The Price of Justice in Machine Learning: Fair Division with Subjective Value under Bounded Rationality

Anonymous authors
Paper under double-blind review

Abstract

Statistical fairness criteria such as Demographic Parity, Equalized Odds, and Calibration are widely used in machine learning, but they constrain rates and may fail to capture how burdens are distributed across individuals. We introduce a harm-allocation view of fairness that models heterogeneous subjective error costs and evaluates decisions through fair-division axioms. Within this framework, we characterize when common statistical criteria can serve as valid surrogates for harm-based fairness, construct counterexamples showing severe failures under cost heterogeneity, and reinterpret classic incompatibilities as conflicts between allocation principles. We further establish approximation lower bounds under realistic constraints, including noisy or coarse cost information and restricted policy classes. Experiments support the theory by showing when surrogate metrics align with or diverge from harm-based fairness, revealing approximation floors and showing that these misalignments persist across tasks, models, and diverse fairness interventions under finer-grained harm diagnostics.

1 Introduction

Machine-learning (ML) systems increasingly mediate access to consequential opportunities, including credit, employment, insurance, education, and public services. As a result, a large body of work has proposed *statistical* fairness criteria such as Demographic Parity (DP), Equalized Odds (EO), and Calibration, and has developed algorithms that enforce these constraints during training or through post-processing. These tools are valuable for auditing and for mitigating certain forms of disparate impact. Yet, in both research and practice, fairness is often treated as synonymous with *matching a metric*, even when it remains unclear what, exactly, that metric is intended to proxy.

This paper argues that the resulting gap is not merely empirical but conceptual. In many deployment settings, a decision rule allocates not only *goods* (approvals, access, and opportunities) but also *bads* (false positives and false negatives, additional scrutiny, delay, stigma, and opportunity costs). From a normative perspective, what matters is how these burdens and benefits are distributed across individuals. Crucially, the burden associated with the same statistical event may differ substantially from one person to another. A false negative may be catastrophic for one applicant but relatively minor for another, while a false positive may generate different downstream risks depending on context. This motivates a *subjective theory of value*: individuals attach heterogeneous, and often unobserved, values and costs to outcomes, so identical confusion-matrix statistics can correspond to very different realized harms.

Once subjective valuation is acknowledged, the appropriate normative language is not simply a collection of rate constraints, but the language of *fair division*. Rather than treating DP, EO, and calibration as fairness itself, we treat a policy as inducing a vector of individual expected harms and evaluate that harm allocation using standard fair-division axioms, such as envy-freeness and proportionality. From this perspective, DP, EO, and calibration become *candidate surrogates*: they are normatively meaningful only to the extent that they preserve the relevant axioms on harm allocations.

A second ingredient of our framework is that fairness in practice is rarely implemented by an omniscient social planner. Instead, it is mediated by an institution or platform operating under *bounded rationality*:

it does not observe individuals' subjective costs directly, it has access only to noisy or coarse estimates, and it is limited to restricted policy classes (e.g., threshold rules, group-specific thresholds, or other low-complexity post-processing rules). Under bounded rationality, the central question is not whether an axiom can be satisfied exactly, but what *approximation floor* remains once informational and policy constraints are imposed. This paper makes that floor explicit and shows that it is generically nonzero.

Our framework provides a unified account of why metric-centric fairness can fail. Statistical constraints operate in a low-dimensional space of rates, whereas harm-based axioms operate in a high-dimensional space shaped by heterogeneous valuations. Bridging the gap between these spaces requires strong implicit assumptions. When those assumptions do not hold, surrogate constraints may be satisfied even while harm-based axioms are substantially violated. We formalize these claims theoretically and examine them empirically through theorem-driven experiments that trace surrogate–axiom alignment under increasing heterogeneity and under informational and policy constraints.

Contributions.

- **Fairness as harm allocation under a subjective theory of value.** We formalize individual expected harm as a function of heterogeneous, person-specific costs associated with different error types, and we view any decision policy as inducing a population-level vector of harms. This makes explicit what is being allocated when an ML system makes decisions.
- **Axiomatizing fairness through fair division.** We adopt fair-division axioms—including envy-freeness and proportionality-style protections—as the primary normative criteria for evaluating harm allocations. Statistical notions such as DP, EO, and Calibration are thereby repositioned as *surrogates*, rather than fairness itself.
- **Representation conditions for surrogate validity.** We show that, for a statistical fairness constraint to function as a valid surrogate for axiomatic harm fairness, strong implicit restrictions on subjective costs are generally required. This identifies the hidden assumptions under which metric-centric fairness can be given a normative interpretation.
- **Counterexamples demonstrating structural surrogate failure.** We construct settings with heterogeneous subjective costs in which a policy satisfies standard statistical criteria while violating harm-based fair-division axioms by a large margin. These examples show that surrogate failure is not a pathological corner case, but a structural possibility once subjective values differ.
- **Reinterpreting classic incompatibilities as conflicts among allocation principles.** We connect well-known impossibility and incompatibility results in ML fairness to conflicts among fair-division axioms defined over harm allocations, thereby clarifying the normative content of these trade-offs beyond rate matching.
- **Bounded rationality and lower bounds on approximate fairness.** We model an intermediary that observes only noisy or coarse proxies of subjective costs and is restricted to simple policy classes. Under these constraints, we show that the best achievable violation level of harm-based axioms generally admits a nontrivial lower bound, reframing the problem as one of minimizing an irreducible approximation floor.
- **Theorem-driven empirical analysis.** We empirically characterize when surrogate metrics align with or diverge from harm-based axioms as value heterogeneity increases, demonstrate approximation floors induced by informational and policy constraints, construct "axiomatic profiles" of DP, EO, and calibration in harm space, and show that the structural misalignment persists across a diverse set of fairness interventions.

Overall, this work does not propose another fairness algorithm. Instead, it provides a principled normative and operational framework—grounded in a subjective theory of value, fair division, and bounded rationality—for understanding what statistical fairness criteria can and cannot guarantee, and for diagnosing the irreducible limitations that arise when fairness is interpreted as harm allocation under heterogeneous values and mediated constraints.

2 Related Work

2.1 Statistical Fairness as Surrogate Constraints

A dominant tradition in fair machine learning defines fairness through constraints on observable prediction statistics, rather than through a direct accounting of the downstream harms and benefits induced by decisions. Early work established the basic vocabulary of this paradigm by distinguishing individual and group notions of fairness and by operationalizing parity-style criteria in statistical terms (Dwork et al., 2012; Feldman et al., 2015; Zemel et al., 2013; Madras et al., 2018a; Zhao et al., 2019; Jang et al., 2024). In this line of work, the primary technical object is a measurable relation over predictions, labels, scores, and sensitive attributes.

Within this tradition, one major family of criteria requires parity in selection rates or error rates across protected groups. Demographic parity, equal opportunity, equalized odds, and related misclassification-based notions were formalized and optimized in a large body of work (Hardt et al., 2016; Zafar et al., 2017b;a; Agarwal et al., 2018; Celis et al., 2019). A parallel line studied calibration-based criteria and the tensions among these statistical requirements, showing that prominent fairness metrics can be mutually incompatible except under special conditions (Kleinberg et al., 2016; Pleiss et al., 2017; Canetti et al., 2019; Hébert-Johnson et al., 2018; Hsu et al., 2022). Subsequent work also emphasized that coarse group parity can mask substantial disparities on structured subpopulations, motivating subgroup-level auditing and richer notions of statistical coverage (Kearns et al., 2018; Speicher et al., 2018; Hashimoto et al., 2018).

