low"), but actionable and minimal ("raise the verifiable causal attribute by exactly this amount").

At the same time, the sharpness of the result is inseparable from the sharpness of the assumptions. We rely on deterministic qualification in verifiable causal features, and on single-crossing linear costs that make the cheapest improvement direction common across types. Once relative costs differ across individuals—so that different agents prefer different "improvement levers"—a singleton recourse recommendation need no longer be close to optimal. The next section takes up that issue directly by treating recourse design as a menu-selection problem, clarifying when $K = 1$ suffices, when $K = d_C$ guarantees coverage, and why intermediate K generally becomes a submodular optimization problem rather than a closed-form construction.

7 Extensions: menu complexity when relative costs vary across agents

The construction in the previous section leans heavily on a knife-edge but economically transparent property: *single-crossing* makes the cheapest way to increase the causal score agree across types. Once we relax that property, recourse design stops being "one message fits all." In many applied environments, different individuals can improve different verifiable causal attributes at very different relative costs (time constraints, liquidity constraints, access

to training, disability accommodations, local availability of courses, etc.). In that case, a principal who insists on giving only a single recommended action can easily leave many *improvable* agents behind, not because they cannot meet the standard, but because the recommended path is not their cheapest.

We formalize this by replacing the separable single-index cost scale with heterogeneous relative costs. Let a type now be a vector of marginal costs $\kappa \in \mathbb{R}_+^{d_C}$ and let

$$c_C(e;\kappa) = \sum_{j=1}^{d_C} \kappa_j e_j.$$

The qualification constraint remains $\beta^\top(x^C+e) \geq \tau$, and acceptance remains "safe" only if it never admits a report with $\beta^\top \hat{x}^C < \tau$. The key difference is that the optimal improvement direction now depends on κ: the relevant "price per unit causal score" for coordinate j is κ_j/β_j, and the cheapest coordinate can vary widely across agents.

7.1 Recourse as a finite menu

A useful way to think about recourse in this environment is as a *menu* of verifiable causal action plans. Upon rejection at $\hat{x}$, the principal reveals a finite set $r(\hat{x}) = \{e^{(1)}(\hat{x}), \ldots, e^{(K)}(\hat{x})\}$. The agent then chooses whether to follow any element of the menu (in a two-stage variant, to reapply after doing so), or to deviate to some other action. Correctness requires that every recommended action clears the causal threshold:

$$\forall e \in r(\hat{x}): \qquad \beta^\top(\hat{x}^C + e) \geq \tau.$$

Because eligibility is pinned to $\beta^\top \hat{x}^C$, the role of the menu is not to "justify" rejection but to reduce *search and uncertainty* about which verifiable improvements will be recognized as sufficient, and (if we restrict attention to menu-following behavior) to provide a small set of candidate paths among which the agent can select her cheapest.

To make the menu problem crisp, fix a rejected $\hat{x}$ with score gap

$$\Delta(\hat{x}) \equiv (\tau - \beta^\top \hat{x}^C)_+ > 0.$$

Any recourse action must satisfy $\beta^\top e \geq \Delta(\hat{x})$. For a given menu $r(\hat{x})$, an agent of type κ will (weakly) prefer the cheapest menu element that achieves acceptance, i.e.

$$\min_{e \in r(\hat{x})} \kappa^\top e \;\leq\; \gamma \quad \Longrightarrow \quad \text{take up recourse (and become qualified).}$$

Thus, for each $\hat{x}$, the menu induces a "covered set" of types: those for whom at least one recommended plan costs at most γ. This immediately highlights

why heterogeneity creates menu complexity: a single plan can be cheap for some types and expensive for others, even when *everyone* shares the same causal threshold $\Delta(\hat{x})$.

7.2 Two polar cases: $K = 1$ and $K = d_C$

Case 1: $K = 1$ (a single recommendation). Suppose we insist on a singleton menu $r(\hat{x}) = \{e(\hat{x})\}$ with $\beta^\top e(\hat{x}) \geq \Delta(\hat{x})$. Under heterogeneous κ, there is no longer a universally optimal direction. Any fixed $e(\hat{x})$ implicitly privileges some cost profiles over others.

A natural (though not universally optimal) benchmark is to recommend a *population-optimal* direction, e.g. choose

$$e^{\text{pop}}(\hat{x}) \in \arg \min_{e \geq 0:\, \beta^\top e \geq \Delta(\hat{x})} \mathbb{E}[\kappa^\top e \mid \hat{x}],$$

which, by linear programming, concentrates all effort on a coordinate

$$j^{\text{pop}} \in \arg\min_j \frac{\mathbb{E}[\kappa_j \mid \hat{x}]}{\beta_j}, \qquad e^{\text{pop}}_{j^{\text{pop}}}(\hat{x}) = \frac{\Delta(\hat{x})}{\beta_{j^{\text{pop}}}}, \;\; e^{\text{pop}}_{k \neq j^{\text{pop}}}(\hat{x}) = 0.$$

This is the direct analogue of the single-crossing construction, except that "cheapest" is now defined in expectation rather than uniformly across types. The limitation is immediate: even if this minimizes average cost, it may deliver poor take-up among types whose cheap coordinate is different. In other words, $K = 1$ can remain safe and simple, but it is generally *not* fully implementable in the sense of inducing all privately improvable agents to become qualified.