A further strand of the literature turned these statistical criteria into trainable or certifiable constraints. This includes reductions-based methods, constrained empirical risk minimization, generalization analyses for fairness constraints, extensions to regression, and settings with limited or noisy access to protected attributes (Cotter et al., 2019; Agarwal et al., 2019; Mozannar et al., 2020; Celis et al., 2021; Mukherjee et al., 2020; Wu et al., 2022b; Bendekgey & Sudderth, 2021; Taufiq et al., 2024). Recent surveys continue to organize much of the field around the design, optimization, and comparison of such measurable statistical criteria (Caton & Haas, 2024; Tang et al., 2023).

From the standpoint of the present paper, this literature is best understood as developing *surrogate constraints*: criteria defined over observables that can be optimized, audited, and compared at scale, but that do not by themselves specify what burdens are being allocated, how those burdens are valued by different individuals, or when parity in metric space constitutes a valid proxy for fairness in harm space.

2.2 Welfare, Utility, and Preference-Aware Fairness

A complementary line of work argues that fairness should be evaluated not only through parity in observable rates, but through the utilities, burdens, preferences, and welfare consequences induced by decisions. In this view, the central question is not merely whether a classifier equalizes a statistic, but how the resulting policy changes the distribution of benefits, risks, and opportunities across individuals and groups (Heidari et al., 2018; Cousins, 2021; Hu & Chen, 2020; Kim et al., 2019; Milli et al., 2019; Williamson & Menon, 2019; Li & Liu, 2022; Liang & Lu, 2024; Pardeshi et al., 2024).

This perspective extends naturally to online, sequential, and allocation settings, where fairness is often expressed through welfare aggregation over rewards, regret, or feasible assignments rather than through static parity constraints alone (Hossain et al., 2021; Barman et al., 2023; Sawarni et al., 2023; Zhang et al., 2024; Siddique et al., 2020; Sim et al., 2021; Devic et al., 2023; Lobo et al., 2024).

Utility- and preference-aware formulations are also central in ranking, recommendation, and multi-sided platforms, where algorithmic decisions allocate exposure, attention, and market opportunity across multiple stakeholders with potentially conflicting objectives (Ge et al., 2022; Singh & Joachims, 2018; Wang et al., 2021; Do et al., 2021; Ge et al., 2021; Wu et al., 2022a; Greenwood et al., 2024; Chen et al., 2024). Taken together, this literature moves substantially closer to the normative object of interest than metric-only approaches do: it asks who benefits, who pays, whose preferences are respected, and which welfare aggregator should guide trade-offs. Our paper is closest in spirit to this tradition, but differs in a crucial respect. We do not propose another welfare-aware objective; instead, we study when familiar statistical fairness metrics

can or cannot serve as valid surrogates for axiomatic fairness over harm allocations under heterogeneous subjective costs.

2.3 Axiomatic Fairness and Fair Division

A different tradition studies fairness not as parity in observable prediction statistics, nor primarily as welfare optimization, but as a property of the *allocation itself*. In this literature, fairness is specified through axioms such as envy-freeness, proportionality, maximin-share guarantees, competitive-equilibrium-based notions, and ordinal assignment principles, and the central task is to characterize when such guarantees exist, how they relate to one another, and how they interact with efficiency. This perspective is especially salient for our purposes because it treats fairness as an object-level claim about who receives what, under which comparison benchmark, and according to which normative principle (Amanatidis et al., 2022; 2023; Aziz et al., 2022b; Lipton et al., 2004; Bouveret et al., 2005; Bouveret & Lemaître, 2016; Caragiannis et al., 2023; Budish, 2011; Bogomolnaia & Moulin, 2001; Segal-Halevi, 2020).

Within algorithmic fair division, a major line of work investigates the existence and computation of approximate fairness guarantees for indivisible goods. These include relaxations of envy-freeness such as EF1 and EFX, maximin-share guarantees and their approximations, Nash-social-welfare-based rules, and fairness under asymmetric entitlements or combinatorial constraints. The resulting literature has shown that approximate fairness is often achievable even when exact envy-freeness or exact proportionality is impossible, and that some allocation rules exhibit striking simultaneous fairness–efficiency properties (Caragiannis et al., 2019b; Cole & Gkatzelis, 2015; Caragiannis et al., 2019a; Kurokawa et al., 2016; 2018; Garg & Taki, 2020; Farhadi et al., 2019; Aziz et al., 2022a; Barman & Biswas, 2018; Li & Vetta, 2021; Greco & Scarcello, 2020; Baklanov et al., 2021; Oh et al., 2021).

The same axiomatic viewpoint has also been extended beyond the classical private-goods setting to mixed divisible–indivisible resources, public decisions, public goods, and constrained random assignment. These extensions are important because they show that fair-division axioms are not tied to a single mathematical environment, but provide a general language for reasoning about claims, entitlements, and permissible complaints across diverse allocation problems (Bei et al., 2021b;a; Babichenko et al., 2024; Conitzer et al., 2017; Fain et al., 2018; Garg et al., 2021; Kawase et al., 2022). Our paper is closest in spirit to this literature. However, rather than directly designing an allocation rule that satisfies a chosen fair-division axiom, we use axiomatic fairness as a normative benchmark for evaluating whether familiar statistical fairness metrics can be interpreted as valid surrogates for fairness in the allocation of harms.

2.4 Fairness under Mediation, Uncertainty, and Bounded Rationality

A further line of work studies fairness when decisions are made under imperfect information, mediated through institutional pipelines, or shaped by downstream decision-makers whose knowledge and incentives differ from those of the model designer. In these settings, fairness cannot be analyzed solely from a static mapping from features to predictions, because labels may be selectively observed, target variables may be imperfect proxies, sensitive attributes may be noisy, and final outcomes may depend on how models interact with human experts or organizational procedures (Kilbertus et al., 2020; Wei, 2021; Guerdan et al., 2023; Chouldechova et al., 2018; Robertson et al., 2021; Madras et al., 2018b; Mozannar & Sontag, 2020; Charusaie et al., 2022; Mozannar et al., 2023; Ghoummaid & Shalit, 2024; Mehrotra & Vishnoi, 2022).

Related work also emphasizes that fairness may be unstable over time when decisions reshape the population, the data distribution, or the institutional environment in which future decisions are made. This includes work on delayed and long-run effects of fair decision rules, sequential user dynamics, performative prediction, and dynamic fairness constraints in learning and control (Liu et al., 2018; Zhang et al., 2019; Hardt & Mendler-Dünner, 2025; Brown et al., 2022; Deng et al., 2022; Rateike et al., 2024; Somerstep et al., 2024). The common lesson of this literature is that fairness assessments based on a single static dataset may be fragile once deployment changes behavior, representation, or future observability.