Case 2: $K = d_C$ (coordinate menu, closed form). At the other extreme, a simple menu can guarantee broad coverage even under arbitrary heterogeneity by offering one option per causal coordinate. Define coordinate actions

$$e^{(j)}(\hat{x}) \;\text{ by }\; e^{(j)}_j(\hat{x}) = \frac{\Delta(\hat{x})}{\beta_j}, \qquad e^{(j)}_{k \neq j}(\hat{x}) = 0,$$

for every j with $\beta_j > 0$, and set

$$r^{\text{coord}}(\hat{x}) = \{e^{(j)}(\hat{x})\}_{j:\beta_j > 0}.$$

This menu is correct by construction: each option clears the threshold with equality. Moreover, for any type κ, the cheapest coordinate plan in the menu has cost

$$\min_{j:\beta_j>0} \kappa^\top e^{(j)}(\hat{x}) = \Delta(\hat{x}) \cdot \min_{j:\beta_j>0} \frac{\kappa_j}{\beta_j}.$$

But that expression is exactly the value of the underlying linear program

$$\min_{e\geq 0:\, \beta^\top e \geq \Delta(\hat{x})} \kappa^\top e,$$

since the feasible region is a single covering constraint with nonnegativity, so an extreme point solution is always a single coordinate. Hence the coordinate menu does more than provide "some" good options: it includes an agent's *globally minimal-cost* way to qualify. As a result, in a two-stage variant where agents reapply after following a recommended plan, r^{coord} restores the full-take-up property: every agent who can profitably become qualified (i.e. whose minimal qualifying cost is at most γ) can do so by selecting the appropriate coordinate option.

The practical message is that a menu of size d_C can be enough to recover the sharp implementability conclusion without any single-crossing restriction. Importantly, the menu remains operationally simple: each option is "increase feature j by $\Delta(\hat{x})/\beta_j$." The tradeoff is communicational and administrative rather than computational: as d_C grows, presenting and validating many distinct improvement pathways may be burdensome.

7.3 Intermediate K: coverage and submodular structure

Most institutional settings live between these extremes. We might be willing to offer K options, but not d_C, and we want to choose which options to include to maximize the number of agents who will take up recourse (and hence become qualified) subject to safety.

To see the combinatorial structure cleanly, it helps to discretize candidate actions. Consider a finite library $\mathcal{E}(\hat{x}) = \{e^{(1)}(\hat{x}), \ldots, e^{(N)}(\hat{x})\}$ of correct actions (each clears τ). A menu is then a subset $S \subseteq \{1, \ldots, N\}$ with $|S| \leq K$. For a given type κ, define the minimal menu cost

$$\text{cost}(\kappa; S, \hat{x}) \equiv \min_{i \in S} \ \kappa^\top e^{(i)}(\hat{x}),$$

with the convention that the minimum is $+\infty$ if $S = \varnothing$. Take-up occurs when $\text{cost}(\kappa; S, \hat{x}) \leq \gamma$. Let $G(S)$ denote the expected take-up (or expected true positives induced) from menu S, e.g.

$$G(S) = \mathbb{E}\Big[\mathbf{1}\{\text{cost}(\kappa; S, \hat{x}) \leq \gamma\} \,\Big|\, \hat{x}\Big],$$

or its sample-average analogue over a dataset of observed κ-draws or proxy clusters.

Two general facts are worth emphasizing.

First, selecting the best size-K menu is in general computationally hard. When the library consists of coordinate actions and types effectively have "acceptable" subsets (those coordinates for which $\Delta(\hat{x})\kappa_j/\beta_j \leq \gamma$), the menu

design problem becomes a form of *maximum coverage*: choose K coordinates to cover as many types as possible. This inherits NP-hardness in the worst case, so we should not expect a closed-form analogue of $j^\dagger$ for general K.

Second, despite hardness, the objective has a helpful diminishing-returns property. When agents choose the cheapest available option, the incremental benefit of adding a new option to a larger menu is typically smaller than adding it to a smaller menu. Under mild regularity (e.g. working with a smoothed take-up probability $\Pr(\text{cost} \leq \gamma)$ or a welfare proxy like $-\mathbb{E}[\min\{\text{cost}, \gamma\}]$), $G(S)$ is monotone and (approximately) submodular in S. This places menu design in the well-studied class of submodular maximization problems, for which a greedy algorithm yields a constant-factor approximation: iteratively add the option with the largest marginal gain

$$i_t \in \arg\max_{i \notin S_{t-1}} \big(G(S_{t-1} \cup \{i\}) - G(S_{t-1})\big),$$

until $|S_t| = K$. In the canonical monotone submodular case, this achieves a $(1 - 1/e)$-approximation to the optimal menu value. The economic interpretation is appealing: each additional recourse option "covers" a set of types for whom that option is (sufficiently) cheap, and overlap in coverage generates diminishing returns.

7.4 What is gained and what is lost by limiting K

This extension clarifies a design frontier that is easy to miss in the single-crossing benchmark. Safety continues to push us toward an acceptance rule depending only on verifiable causal features, but implementability now comes in degrees: the principal can trade off the simplicity of recourse communication against the breadth of types who can find a personally inexpensive path to qualification.

When K is small, the principal effectively chooses which improvement pathways to "institutionalize." This can be normatively fraught: if different demographic groups face systematically different cost vectors κ (because of access, geography, discrimination, or wealth constraints), then a small menu can generate disparate take-up even though the acceptance rule is causally correct. In that sense, menu complexity is not merely a technicality—it is a channel through which seemingly neutral standards interact with heterogeneous constraints.

At the same time, offering a very large menu is not free. Beyond cognitive load for agents, each option may require administrative verification, auditing, and enforcement (to preserve verifiability and prevent proxy substitution masquerading as causal improvement). Thus, even in a world where $K = d_C$ is theoretically attractive, real institutions may rationally choose intermediate K, at which point approximation methods and empirical calibration become relevant.

Finally, we stress a limitation of the coordinate-menu conclusion: it relies on the structure of the qualification constraint as a single linear threshold $\beta^\top(x^C + e) \geq \tau$ and on linear costs. If causal qualification depends on multiple constraints (e.g. minimum requirements in several dimensions) or costs are nonlinear (fixed costs, complementarities, or capacity constraints), then even $K = d_C$ may fail to contain each agent's globally cheapest qualifying action. In such richer models, menu design becomes genuinely multidimensional, and the computational boundary becomes sharper.

In sum, once relative costs are heterogeneous, recourse design shifts from a closed-form prescription to a menu-selection problem: $K = 1$ is simple but generally leaves attainable true positives unrealized; $K = d_C$ restores full coverage in our linear-threshold benchmark; and intermediate K naturally leads to coverage-like, submodular optimization where greedy methods provide principled approximations. This sets the stage for the next extension, where the principal may not even know (β, τ) and must learn the causal standard from the strategic improvements agents undertake in response to recourse.

8 Extensions: learning the causal standard when (β, τ) are unknown

Thus far we have treated the causal qualification rule

$$y = h^\star(x^C) = \mathbf{1}\{\beta^\top x^C \geq \tau\}$$

as known to the principal. This is the right benchmark for isolating the strategic role of proxies and the logic of "safe sufficiency." But in many real deployments the principal does *not* know the causal weights β nor the threshold τ ex ante. The institution may know which features are causally relevant and verifiable (e.g., completed courses, certified skills, lab measurements), yet be uncertain about (i) how these features trade off, and (ii) where the true boundary between qualified and unqualified lies.

Once we admit this uncertainty, we face a familiar econometric tension in an unfamiliar strategic wrapper: we want to *learn* the causal decision boundary from observed data, but the data are *endogenously shaped* by the very rule and recourse guidance we deploy. In other words, we are in a setting of *strategic regression*: observed post-decision covariates $x^C + e$ are not passively drawn; they are equilibrium objects. The key question for this extension is whether the principal can nevertheless identify and learn (β, τ) *from observed strategic improvements*, and if so how this interacts with the "no false positives" safety constraint.

8.1 Why strategic improvements are informative (intuition)

Our baseline mechanism does more than classify agents; it also induces some agents to move in x^C by choosing costly investment e. These movements are not arbitrary. They are disciplined by (a) the geometry of the qualification set $\{\beta^\top x^C \geq \tau\}$, and (b) the agent's optimization problem, which under linear costs yields corner solutions and tight satisfaction of constraints.

This generates a useful "revealed boundary" idea: when an agent invests *just enough* to become qualified, her post-investment causal features x^C+e lie *on* (or very near) the true boundary $\beta^\top x^C = \tau$. If we can observe a collection of such boundary points spanning the space, then learning (β, τ) becomes conceptually similar to learning a separating hyperplane—except the points were endogenously produced by incentives rather than exogenously sampled.

The role of recourse is pivotal here. A well-designed recourse policy does not merely "tell agents what to do"; it creates structured variation in e and reduces slackness (over-investment). In effect, recourse can turn strategic behavior into an identification device.

8.2 A simple identification sketch under single-crossing and a two-stage variant

To keep the sketch crisp, suppose we are in the single-crossing benchmark $c_C(e;\theta) = \theta \sum_j w_j e_j$ (unknown θ, known w), and consider a two-stage implementation where a rejected agent may follow a recommended action and reapply. Suppose further that the principal can observe: baseline x^C, chosen investment e, and an eventual outcome label y (e.g., an externally realized "success" indicator that corresponds to true qualification).[1]

Under the ground truth, the set of causal feature vectors with $y = 1$ is exactly the halfspace $\{\beta^\top x^C \geq \tau\}$. If agents sometimes arrive with $y = 1$ already (i.e., $\beta^\top x^C \geq \tau$ at baseline), these provide positive examples; if some arrive with $y = 0$ and do not invest, these provide negative examples. That is standard classification data.

The distinctive feature here is that we also observe strategic transitions $x^C \mapsto x^C + e$. Under the linear program

$$\min_{e \geq 0} \sum_j w_j e_j \quad \text{s.t.} \quad \beta^\top (x^C + e) \geq \tau,$$

optimality generically implies *tightness* at the constraint for invest-then-qualify agents:

$$\beta^\top (x^C + e) = \tau,$$

[1] Without some source of labels, learning is ill-posed: if the principal never observes whether an agent is truly qualified, (β, τ) are not identified from actions alone because many boundaries can rationalize the same accept/reject decisions. In practice, labels may come from repayment, job retention, audits, exams, or delayed performance measures.

because any slack would waste cost. Hence each such agent supplies a linear equation in (β, τ) evaluated at an observed point $x^C + e$. With enough distinct boundary points (and a normalization for scale, e.g. $\|\beta\|_1 = 1$ or fixing one coordinate of β), (β, τ) are identified in principle.

The conceptual "mild richness" requirement is that the distribution of these boundary points has support that is not confined to a lower-dimensional set. Put informally: we need enough variation in x^C among those who invest, so that the induced boundary points are not all collinear in feature space.

8.3 Learning the direction of β via recourse-induced experiments

In practice, the principal cannot rely on boundary points "appearing on their own," especially early on when the deployed rule may be misspecified and agents' incentives may generate selection. A natural remedy is to use the recourse channel as a controlled source of variation—analogous to an instrument in causal inference.