A closely related body of work studies strategic response and recourse, where affected individuals adapt to the rule, attempt to manipulate observable features, or face unequal abilities to comply with model-

induced demands. Here, fairness depends not only on the deployed rule itself, but also on the response costs, competitive environment, and actionability of the interventions the rule induces (Dong et al., 2018; Sundaram et al., 2023; Liu et al., 2022; Estornell et al., 2023; Von Kügelgen et al., 2022; Fokkema et al., 2024; Perello et al., 2025). Our paper is closest to this broader tradition of mediated and information-constrained decision-making. However, rather than focusing on a particular dynamic protocol, deferral architecture, or strategic-response model, we isolate a more general point: even when one fixes a harm-based normative benchmark, limited information and restricted policy classes can prevent a mediator from implementing or even reliably approximating the corresponding fair allocation.

3 Preliminaries: Decision Rules and Harm Allocation Setup

We consider a binary decision problem over a population of individuals indexed by $i \in \{1, \ldots, n\}$. Each individual is associated with features $X \in \mathcal{X}$, a (possibly sensitive) group attribute $A \in \mathcal{A}$, and a ground-truth label $Y \in \{0, 1\}$. We write an observed instance as (x_i, a_i, y_i), and use uppercase letters (X, A, Y) to denote random variables drawn from an underlying distribution.

Scores and decision rules. A (possibly learned) scoring model produces a score $s : \mathcal{X} \to [0, 1]$, where larger values indicate a greater predicted likelihood that $Y = 1$. A *decision rule* (or *policy*) π maps the available information to an action $\hat{Y} \in \{0, 1\}$, where $\hat{Y} = 1$ denotes a positive decision (e.g., approve or accept) and $\hat{Y} = 0$ denotes a negative decision (e.g., deny or reject). To accommodate randomized post-processing, which is common in fairness interventions, we model π as a Markov kernel:

$$\pi : \mathcal{X} \times \mathcal{A} \to [0, 1], \qquad \pi(x, a) \;=\; \Pr_{\pi}(\hat{Y} = 1 \mid X = x, A = a).$$

Deterministic rules arise as the special case in which $\pi(x, a) \in \{0, 1\}$ for all (x, a), for example under threshold-based decision rules.

Errors as allocative bads. We focus on the *harms* (bads) generated by decisions. In the basic binary setting, the primary harms arise from classification errors:

$$\mathrm{FP} := \{\hat{Y} = 1,\, Y = 0\}, \qquad \mathrm{FN} := \{\hat{Y} = 0,\, Y = 1\}.$$

When π is randomized, each individual instance (x_i, a_i, y_i) induces a distribution over $\hat{Y}$. Fix an instance i and treat y_i as part of the realized context. The relevant per-instance error probabilities are then

$$\Pr_{\pi}(\mathrm{FP} \mid i) = \Pr_{\pi}(\hat{Y} = 1 \mid x_i, a_i, y_i)\, \mathbb{1}\{y_i = 0\} = \mathbb{1}\{y_i = 0\}\, \pi(x_i, a_i),$$

$$\Pr_{\pi}(\mathrm{FN} \mid i) = \Pr_{\pi}(\hat{Y} = 0 \mid x_i, a_i, y_i)\, \mathbb{1}\{y_i = 1\} = \mathbb{1}\{y_i = 1\}\, (1 - \pi(x_i, a_i)).$$

These per-instance error probabilities serve as the basic primitives for our harm accounting.

Population- and group-level rates. Let $\mathcal{D}$ denote the population distribution over (X, A, Y). For each group value $a \in \mathcal{A}$, we use the standard conditional error rates under π:

$$\mathrm{FPR}_a(\pi) := \Pr_{\mathcal{D},\pi}(\hat{Y} = 1 \mid Y = 0, A = a), \qquad \mathrm{FNR}_a(\pi) := \Pr_{\mathcal{D},\pi}(\hat{Y} = 0 \mid Y = 1, A = a).$$

We also use the corresponding true rates $\mathrm{TPR}_a(\pi) = 1 - \mathrm{FNR}_a(\pi)$ and $\mathrm{TNR}_a(\pi) = 1 - \mathrm{FPR}_a(\pi)$ when convenient. Expectations and probabilities subscripted by $(\mathcal{D}, \pi)$ indicate that randomness may arise from both the data-generating process $\mathcal{D}$ and any internal randomization in π.

What is being allocated. A policy π allocates (i) *goods*, such as the positive decision $\hat{Y} = 1$ (access, approval, or opportunity), and (ii) *bads*, such as error events (FP/FN) and their downstream consequences. Our central object of study is the distribution of *harm* induced by π across individuals.

4 Subjective Value: Individualized Error Costs and Harm Vectors

A central departure from standard metric-first approaches to fairness is that *the value of an outcome is not universal*. Different individuals may reasonably assign different costs to the same type of error. We therefore model fairness at the level of *harm allocation*, rather than only at the level of group-aggregated error rates.

Individualized error costs. Each individual i is associated with a pair of nonnegative error costs

$$c_i^{\mathrm{FP}} \geq 0, \qquad c_i^{\mathrm{FN}} \geq 0,$$

interpreted as the subjective disutility associated with a false positive and a false negative, respectively. These costs may reflect heterogeneous downstream consequences, such as financial loss, opportunity cost, stigma, time burden, or safety risk, and they need not be constant across individuals even within the same demographic group.

Individual expected harm under a policy. Fix a decision rule π. Using the individual-level error probabilities introduced in Section 3, we define the expected harm borne by individual i as

$$H_i(\pi) \;:=\; c_i^{\mathrm{FP}} \Pr_{\pi}(\mathrm{FP} \mid i) \;+\; c_i^{\mathrm{FN}} \Pr_{\pi}(\mathrm{FN} \mid i). \tag{4.1}$$

In words, harm is defined as a cost-weighted combination of the two basic error events under π. This formulation naturally accommodates randomized policies and allows policies to be compared in terms of how they distribute expected harm across individuals.

Harm-allocation vector. The *harm allocation* induced by policy π is the vector of individual harms

$$H(\pi) \;:=\; \big(H_1(\pi), \ldots, H_n(\pi)\big) \in \mathbb{R}_+^n. \tag{4.2}$$

Our fairness objectives will be formulated as axioms and constraints on $H(\pi)$, rather than solely as constraints on group-level confusion-matrix statistics.

Heterogeneity as a source of misalignment. The key difficulty is that statistical fairness constraints often implicitly treat the "unit" of harm as comparable across individuals, whereas subjective costs generally undermine that comparability. We highlight two common forms of heterogeneity:

- **Between-group heterogeneity:** $\{c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}}\}$ differ systematically across demographic groups, for example because of differential exposure to downstream sanctions or different opportunity costs.
- **Within-group heterogeneity:** even holding $A = a$ fixed, individuals may vary substantially in c_i^{FP} and c_i^{FN}, for example because of differences in risk tolerance, economic vulnerability, or health constraints, so group-average costs can be a poor proxy for individual burden.

Under either type of heterogeneity, equalizing error rates or selection rates across groups may fail to equalize, or even meaningfully constrain, the distribution of harm across individuals.