One particularly transparent approach is to occasionally deploy *multiple* correct candidate improvement directions (or approximate directions) and observe which one agents choose and how outcomes respond. Concretely, suppose that on a small fraction of rounds the principal posts a recourse menu of K alternative action plans $\{e^{(1)}, \dots, e^{(K)}\}$, each of which is verifiable and moves different coordinates of x^C. Let $Z \in \{1, \dots, K\}$ denote the index of the recommended plan emphasized (or randomly highlighted), and let e be the realized investment. Variation in Z shifts e through the agent's best response but (by construction) does not directly affect the true outcome except through $x^C + e$.

This is the strategic analogue of a first stage:

$$e = \mathrm{BR}(x^C, Z, \theta) \quad \Rightarrow \quad x^{C\prime} = x^C + e,$$

followed by a structural second stage

$$y = \mathbf{1}\{\beta^\top x^{C\prime} \geq \tau\}.$$

With sufficient exploration over Z and support in baseline x^C, one can in principle recover the orientation of β (up to scale) from how outcome probabilities change with induced movements in different directions. In the deterministic-threshold model, the relevant variation is not in smooth probabilities but in which induced moves cross the boundary.

A pragmatic way to implement this is to reduce the learning target to a finite-dimensional parameter estimation problem and fit (β, τ) by enforcing consistency with observed outcomes:

$$y_i = 1 \implies \beta^\top (x_i^C + e_i) \geq \tau, \qquad y_i = 0 \implies \beta^\top (x_i^C + e_i) < \tau,$$

and then use a large-margin or maximum score criterion (e.g., a hinge-loss objective) to handle inevitable slack, noise, and discreteness:

$$\min_{\beta,\tau} \sum_i \ell\big(y_i(\beta^\top(x_i^C + e_i) - \tau)\big) \quad \text{s.t.} \quad \beta \geq 0,\ \|\beta\|_1 = 1.$$

What makes this "strategic regression" rather than ordinary regression is that the regressor $x_i^C + e_i$ is endogenous to the deployed mechanism. The justification for the above fitting step is therefore structural: we are explicitly modeling (and exploiting) the equilibrium mapping from policy to e_i, rather than assuming i.i.d. covariates.

8.4 Safety during learning: conservative acceptance and robust recourse

Learning raises a constraint interaction that is easy to overlook. Our earlier safety guarantee relied on the principal *knowing* the true score $\beta^\top x^C$ and threshold τ, so that the acceptance rule $f(\hat{x}) = \mathbf{1}\{\beta^\top \hat{x}^C \geq \tau\}$ was pointwise safe. If (β, τ) are uncertain, naively plugging in estimates $(\hat{\beta}, \hat{\tau})$ can produce false positives.

A standard remedy in "safe learning" is to maintain a confidence set $\mathcal{C}_t$ of plausible parameters given data up to time t, and to accept only when qualification is guaranteed for all parameters in the set:

$$f_t(\hat{x}) = \mathbf{1}\Big\{ \min_{(\beta,\tau)\in\mathcal{C}_t} \big(\beta^\top \hat{x}^C - \tau\big) \geq 0 \Big\}.$$

If $\mathcal{C}_t$ is a valid (high-probability) confidence region, this ensures a high-probability analogue of no false positives. The cost is conservatism: early on, $\mathcal{C}_t$ is large, and acceptance may be rare.

Recourse must also be adapted. "Correctness" of recourse actions now means: recommended actions should (with high probability) lead to true qualification, not merely estimated qualification. A robust counterpart is

$$r_t(\hat{x}) \subseteq \Big\{ e \geq 0 : \min_{(\beta,\tau)\in\mathcal{C}_t} \big(\beta^\top(\hat{x}^C + e) - \tau\big) \geq 0 \Big\}.$$

This robustification again trades off sharpness for safety: to guarantee qualification uniformly over $\mathcal{C}_t$, the recommended e may be larger than what is truly necessary for the realized (β, τ).

The economic point is that we should expect a *safety–learning* tradeoff: insisting on strict no-false-positives while parameters are unknown forces either (i) conservative acceptance (fewer opportunities to observe y), or (ii) conservative recourse (higher induced investment costs), or both. In many applications, institutions manage this tradeoff implicitly via pilot programs, staged rollouts, audits, or temporary human review; our framework makes the channel explicit.

8.5 Connecting to instrumental variables and "strategic IV"

It is useful to relate the above to instrumental variables, because the core empirical difficulty is endogeneity of $x^C + e$. In observational data, people invest in skills, health, or creditworthiness based on unobserved traits that also affect outcomes; naive regression confounds these channels. In our mechanism design setting, we can sometimes *create* exogenous shifts in investment incentives through recourse, and thereby generate a policy-driven instrument.

Formally, let Z be a randomized recourse recommendation (or subsidy, or deadline) that affects the agent's chosen e but is independent of the agent's latent determinants of y except through $x^C + e$. When such exclusion is plausible, Z provides leverage to identify the causal effect of moving different coordinates of x^C on the outcome boundary. While the ground-truth model here is a linear threshold rather than a linear outcome equation, the logic is parallel: recourse-induced variation can play the role of a first stage that is both policy-relevant and verifiable.

This "strategic IV" interpretation also clarifies why proxies are a distraction in the learning problem. Because proxies do not affect y, any proxy shifts induced by manipulation are invalid instruments. By designing the mechanism to focus on verifiable causal features, we align the variation we create (and the variation agents choose) with the outcome-relevant state, which is precisely what identification requires.