Aggregate harm summaries (for efficiency comparisons). For later comparisons involving efficiency and welfare trade-offs, we will occasionally use scalar summaries of $H(\pi)$, such as mean harm and tail or worst-case harm:

$$\overline{H}(\pi) := \frac{1}{n}\sum_{i=1}^{n} H_i(\pi), \qquad H^{(q)}(\pi) := \mathrm{Quantile}_q\big(\{H_i(\pi)\}_{i=1}^n\big),$$

where $H^{(q)}(\pi)$ denotes the q-quantile (e.g., $q = 0.9$ for the "worst 10%" of harms). These summaries are not fairness notions in themselves; rather, they serve as complementary efficiency signals when we study the cost of enforcing fairness axioms.

5 Axiomatic Fair Division of Harm

This section treats fairness as an axiomatic property of the *harm-allocation vector* $H(\pi)$, rather than as a constraint on any particular statistical rate. We adapt standard fair-division axioms to a setting in which each individual is characterized by subjective error costs.

Why envy and proportionality (two complementary protections). We focus on two canonical axioms that capture complementary ways in which a harm allocation may be objectionable. *Envy-freeness* is *relational*: it detects comparative grievances by asking whether an individual would prefer another person's error lottery when evaluated using their own subjective costs. *Proportionality* is *protective*: it provides an individualized upper-bound guarantee against highly concentrated harm by comparing each person's realized burden with their own valuation of the total harm. Taken together, EF and PROP distinguish two different failure modes—"someone has reason to complain by comparison" versus "someone bears an outsized share of the burden"—a distinction that statistical rate constraints can obscure under heterogeneous valuations. We do not claim that these axioms exhaust the normative space; rather, they serve as a minimal and standard pair of fair-division-style protections for *bads* that can be measured from error events.

Bundles of harm and subjective comparison. A policy π induces, for each individual j, an *error-probability vector*

$$e_j(\pi) := \big(\Pr_\pi(\mathrm{FP} \mid j),\ \Pr_\pi(\mathrm{FN} \mid j)\big),$$

where the realized label y_j determines which component can be nonzero. To express interpersonal comparisons under *subjective* costs, we define how individual i evaluates the bundle induced for individual j:

$$H_{i\leftarrow j}(\pi) \ := \ c_i^{\mathrm{FP}} \Pr_\pi(\mathrm{FP} \mid j) \ + \ c_i^{\mathrm{FN}} \Pr_\pi(\mathrm{FN} \mid j). \tag{5.1}$$

By definition, $H_j(\pi) = H_{j\leftarrow j}(\pi)$. The quantity $H_{i\leftarrow j}(\pi)$ serves as a comparison device: it measures j's error profile in i's cost units.

Envy-freeness (harm version). Envy-freeness requires that no individual would strictly prefer another person's induced harm bundle when evaluated using their own costs.

Definition 5.1 (Envy-freeness)**.** A policy π is *envy-free* (EF) if, for all individuals i, j,

$$H_{i\leftarrow i}(\pi) \ \leq \ H_{i\leftarrow j}(\pi).$$

We also use the following relaxed notion.

Definition 5.2 (δ-envy-freeness)**.** For $\delta \geq 0$, a policy π is *δ-envy-free* if, for all i, j,

$$H_{i\leftarrow i}(\pi) \ \leq \ H_{i\leftarrow j}(\pi) + \delta.$$

Proportionality / upper-bound guarantees (harm version). In fair division, proportionality compares an individual's bundle with a $1/n$-share of the total, as valued by that individual. Here, individual i's valuation of the entire harm allocation is $\sum_{j=1}^{n} H_{i\leftarrow j}(\pi)$.

Definition 5.3 (Proportionality)**.** A policy π is *proportional* (PROP) if, for all i,

$$H_{i\leftarrow i}(\pi) \ \leq \ \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi).$$

Definition 5.4 (δ-proportionality)**.** For $\delta \geq 0$, a policy π is *δ-proportional* if, for all i,

$$H_{i\leftarrow i}(\pi) \ \leq \ \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi) + \delta.$$

Interpretation. This condition functions as an individualized upper-bound guarantee: no individual bears more than a $1/n$-share of the total harm, measured in their own cost units. For *bads*, such a guarantee provides a conservative safeguard against scapegoating and extreme burden concentration: even when envy is low, and thus comparative grievances are muted, a policy may still be normatively objectionable if it concentrates severe harm on a minority. We therefore use PROP primarily as a *cap-style protection* complementary to envy-freeness, rather than as a claim that proportionality is the unique or complete normative principle for allocating harm.[1]

Violation measures for theory and experiments. To quantify how far a policy is from satisfying these axioms, we use nonnegative violation functionals. For envy-freeness, define pairwise envy and its aggregates:

$$\begin{aligned}\mathrm{EnvyGap}_{i,j}(\pi) &:= \big[H_{i\leftarrow i}(\pi) - H_{i\leftarrow j}(\pi)\big]_+,\\ \mathrm{Envy}^{\max}(\pi) &:= \max_{i,j} \mathrm{EnvyGap}_{i,j}(\pi),\\ \mathrm{Envy}^{\mathrm{avg}}(\pi) &:= \frac{1}{n(n-1)} \sum_{i\neq j} \mathrm{EnvyGap}_{i,j}(\pi).\end{aligned}$$

We also define the fraction of violating pairs:

$$\mathrm{EnvyFrac}(\pi) := \frac{1}{n(n-1)} \sum_{i\neq j} \mathbb{1}\{\mathrm{EnvyGap}_{i,j}(\pi) > 0\}.$$

For proportionality, define individualized excess harm:

$$\begin{aligned}\mathrm{PropExcess}_i(\pi) &:= \Big[H_{i\leftarrow i}(\pi) - \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi)\Big]_+,\\ \mathrm{Prop}^{\max}(\pi) &:= \max_i \mathrm{PropExcess}_i(\pi),\\ \mathrm{Prop}^{\mathrm{avg}}(\pi) &:= \frac{1}{n}\sum_{i=1}^{n} \mathrm{PropExcess}_i(\pi).\end{aligned}$$

Finally, to track harm concentration—which is useful in practice even when it is not itself an axiom—we report tail summaries such as

$$H^{(q)}(\pi) := \mathrm{Quantile}_q\big(\{H_i(\pi)\}_{i=1}^n\big), \qquad H^{\max}(\pi) := \max_i H_i(\pi).$$

Cost-stratum envy proxy used in experiments. Computing $\mathrm{Envy}^{\max}(\pi)$ over all individual pairs is an $O(n^2)$ operation, so in the experiments we use a scalable coarse-grained proxy that aggregates individuals into *cost strata*.