8.6 Limitations and what we would need for a full theorem

The discussion above is intentionally a sketch, because turning it into a theorem requires committing to additional structure that is context dependent:

(i) **Outcome observability and delay.** If y is only observed for accepted agents (selective labels), learning must address selection. Recourse can help by moving some rejected agents into acceptance and thereby revealing their y, but the induced sample is still policy-dependent.

(ii) **Noise and model misspecification.** Real outcomes are noisy and rarely exact thresholds. With stochastic y (e.g., $y \sim \text{Bernoulli}(\sigma(\beta^\top x^C - \tau))$), learning becomes smoother but safety becomes inherently probabilistic; exact no-false-positives is then unattainable without rejecting almost everyone.

(iii) **Nonlinear qualification and multi-constraint standards.** If qualification is not a single halfspace—e.g., it involves complementarities, minimum requirements in multiple dimensions, or nonconvexities—then boundary-point logic and linear identification can fail, and recourse-induced experiments must be redesigned.

(iv) **Equilibrium shifts during learning.** Because agents respond to the mechanism, the data-generating process shifts as estimates change. A

full analysis would treat this as an adaptive control problem: policy affects behavior, behavior affects data, data affect policy.

Despite these caveats, the central message of this extension is robust: strategic improvements are not merely an obstacle to learning; when investments are verifiable and structured by recourse, they can be an informative signal about the causal standard itself. Recourse is therefore dual-purpose: it is a tool for enabling qualified behavior *and* a tool for uncovering the qualification boundary that the institution ultimately cares about.

9 Welfare and policy interpretation: what "no false positives" buys, when to relax it to ε-FP, and recourse as regulated improvement pathways

Our baseline results were stated in the language of implementability and safe maximization of true positives. In practice, however, the constraint we impose—*no false positives in equilibrium*—is not merely a technical convenience. It is a normative and institutional choice about which mistakes are admissible, how much strategic behavior we are willing to tolerate, and what kind of accountability an automated (or semi-automated) decision system should provide.

What the no-false-positives constraint buys. The most direct interpretation is *risk containment*: acceptance is a form of permission (to borrow, to enroll, to be hired, to access a scarce medical intervention), and a false positive is an instance of permission granted to someone who is not, in the relevant causal sense, qualified. When the downstream harm of such permission is large or external—default risk in credit, safety risk in critical jobs, clinical risk in medicine, congestion in scarce programs—institutions routinely adopt a hard safety constraint that prioritizes avoiding false positives over minimizing false negatives. Our equilibrium no-FP constraint is exactly this logic: it insists that, taking strategic responses seriously, the mechanism should never induce an unqualified agent to be accepted.

A second benefit is *strategic robustness*. Once proxies are manipulable, any eligibility rule that (even accidentally) rewards proxies creates a wedge between the agent's private optimization problem and the principal's causal objective. That wedge produces two social costs. First, it creates *direct deadweight loss* through wasteful proxy manipulation: agents pay $c_P(m)$ for actions that do not change y. Second, it creates an *indirect adverse-selection channel*: the mechanism ends up selecting agents who are best at manipulating proxies (low c_P) rather than those who can cheaply become causally qualified (low c_C). The safe-sufficiency logic can therefore be read as a welfare statement: by designing acceptance to depend only on verifiable causal

score, we shut down a rent-seeking technology and align private incentives with socially relevant improvement.

Third, the constraint buys a form of *procedural legitimacy*. If acceptance is guaranteed to imply true qualification (relative to the institution's articulated causal standard), then an acceptance decision is defensible as rule-following rather than discretionary or opaque. This matters in regulated domains where decision-makers must justify approvals to auditors, courts, or the public. In such settings, the relevant question is often not "did we maximize accuracy?" but "can we certify that approvals satisfy a stated standard?" Our model formalizes one route to certification: if $f(\hat{x}) = 1 \Rightarrow \beta^\top \hat{x}^C \geq \tau$ holds pointwise in equilibrium, then the institution can credibly claim that approvals meet the causal threshold, regardless of the strategic ingenuity of applicants.

The welfare cost of strict safety: false negatives and costly compliance. The flip side is that strict no-FP necessarily tolerates false negatives. Even in the deterministic benchmark where $y = \mathbf{1}\{\beta^\top x^C \geq \tau\}$, false negatives arise because (i) some agents are qualified but may be rejected if the mechanism is conservative for other reasons (e.g., capacity constraints), and (ii) some agents could become qualified but do not invest because $c_C(e;\theta) > \gamma$. The latter is particularly salient: our equilibrium cutoff $\gamma \geq C^\star(x^C, \theta)$ is privately rational, but it need not be socially efficient. If qualifying generates positive externalities (more productive workers, healthier patients, lower default risk that benefits the pool), society may want more investment than individual agents are willing to undertake.

This observation reframes recourse. In our mechanism, recourse is informational: it reveals a cheapest verifiable improvement path. From a welfare perspective, it is also a *compliance technology* that can reduce waste (by preventing over-investment and proxy manipulation) while potentially increasing productive investment. Yet it can shift burdens onto agents: when the "safe path" requires costly investments, recourse resembles a regulated hurdle. Whether that is acceptable depends on a policy judgment about who should bear qualification costs. A planner who internalizes social surplus might instead consider subsidies, financing, or institutional provision of the required investments. In our notation, one can imagine augmenting agent utility with transfers $T(e)$ (e.g., training vouchers, matched savings, health subsidies), effectively lowering the private marginal cost w and expanding the set $\{(x^C, \theta) : C^\star(x^C, \theta) \leq \gamma\}$ who take up recourse. The key point is that no-FP does not by itself resolve distributive concerns; it primarily disciplines *which* behaviors are rewarded.

When to relax no-FP to ε-FP. There are at least three reasons an institution may rationally relax strict no-FP.

(1) Inherently noisy qualification. In many environments, there is no deterministic $h^\star$: even with perfect causal features, outcomes are stochastic. Then "no false positives" is either infeasible or vacuous. A natural replacement is a probabilistic safety constraint, such as

$$\Pr(y = 0 \mid f(\hat{x}) = 1) \leq \varepsilon \quad \text{or} \quad \Pr(f(\hat{x}) = 1,\, y = 0) \leq \varepsilon,$$

for a tolerance level $\varepsilon > 0$. This is not merely a technical change; it shifts the institution from certifying qualification to managing risk. In such settings, the design problem resembles constrained classification (or risk-limiting decision rules), and recourse becomes a means of *increasing the conditional success probability* among the accepted by inducing movement in causal features.

(2) Unverifiable or partially verifiable causal features. Our baseline safety argument leans heavily on verifiability of $x^C + e$. When causal features are measured with error, can be strategically misreported, or are only observed through noisy tests, insisting on pointwise no-FP may force extreme conservatism. Allowing ε-FP can be understood as acknowledging measurement limits: the institution commits to keeping error within a regulated tolerance rather than pretending it can certify a sharp boundary.

(3) High social cost of false negatives. Some domains place asymmetric weight on missed opportunities. For example, denying access to an education program or preventive health service may have large long-run costs. If these costs exceed the expected harms from occasional false positives, the welfare optimum may involve an interior tradeoff. A convenient way to express this is to replace the hard constraint by a penalty (a Lagrangian relaxation):

$$\max_{M} \; \mathbb{E}\big[\mathbf{1}\{f(\hat{x}) = 1\} \cdot y\big] \;-\; \lambda \cdot \mathbb{E}\big[\mathbf{1}\{f(\hat{x}) = 1\} \cdot (1 - y)\big],$$

where λ encodes the marginal social cost of a false positive. As $\lambda \to \infty$ we recover the hard no-FP regime; for finite λ the mechanism may rationally accept some borderline cases.

Importantly, relaxing to ε-FP reopens the door to proxy dependence. If a proxy is predictive of y (even if not causally relevant), a purely statistical designer might want to use it. Our framework suggests a cautionary corollary: when proxies are manipulable, the *predictive* value of a proxy does not imply *mechanism* value. A proxy that is informative under passive observation can become actively misleading once it is rewarded. Thus, even under ε-FP, a robust policy stance is to use proxies only in ways that do not create strong manipulation incentives—for example, as inputs to auditing, to random checks, or to post-acceptance monitoring rather than to eligibility itself.

Recourse menus as regulated improvement pathways. A central practical contribution of recourse is that it turns an abstract eligibility

threshold into a *pathway*—a concrete, verifiable set of actions that, if taken, will lead to acceptance. This is closely aligned with how institutions already operate in regulated settings.

In consumer credit, for example, "credit-building" programs implicitly define acceptable pathways: make on-time payments for T months, reduce utilization below a threshold, establish a secured card, or complete a counseling program. These are not merely tips; they are quasi-regulatory standards that shape what behaviors are legible and rewarded. In workforce policy, training and credentialing programs play the same role: complete a certified course, pass an exam, log supervised hours. In health, compliance pathways are ubiquitous: adhere to a medication regimen, achieve lab targets, attend follow-up visits. In each case, the institution is not just predicting outcomes; it is *governing* the space of improvements.

Seen this way, a recourse menu $r(\hat{x})$ is a form of *regulated action set*. It specifies which changes count, how they will be verified, and how much improvement is needed. This perspective clarifies both the promise and the risk.

The promise is *reduction in arbitrary barriers*. If the mechanism is safe-sufficient, then the pathway focuses attention on causally relevant, verifiable improvements and makes explicit that proxy cosmetics (presentation, gaming, performative signals) are irrelevant. This can equalize access to the "rules of the game," especially for agents who lack informal knowledge. It also mitigates the "moving target" problem: if agents fear that standards will change after they invest, they may underinvest; a credible recourse pathway is a commitment device.

The risk is *paternalism and mismatch*. A menu is inevitably incomplete, particularly when relative costs are heterogeneous (as in our menu-complexity extension). A narrow menu may force some agents into inefficient improvement directions, raising their compliance costs and potentially exacerbating inequality. More subtly, regulated pathways can crowd out innovation: if only certain credentials or programs are recognized, alternative ways of becoming qualified may be ignored even if causally effective. This is a familiar tension in licensing and accreditation. Our model isolates the mechanism-design analogue: a smaller menu is simpler and easier to verify, but it may leave welfare on the table for types whose cheapest causal route is not offered.

Policy levers: shifting the burden and shaping incentives. Interpreting recourse as regulated pathways highlights a menu of policy levers beyond the acceptance rule itself. Institutions can (i) subsidize specific pathway actions (lowering effective w), (ii) expand the menu (increasing matching to heterogeneous types), (iii) finance investments (turning up-front costs into contingent repayments), and (iv) provide complementary supports (informa-

tion, coaching) that reduce non-monetary frictions not captured by θ. Each lever changes equilibrium take-up without relying on proxies.