Let $G(i) \in \{1, \ldots, K\}$ denote a cost-stratum index defined by quantiles of $\log(c_i^{FP}/c_i^{FN})$, and let $S_g = \{i : G(i) = g\}$. Let $\bar{c}_g = (\bar{c}_g^{FP}, \bar{c}_g^{FN})$ denote the mean costs in stratum g. On a split $\mathcal{D}$, define the stratum-level empirical error rates as

$$\widehat{\Pr}(FP \mid g) = \frac{1}{|S_g|} \sum_{i\in S_g} \mathbf{1}\{y_i = 0, \hat{y}_i = 1\}, \quad \widehat{\Pr}(FN \mid g) = \frac{1}{|S_g|} \sum_{i\in S_g} \mathbf{1}\{y_i = 1, \hat{y}_i = 0\}.$$

We then define the cross-evaluated *stratum harm*

$$\widetilde{H}_{g\leftarrow g'}(\pi) := \bar{c}_g^{FP}\, \widehat{\Pr}(FP \mid g') + \bar{c}_g^{FN}\, \widehat{\Pr}(FN \mid g'),$$

[1] In the fair-division literature, proportionality is one of several standard protections, alongside envy-freeness, maximin-style guarantees, and share-based guarantees. Our results concern surrogate failure under heterogeneous valuations; the role of PROP here is to provide a simple, measurable bound against concentrated harm that is not captured by relational envy alone.

and the corresponding stratum-level envy gap and maximum aggregate

$$\widetilde{\mathrm{EnvyGap}}_{g,g'}(\pi) := \left[\widetilde{H}_{g\leftarrow g}(\pi) - \widetilde{H}_{g\leftarrow g'}(\pi)\right]_+, \qquad \widetilde{\mathrm{Envy}}^{\max}(\pi) := \max_{g,g'} \widetilde{\mathrm{EnvyGap}}_{g,g'}(\pi).$$

This proxy measures comparative grievances *between cost strata* induced by π; it does not enumerate individual-pair envy within the same stratum. Consequently, if $\log(c^{FP}/c^{FN})$ is nearly constant and the stratum partition effectively collapses, $\widetilde{\mathrm{Envy}}^{\max}(\pi)$ becomes mechanically small. This should therefore be interpreted as indicating "no detectable *cross-stratum* envy under the chosen partition," rather than as establishing individual-level envy-freeness. Unless stated otherwise, the term "envy" in the experiments refers to $\widetilde{\mathrm{Envy}}^{\max}(\pi)$ with $K = 80$.

Price of fairness (efficiency loss). When comparing fair policies with an unconstrained benchmark, we use a simple price-of-fairness quantity. Let $\pi^\star$ denote a reference policy (e.g., a risk-minimizing policy within a chosen policy class), and let π^{fair} denote a policy satisfying a target axiom, or achieving the smallest attainable violation. We measure efficiency loss via mean harm:

$$\mathrm{PoF}(\pi^{\mathrm{fair}}; \pi^\star) := \overline{H}(\pi^{\mathrm{fair}}) - \overline{H}(\pi^\star), \qquad \overline{H}(\pi) := \frac{1}{n}\sum_{i=1}^{n} H_i(\pi),$$

or, when convenient, via the corresponding ratio $\overline{H}(\pi^{\mathrm{fair}})/\overline{H}(\pi^\star)$.

6 Bounded Rationality and Intermediation Constraints

Fairness constraints in deployed ML systems are typically implemented by an intermediary (e.g., a platform, agency, or regulator) that selects and enforces a policy π. Crucially, such an intermediary does not optimize under full information or unrestricted computational capacity. We model these frictions explicitly and treat fairness as a *robust* objective under uncertainty and restricted policy classes.

Intermediary choice problem. Let Π denote a set of feasible decision rules (policies). The intermediary selects a policy $\pi \in \Pi$ after observing the data and whatever information is available about individuals' subjective costs. The induced harm allocation is $H(\pi) \in \mathbb{R}^n_+$ (Section 4), and fairness is evaluated through the axiom-violation functionals introduced in Section 5 (e.g., $\mathrm{Envy}^{\max}(\pi)$ and $\mathrm{Prop}^{\max}(\pi)$).

Information constraints. The intermediary typically does not observe the true cost vectors $c_i = (c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}})$. Instead, it observes a proxy $\hat{c}_i$ subject to bounded error. We model this through an uncertainty set:

$$\mathcal{C}(\hat{c}, \epsilon) := \Big\{ c = \{c_i\}_{i=1}^n \ : \ \|c_i - \hat{c}_i\| \le \epsilon \text{ for all } i \Big\}, \tag{6.1}$$

where $\|\cdot\|$ denotes a chosen norm (e.g., ℓ_∞ or ℓ_2), and $\epsilon \ge 0$ quantifies the degree of informational imperfection.

Remark 6.1. Experiment 2 (Section 8.2) uses multiplicative noise of the form $\hat{c} = c\exp(z)$, that is, $|\log(\hat{c}/c)| \le \epsilon$. Equivalently, letting $d = \log c$ and $\hat{d} = \log \hat{c}$, this can be written as an additive bound $\|d - \hat{d}\|_\infty \le \epsilon$ in log-cost space.

A more severe restriction is *coarse stratification*: the intermediary observes only a stratum label $g(i) \in \{1, \dots, k\}$, possibly derived from limited surveys or coarse proxies, and maintains a stratum-level cost estimate $\hat{c}_g$. In this case, all individuals in the same stratum share the same observed proxy, and heterogeneity within a stratum remains unobserved:

$$\hat{c}_i = \hat{c}_{g(i)}, \qquad c_i \in \mathcal{C}_{g(i)} \ \text{ (unknown within-stratum variation).}$$

Both models formalize the same basic principle: the intermediary must choose a single policy using imperfect and aggregated information about subjective harms.

Computation and policy-class constraints. Even if the intermediary's objective were specified precisely, the set of implementable policies is often restricted. We model this by restricting π to a simple family $\Pi_{\text{simple}} \subseteq \Pi$, for example:

- **Single-threshold rules:** $\hat{Y} = \mathbb{1}\{s(X) \geq \tau\}$ for some τ.
- **Group-specific thresholds:** $\hat{Y} = \mathbb{1}\{s(X) \geq \tau_A\}$ with one threshold parameter per group.
- **Low-parameter post-processing:** randomized mappings $\pi(s, a)$ with only a small number of degrees of freedom.

Such restrictions capture limited computational resources, interpretability budgets, legal constraints, or operational requirements, such as the need to deploy a small set of auditable thresholds.

Robust/minimax achievable fairness. Under information uncertainty, fairness should be interpreted as a *robust guarantee*. Let $\text{Viol}(\pi; c)$ denote an axiom-violation measure when the true costs are c (e.g., $\text{Envy}^{\max}(\pi)$ or $\text{Prop}^{\max}(\pi)$, both of which depend on c through $H_{i \leftarrow j}(\pi)$). The intermediary's best achievable violation level under information uncertainty and policy constraints is

$$\text{Ach}(\hat{c}, \epsilon) \;:=\; \inf_{\pi \in \Pi_{\text{simple}}} \; \sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \; \text{Viol}(\pi; c). \tag{6.2}$$

This objective has a worst-case interpretation: given only $\hat{c}$ (and an error radius ϵ), the intermediary chooses a simple policy that minimizes the largest possible fairness violation over all true cost realizations compatible with the observed proxies.

Why approximate fairness and lower bounds are inevitable. The presence of both Π_{simple} and $\mathcal{C}(\hat{c}, \epsilon)$ implies that exact satisfaction of the axioms (i.e., zero violation) is generally unattainable:

- If $\epsilon > 0$, or if value information is available only through coarse cost stratification, then different realizations of c consistent with the observations can induce conflicting fairness requirements, so any fixed policy must violate some of them.
- If Π_{simple} is restricted, then even with perfect information the policy class may not be expressive enough to implement the allocation required to satisfy the axioms.

Accordingly, the relevant primitive is *approximate* fairness, that is, small but nonzero violation. The central theoretical question therefore becomes: how does $\text{Ach}(\hat{c}, \epsilon)$ scale with the severity of the informational and computational constraints? This perspective leads naturally to lower-bound results of the form $\text{Ach}(\hat{c}, \epsilon) \geq f(\epsilon)$ for some nontrivial function f, thereby formalizing the idea that bounded rationality imposes an irreducible floor on harm-based fairness violations.

7 Statistical Fairness Metrics as Surrogates: Representation, Counterexamples, and Lower Bounds

Statistical fairness metrics such as demographic parity, equalized odds, and calibration are typically formulated as constraints on *group-aggregated* quantities, such as rates or conditional distributions. By contrast, our framework defines fairness axiomatically on the *individual* harm allocation $H(\pi)$. This section treats DP, EO, and calibration as *surrogates*: under restrictive structural assumptions, they may proxy axiomatic harm fairness, but in general they do not characterize it.

Statistical metrics (notation). Let $A \in \mathcal{A}$ denote a group attribute and let π be a (possibly randomized) policy. We use the following standard group-level definitions:

- *Demographic parity (DP):* $\Pr_{\mathcal{D},\pi}(\hat{Y} = 1 \mid A = a)$ is equal across all $a \in \mathcal{A}$.

- *Equalized odds (EO):* $\Pr_{\mathcal{D},\pi}(\hat{Y} = 1 \mid Y = y, A = a)$ is equal across groups for each $y \in \{0, 1\}$ (equivalently, $\mathrm{FPR}_a(\pi)$ and $\mathrm{TPR}_a(\pi)$ are equal across groups).
- *Equal opportunity (EOp):* $\mathrm{TPR}_a(\pi)$ is equal across groups.
- *Groupwise calibration:* for $S = s(X) \in [0, 1]$, $\Pr(Y = 1 \mid S = u, A = a) = u$ for all u and a (up to discretization or binning in practice).

A bridge quantity: group-average harm under cost restrictions. To relate rate constraints to harm, consider the group-average harm $\overline{H}_a(\pi) := \mathbb{E}[H_i(\pi) \mid A = a]$. When subjective costs are *group-homogeneous*, that is, when $c_i^{\mathrm{FP}} = c_a^{\mathrm{FP}}$ and $c_i^{\mathrm{FN}} = c_a^{\mathrm{FN}}$ for all i with $A = a$, we have

$$\overline{H}_a(\pi) = c_a^{\mathrm{FP}} \Pr(Y = 0 \mid A = a)\, \mathrm{FPR}_a(\pi) + c_a^{\mathrm{FN}} \Pr(Y = 1 \mid A = a)\, \mathrm{FNR}_a(\pi). \tag{7.1}$$

This identity provides the main channel through which rate constraints can, in some cases, proxy harm-related notions.

7.1 Representation conditions: when do statistical constraints proxy harm axioms?

Theorem 7.1 (Representation conditions)**.** *Fix an axiom family on harm allocations (Section 5). If a statistical constraint family $M \in \{\mathrm{DP}, \mathrm{EO/EOp}, \textit{Calibration}\}$ is to imply small axiom violations* uniformly *over admissible instances, then additional structural restrictions are necessary, including at least restrictions that collapse subjective costs into a comparable unit (e.g., homogeneity within groups or a low-dimensional cost family aligned with the metric).*

Proof. Fix $(\mathcal{D}, \pi)$ such that the metric constraint $M(\pi)$ holds. Pick any two individuals $i \neq j$ within the same group $A = a$.

Assume that their individual error lotteries differ, which is generic unless π equalizes individuals:

$$\Delta_{\mathrm{FN}} := \Pr_{\pi}(\mathrm{FN} \mid i) - \Pr_{\pi}(\mathrm{FN} \mid j) \neq 0 \quad \text{or} \quad \Delta_{\mathrm{FP}} := \Pr_{\pi}(\mathrm{FP} \mid i) - \Pr_{\pi}(\mathrm{FP} \mid j) \neq 0.$$

Since M depends only on group-level rates or distributions, modifying the cost profile $\{c_\ell\}_{\ell=1}^n$ does not affect $M(\pi)$.

Without loss of generality, suppose $\Delta_{\mathrm{FN}} > 0$. Define a cost family in which only individual i is scaled:

$$c_i^{\mathrm{FN}} = C, \quad c_i^{\mathrm{FP}} = 0, \qquad c_\ell^{\mathrm{FN}} = c_\ell^{\mathrm{FP}} = 1 \;\; (\ell \neq i),$$

and let $C \to \infty$. Then

$$\begin{aligned}
H_{i \leftarrow i}(\pi) &= C \Pr_{\pi}(\mathrm{FN} \mid i), \\
H_{i \leftarrow j}(\pi) &= C \Pr_{\pi}(\mathrm{FN} \mid j), \\
\mathrm{EnvyGap}_{i,j}(\pi) &= \big[H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi)\big]_+ \\
&= \big[C(\Pr_{\pi}(\mathrm{FN} \mid i) - \Pr_{\pi}(\mathrm{FN} \mid j))\big]_+ \\
&= C\, \Delta_{\mathrm{FN}} \xrightarrow[C \to \infty]{} \infty.
\end{aligned}$$

Hence, without structural restrictions on costs (e.g., homogeneity, bounded cost ratios, or restricted cost families), a group-rate constraint M cannot uniformly upper bound harm-axiom violations. □

7.2 Counterexamples: metric satisfaction does not imply axiomatic harm fairness

Theorem 7.2 (Counterexample under heterogeneous costs)**.** *For each metric family $M \in \{\mathrm{DP}, \mathrm{EO/EOp}, \textit{Calibration}\}$ and any tolerance $\eta \geq 0$, there exist a distribution $\mathcal{D}$, heterogeneous subjective costs $\{c_i\}_{i=1}^n$, and a policy π such that $M(\pi) \leq \eta$ while $\mathrm{Envy}^{\max}(\pi)$ and $\mathrm{Prop}^{\max}(\pi)$ are arbitrarily large.*

Proof. We give explicit constructions. In both cases, the harm-axiom violations are driven by within-group cost heterogeneity together with the fact that M is blind to costs.

Case 1: $M \in \{\mathrm{DP}, \mathrm{EO}/\mathrm{EOp}\}$ **(two-group, nondegenerate construction).** Let $A \in \{a, b\}$ and $Y \in \{0, 1\}$ with $\Pr(Y = 1 \mid A = a) = \Pr(Y = 1 \mid A = b) = 1/2$. Let a discrete score $S \in \{0, 1\}$ satisfy

$$\Pr(S = 1 \mid Y = 1, A = a) = 0.8, \quad \Pr(S = 1 \mid Y = 0, A = a) = 0.2,$$

$$\Pr(S = 1 \mid Y = 1, A = b) = 0.4, \quad \Pr(S = 1 \mid Y = 0, A = b) = 0.1.$$

Define a group-wise randomized post-processing policy π by

$$\Pr_\pi(\hat{Y} = 1 \mid S = 1, A = a) = r_a := \tfrac{1}{2}, \qquad \Pr_\pi(\hat{Y} = 1 \mid S = 0, A = a) = 0,$$

$$\Pr_\pi(\hat{Y} = 1 \mid S = 1, A = b) = r_b := 1, \qquad \Pr_\pi(\hat{Y} = 1 \mid S = 0, A = b) = 0.$$