Finally, strict no-FP should be understood as a *commitment about what will not be rewarded.* In a world where applicants can and will respond, that commitment has independent value: it channels effort away from manipulation and toward improvements that society can defend as genuinely relevant. Whether one ultimately adopts strict no-FP or an ε-FP relaxation depends on domain-specific harms, noise, and institutional capacity for verification and support. But in either regime, recourse is not an afterthought; it is the interface through which an eligibility standard becomes an implementable, governable, and contestable policy.

10 Discussion and open problems: verifiability, measurement error, dynamics, group constraints, and deployment

Our main theorem is intentionally clean: deterministic qualification, verifiable causal features, and a separable single-crossing cost technology yield a mechanism that is simultaneously safe (no proxy false positives), simple (a score threshold), and constructive (a concrete recourse action). The clarity is useful precisely because it highlights where the engineering and policy difficulty actually lies. In this section we discuss the most consequential gaps between the benchmark and real deployments, and we flag open problems that, in our view, determine whether "safe recourse" is a theoretical curiosity or a practical governance tool.

1. Verifiability is the hinge: what does it mean, and who bears the burden? The benchmark assumes that once an agent changes a causal feature, the principal observes it without strategic distortion. In practice, verifiability is neither binary nor free. It is an institutional arrangement involving (i) a measurement technology (tests, transcripts, payroll records, lab results), (ii) an auditing or attestation process (third parties, document checks, tamper resistance), and (iii) a dispute-resolution protocol when evidence is contested.

A first open problem is to model *partial verifiability.* Suppose a causal feature is only verifiable with probability $p \in (0, 1)$, or only within an interval. Then "no false positives" cannot be interpreted pointwise in the feature space; it must be interpreted as a joint property of the decision rule and the verification process. Mechanism design with endogenous verification suggests that optimal policy may trade off a stricter acceptance threshold against more aggressive auditing, e.g.,

$$\text{accept if } \beta^\top \hat{x}^C \geq \tau \text{ and pass audit},$$

where the audit has power that depends on resources and on the reported feature. The recourse problem becomes richer: providing a pathway may simultaneously increase true qualification *and* increase verifiability by steering agents toward actions that are easiest to attest (completing a certified program rather than claiming experience). This points to a normative tension: "verifiable" often means "institutionally legible," which can exclude informal but genuine routes to qualification.

A second open problem is to incorporate *third-party and platform intermediaries*. Many agents interact with decision systems through brokers (test-prep companies, credential vendors, credit-repair firms). These intermediaries can reduce real costs of causal investment (good) but can also supply proxy manipulation technologies (bad). A realistic equilibrium analysis should treat the manipulation cost c_P as endogenous to the surrounding market, potentially responding to the mechanism. The safe-sufficiency result then looks like a commitment device not just against agents, but against an entire manipulation industry: by making proxies payoff-irrelevant, the mechanism shrinks the demand for proxy gaming tools.

2. Measurement error in causal features: robust safety versus access. Even when we restrict to causal features, the principal typically observes them with noise. Let $\tilde{x}^C$ denote the measurement used for decision-making, with

$$\tilde{x}^C = x^C + e + \eta, \qquad \eta \sim \mathcal{D} \text{ (possibly heteroskedastic).}$$

If we maintain the causal notion of true qualification,

$$y = \mathbf{1}\{\beta^\top (x^C + e) \geq \tau\},$$

then an acceptance rule of the form $\mathbf{1}\{\beta^\top \tilde{x}^C \geq \tau\}$ generally violates no-FP because positive noise η can push an unqualified agent across the observed threshold. Enforcing safety now requires accepting only when qualification is implied *for all plausible realizations* of η (worst-case) or with high probability (risk-limiting). A canonical robust rule is a "margin" requirement:

$$f(\tilde{x}) = \mathbf{1}\{\beta^\top \tilde{x}^C \geq \tau + \kappa\},$$

where κ is chosen so that $\Pr(\beta^\top \eta \geq \kappa) \leq \varepsilon$; then $\Pr(y = 0 \mid f(\tilde{x}) = 1) \leq \varepsilon$ under distributional assumptions. This immediately raises a design question: *who pays for robustness*? Increasing κ preserves safety but expands false negatives and increases required investment. Recourse can partially mitigate this by explicitly telling agents the *measured* target they must exceed, but that can be perverse if agents then invest to "beat the test" rather than to become truly qualified. In domains like education and healthcare, this is not hypothetical: systems that couple eligibility to noisy measurements induce Goodhart effects even when the measured quantity is causally relevant.

An open problem we find especially important is *joint design of measurement and recourse*. If the institution can choose a more precise test (reducing the variance of η) at some cost, then the optimal policy may be to invest in measurement rather than to tighten thresholds. This is a concrete way to connect our framework to practice: many fairness and accountability debates revolve around whether to improve measurement (better data, better audits) or to adjust decision rules. In our language, improving measurement can expand the set of agents for whom a safe acceptance rule is not overly conservative, increasing true positives without relaxing safety.

3. Dynamic and multi-stage investment: recourse as a contract, not a message. Our equilibrium analysis is essentially static: agents choose e once, acceptance is decided once, and recourse is informational. Yet in most applications, investment is dynamic. Training takes months, credit-building takes repeated payments, health improvement takes sustained adherence. A natural extension is a two-stage (or infinite-horizon) model where agents can invest over time and reapply, with discounting and possible depreciation:

$$x_{t+1}^C = x_t^C + e_t - \delta x_t^C, \qquad U_A = \sum_{t \geq 0} \rho^t \big(\gamma a_t - c_C(e_t; \theta)\big).$$

In such environments, a recourse policy is closer to a *relational contract*: it is a commitment about future acceptance conditional on future verifiable states. This raises time-consistency issues. If the principal can change τ or β after agents invest, the promise embedded in recourse is not credible, and rational agents underinvest. Thus, the central implementation problem may be less about computing $e^\dagger$ and more about institutional commitment (regulation, published standards, appeal rights).