Then the group-conditional rates are

$$\mathrm{TPR}_a(\pi) = r_a \cdot 0.8 = 0.4, \quad \mathrm{FPR}_a(\pi) = r_a \cdot 0.2 = 0.1,$$

$$\mathrm{TPR}_b(\pi) = r_b \cdot 0.4 = 0.4, \quad \mathrm{FPR}_b(\pi) = r_b \cdot 0.1 = 0.1,$$

so EO (and hence EOp) holds exactly and in a nondegenerate way. Moreover, the acceptance rates also match:

$$\begin{aligned}
\Pr(\hat{Y} = 1 \mid A = a) &= r_a \Big(\Pr(Y = 1 \mid A = a) \Pr(S = 1 \mid Y = 1, A = a) \\
&\qquad + \Pr(Y = 0 \mid A = a) \Pr(S = 1 \mid Y = 0, A = a) \Big) \\
&= \tfrac{1}{2} \Big(\tfrac{1}{2} \cdot 0.8 + \tfrac{1}{2} \cdot 0.2 \Big) = 0.25, \\
\Pr(\hat{Y} = 1 \mid A = b) &= r_b \Big(\tfrac{1}{2} \cdot 0.4 + \tfrac{1}{2} \cdot 0.1 \Big) = 0.25,
\end{aligned}$$

so DP also holds exactly. Hence $M(\pi) = 0 \le \eta$ for every $\eta \ge 0$.

Now pick two individuals $i \neq j$ in group $A = a$ with $y_i = y_j = 1$ and $(s_i, s_j) = (0, 1)$. Then

$$\Pr_\pi(\mathrm{FN} \mid i) = 1, \qquad \Pr_\pi(\mathrm{FN} \mid j) = 1 - r_a = \tfrac{1}{2}.$$

Choose costs

$$c_i^{\mathrm{FN}} = C, \quad c_j^{\mathrm{FN}} = 1, \qquad c_i^{\mathrm{FP}} = c_j^{\mathrm{FP}} = 0,$$

and let all other individuals, if any, have bounded costs (e.g., $c_\ell^{\mathrm{FP}} = c_\ell^{\mathrm{FN}} = 1$). Using the cross-evaluated harms $H_{i \leftarrow k}(\pi) = c_i^{\mathrm{FN}} \Pr_\pi(\mathrm{FN} \mid k)$, we obtain

$$H_{i \leftarrow i}(\pi) = C, \qquad H_{i \leftarrow j}(\pi) = \tfrac{C}{2}.$$

Therefore

$$\mathrm{EnvyGap}_{i,j}(\pi) = \big[H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi) \big]_+ = \tfrac{C}{2} \xrightarrow[C \to \infty]{} \infty.$$

For proportionality, let $\bar{H}_i(\pi) = \frac{1}{n} \sum_{k=1}^n H_{i \leftarrow k}(\pi)$ denote the population mean harm under i's valuation. Since $H_{i \leftarrow k}(\pi) \le C$ for all k and $H_{i \leftarrow j}(\pi) = C/2$, we have

$$\bar{H}_i(\pi) \le \frac{C + (n-2)C + C/2}{n} = C \Big(1 - \frac{1}{2n} \Big),$$

hence

$$\mathrm{PropExcess}_i(\pi) = \big[H_{i \leftarrow i}(\pi) - \bar{H}_i(\pi) \big]_+ \ge \frac{C}{2n} \xrightarrow[C \to \infty]{} \infty,$$

so $\mathrm{Prop}^{\max}(\pi)$ is likewise arbitrarily large.

Case 2: $M =$ **Calibration.** Let there be a single group, or equivalently fix any group $A = a$, and let a score $S \in \{u_0, u_1\} \subset (0, 1)$ satisfy

$$\Pr(Y = 1 \mid S = u) = u \quad \text{for } u \in \{u_0, u_1\},$$

so that S is perfectly calibrated and the groupwise calibration error is zero. Define π to be the randomized classifier

$$\hat{Y} \sim \text{Bernoulli}(S),$$

that is, $\Pr_\pi(\hat{Y} = 1 \mid S = u) = u$. Then calibration holds exactly, and hence $M(\pi) = 0 \leq \eta$.

Pick two individuals $i \neq j$ with $y_i = y_j = 1$ and $(s_i, s_j) = (u_0, u_1)$, where $u_1 > u_0$. Then

$$\Pr_\pi(\text{FN} \mid i) = 1 - u_0, \qquad \Pr_\pi(\text{FN} \mid j) = 1 - u_1,$$

so $\Pr_\pi(\text{FN} \mid i) - \Pr_\pi(\text{FN} \mid j) = u_1 - u_0 > 0$. Choose costs

$$c_i^{\text{FN}} = C, \quad c_j^{\text{FN}} = 1, \qquad c_i^{\text{FP}} = c_j^{\text{FP}} = 0.$$

Then

$$\text{EnvyGap}_{i,j}(\pi) = \big[C(1 - u_0) - C(1 - u_1)\big]_+ = C(u_1 - u_0) \xrightarrow[C \to \infty]{} \infty.$$

Moreover, since $S \in \{u_0, u_1\}$ implies $H_{i \leftarrow k}(\pi) \leq C(1 - u_0) = H_{i \leftarrow i}(\pi)$ for all k and $H_{i \leftarrow j}(\pi) = C(1 - u_1)$, we have

$$\bar{H}_i(\pi) \leq \frac{(n-1)H_{i \leftarrow i}(\pi) + H_{i \leftarrow j}(\pi)}{n} = H_{i \leftarrow i}(\pi) - \frac{C(u_1 - u_0)}{n},$$

hence

$$\text{PropExcess}_i(\pi) = \big[H_{i \leftarrow i}(\pi) - \bar{H}_i(\pi)\big]_+ \geq \frac{C(u_1 - u_0)}{n} \xrightarrow[C \to \infty]{} \infty.$$

Combining the two cases establishes the claim for each $M \in \{\text{DP}, \text{EO/EOp}, \text{Calibration}\}$. □

7.3 Incompatibility reinterpreted as axiom conflict

Proposition 7.3 (Impossibility as axiom conflict)**.** *When groups have different base rates and subjective value is not collapsed into a common unit, statistical requirements such as calibration and EO encode distinct normative constraints that may correspond to conflicting directions in harm-axiom space. In general, they cannot be jointly satisfied without incurring nontrivial violations of at least one harm-axiom family (e.g., envy or proportionality) beyond a fixed tolerance.*

Proof. Let $A \in \{a, b\}$ with distinct base rates

$$\mu_a := \Pr(Y = 1 \mid A = a) \neq \mu_b := \Pr(Y = 1 \mid A = b).$$

Consider a discretized score $S \in \{u_0, u_1\}$ with a deterministic decision rule

$$\hat{Y} := \mathbb{1}\{S = u_1\}.$$

Assume groupwise calibration of the score:

$$\Pr(Y = 1 \mid S = u_1, A = a) = u_1, \qquad \Pr(Y = 1 \mid S = u_1, A = b) = u_1.$$

Then the positive predictive value (PPV) of the induced classifier is

$$\text{PPV}_a := \Pr(Y = 1 \mid \hat{Y} = 1, A = a) = u_1, \qquad \text{PPV}_b := \Pr(Y = 1 \mid \hat{Y} = 1, A = b) = u_1,$$

so $\text{PPV}_a = \text{PPV}_b$.