Dynamics also introduce the possibility of *staged recourse menus*. Instead of recommending the full action to cross the threshold, an institution may recommend incremental steps (e.g., "raise $s(x^C)$ by Δ each quarter"), partly to reduce abandonment and partly to allow learning about agent costs and constraints. Designing such staged pathways resembles optimal stopping and screening: we would like menus that are simple, verifiable, and incentive-compatible over time, while avoiding lock-in to inefficient routes.

A further open problem is financing. When investment costs are upfront but benefits are delayed, many socially desirable improvements fail privately because γ is effectively discounted or liquidity-constrained. Mechanisms that incorporate loans, income-share agreements, or conditional subsidies can be viewed as altering the agent's effective γ and w, but they also create new moral hazard and default considerations. Integrating recourse with financing is, in our view, essential for policy relevance in education and workforce settings.

4. Group constraints and distributive objectives: safety is not fairness. Our baseline objective maximizes true positives subject to safety. This is a plausible institutional stance in high-stakes approval contexts, but it is not a complete welfare criterion. In particular, the distribution of x^C and θ may differ across protected groups, so that the set of agents satisfying $\gamma \geq C^\star(x^C, \theta)$ varies sharply by group. Even if the mechanism is "neutral" in the sense of using only causal score, it can still generate disparate acceptance and disparate investment burdens.

There are (at least) two ways group constraints enter. First, the principal may face *acceptance-rate constraints* (e.g., demographic parity) or *error constraints* (equalized odds) across groups. Under our safety regime, false positives are (approximately) eliminated, so the binding fairness concern is typically about false negatives and access. Imposing group parity in acceptance while maintaining safety may force group-specific thresholds τ_g or group-specific recourse support (subsidies, expanded menus) that equalize the cost of reaching the standard rather than equalizing outcomes mechanically. This connects to a policy distinction: *equality of opportunity through support* versus *equality of outcomes through rule changes*. Our framework naturally emphasizes the former because it preserves the meaning of qualification while addressing heterogeneous costs.

Second, group membership may be correlated with proxy manipulability (different access to coaching, documentation services, or social capital). One practical benefit of safe-sufficient mechanisms is that they reduce returns to such proxy advantages. But if causal investments themselves are unequally accessible, then strict safety can inadvertently entrench inequality. A principled extension would treat the principal's objective as including welfare or burden terms, e.g.,

$$\max_M \ \mathbb{E}[\mathbf{1}\{f(\hat{x}) = 1\}y] - \lambda\, \mathbb{E}[c_C(e^\star; \theta)] \quad \text{and/or group constraints.}$$

The open question is how to design recourse that is both incentive-compatible and burden-aware. For example, if we allow transfers $T(e)$ targeted to actions in the recourse menu, can we guarantee safety while achieving approximate group parity at minimal subsidy cost? This starts to resemble optimal policy design with endogenous effort, and it invites empirical work on which investments are most elastic to support.

5. Practical deployment: from a theorem to an accountable system. Finally, even if one accepts the model's normative stance, implementation requires decisions that are often omitted from formal analyses.

Communication and contestability. Recourse is only valuable if agents understand it and trust it. Communicating a linear score threshold is easier than communicating a complex classifier, but institutions still face choices

about how much to reveal: revealing β can enable gaming of borderline verifications, while hiding it can undermine legitimacy. A promising direction is to publish *actionable* pathways (the recommended $e^\dagger$) without fully revealing the scoring weights, paired with auditable guarantees that completing the pathway implies acceptance.

Heterogeneity beyond single-crossing. Our constructive recourse uses a single cheapest coordinate direction $j^\dagger$. In practice, the cheapest improvement differs across agents (time constraints, geography, disability, labor market conditions). The menu-complexity extension hints at this, but a deployment-oriented theory should address how menus are chosen under limited administrative capacity and how to prevent menus from becoming de facto exclusionary (e.g., only one credential vendor is recognized). This is both a computational and a governance question.

Behavioral frictions. Real agents are not perfect optimizers. Take-up may be low even when $C^\star \leq \gamma$ because of present bias, misinformation, or distrust. In our notation, this is a wedge between modeled θ and realized behavior. Incorporating bounded rationality suggests that recourse should include not only a target action but also supports (reminders, coaching) that effectively reduce non-monetary components of θ. The theoretical challenge is to do so without reintroducing manipulable proxies or discretionary favoritism.

Monitoring and post-acceptance incentives. Many domains care about sustained performance, not one-time qualification. If acceptance changes incentives (e.g., once hired, effort declines), then the principal may want post-acceptance monitoring that depends on outcomes rather than features. Designing such monitoring while preserving the spirit of safe recourse—rewarding real performance rather than cosmetic proxies—is an open avenue.

Summary. We view the benchmark mechanism as a disciplined starting point: it clarifies that, under verifiability and deterministic causal standards, proxy dependence is a design error when proxies are manipulable, and that recourse can be made both safe and actionable. The most urgent open problems are therefore not about deriving yet another optimal classifier, but about institutional complements: verifiable measurement, robust decision rules under noise, credible dynamic commitments, distributive supports under heterogeneous costs, and deployment practices that make pathways legible and contestable. Progress on these fronts would move recourse from an explanatory artifact to a regulated interface between individuals and high-stakes institutions.