Now impose EO:

$$\text{TPR}_a = \text{TPR}_b =: \text{TPR}, \qquad \text{FPR}_a = \text{FPR}_b =: \text{FPR}.$$

By Bayes' rule,

$$\mathrm{PPV}_g = \Pr(Y = 1 \mid \hat{Y} = 1, A = g) = \frac{\mu_g \,\mathrm{TPR}}{\mu_g \,\mathrm{TPR} + (1 - \mu_g)\,\mathrm{FPR}}, \qquad g \in \{a, b\}.$$

Hence, if $\mu_a \neq \mu_b$ and the classifier is nondegenerate (i.e., $\mathrm{TPR} \neq \mathrm{FPR}$),

$$\mathrm{PPV}_a \neq \mathrm{PPV}_b,$$

contradicting $\mathrm{PPV}_a = \mathrm{PPV}_b = u_1$, which is implied by calibration. Therefore, under unequal base rates, calibration and EO cannot simultaneously hold except in degenerate cases (e.g., perfect prediction with $\mathrm{FPR} = 0, \mathrm{TPR} = 1$ or an uninformative classifier with $\mathrm{TPR} = \mathrm{FPR}$).

Finally, whenever one relaxes either constraint, the resulting discrepancy in group-level errors can be translated into axiom violations under suitable cost profiles, whether group-homogeneous or heterogeneous, via the linear harm identity in equation 7.1 together with scaling arguments; see Appendix A and Appendix B for the formal constructions. □

7.4 Lower bounds under bounded rationality: an irreducible floor on harm-based fairness

Theorem 7.4 (A tight lower bound under information and policy constraints)**.** *Fix $\epsilon \in (0, 1)$ and let the mediator observe proxy costs $\hat{c}_1 = \hat{c}_2 = (1, 1)$ with componentwise uncertainty*

$$\mathcal{C}(\hat{c}, \epsilon) = \Big\{ c : \; c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}} \in [1 - \epsilon,\, 1 + \epsilon] \text{ for } i = 1, 2 \Big\}.$$

Consider the restricted policy class Π_{simple} that enforces identical treatment:

$$\Pr_{\pi}(\hat{Y} = 1 \mid 1) = \Pr_{\pi}(\hat{Y} = 1 \mid 2) =: \alpha \in [0, 1].$$

Then there exists a two-individual instance (with $y_1 = 1$ and $y_2 = 0$) such that the minimax envy violation satisfies the tight *equality*

$$\inf_{\pi \in \Pi_{simple}} \sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{Envy}^{\max}(\pi; c) \;=\; \epsilon,$$

where $\mathrm{Envy}^{\max}(\pi; c) = \max\{\mathrm{EnvyGap}_{1,2}(\pi; c),\, \mathrm{EnvyGap}_{2,1}(\pi; c)\}$.

Proof. Let $n = 2$ with labels $y_1 = 1$ and $y_2 = 0$. Under the identical-treatment constraint,

$$\Pr_{\pi}(\mathrm{FN} \mid 1) = 1 - \alpha, \qquad \Pr_{\pi}(\mathrm{FP} \mid 2) = \alpha.$$

Recall that

$$\mathrm{EnvyGap}_{1,2}(\pi; c) = \big[c_1^{\mathrm{FN}}(1 - \alpha) - c_1^{\mathrm{FP}}\alpha\big]_+,$$
$$\mathrm{EnvyGap}_{2,1}(\pi; c) = \big[c_2^{\mathrm{FP}}\alpha - c_2^{\mathrm{FN}}(1 - \alpha)\big]_+.$$

Because c_1^{FN} and c_1^{FP} vary independently over $[1 - \epsilon, 1 + \epsilon]$, the worst case for $\mathrm{EnvyGap}_{1,2}$ is obtained by maximizing the linear form inside $[\cdot]_+$, that is,

$$\sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{EnvyGap}_{1,2}(\pi; c) = \Big[(1 + \epsilon)(1 - \alpha) - (1 - \epsilon)\alpha\Big]_+ = \big[1 + \epsilon - 2\alpha\big]_+,$$

achieved, for example, at $(c_1^{\mathrm{FN}}, c_1^{\mathrm{FP}}) = (1 + \epsilon,\, 1 - \epsilon)$.

Similarly,

$$\sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{EnvyGap}_{2,1}(\pi; c) = \Big[(1 + \epsilon)\alpha - (1 - \epsilon)(1 - \alpha)\Big]_+ = \big[2\alpha - (1 - \epsilon)\big]_+,$$

achieved, for example, at $(c_2^{\mathrm{FP}}, c_2^{\mathrm{FN}}) = (1+\epsilon,\, 1-\epsilon)$.

Since the costs for individuals 1 and 2 are independent within $\mathcal{C}(\hat{c}, \epsilon)$, the worst-case maximum envy is

$$\sup_{c \in \mathcal{C}(\hat{c},\epsilon)} \mathrm{Envy}^{\max}(\pi; c) = \max\Big\{ \big[1+\epsilon-2\alpha\big]_+,\ \big[2\alpha-(1-\epsilon)\big]_+ \Big\}.$$

Minimizing over $\alpha \in [0,1]$ yields

$$\inf_{\alpha \in [0,1]} \max\Big\{ \big[1+\epsilon-2\alpha\big]_+,\ \big[2\alpha-(1-\epsilon)\big]_+ \Big\} = \epsilon,$$

which is achieved at $\alpha = \frac{1}{2}$, where both bracketed terms equal ϵ. Thus

$$\inf_{\pi \in \Pi_{\mathrm{simple}}} \sup_{c \in \mathcal{C}(\hat{c},\epsilon)} \mathrm{Envy}^{\max}(\pi; c) \;=\; \epsilon,$$

completing the proof. □

Interpretation (why this is a "price"). Under policy coupling (identical treatment), an ϵ-radius uncertainty set over values allows Nature to choose costs so that either individual becomes the worst-off agent for the same choice of α. As a result, an ϵ-level residual envy is unavoidable, which can be interpreted as the irreducible price of fairness under simultaneous informational and policy constraints.

8 Experiments

8.1 Experiment 1: Surrogate failure under cost heterogeneity

Experimental setup. We test whether standard statistical fairness objectives provide a reliable guide to *harm-based* (axiomatic) fairness when individuals have heterogeneous error costs. We use the common ACS 2018 (Folktables) setup (CA-only for RAM safety). The sensitive attribute is `SEX`; it is used only as the group label and is removed from the feature set. We run three binary classification tasks (ACSIncome, ACSPublicCoverage, ACSEmployment) with three random seeds; in tables, we abbreviate them as Inc, Cov, and Emp. For each seed, we subsample 50,000 rows *before* converting the dataframe to numpy, and split into train/validation/test (60/20/20, stratified). We train two score models per task: logistic regression (LR) and histogram gradient boosting (HGB). Let $s(x) \in [0,1]$ denote the predicted score.

From each trained score model, we construct families of post-processed decision policies on the validation set and evaluate them on the test set: (i) **BASE**: global threshold at 0.5; (ii) **DP**: group-specific thresholds chosen to achieve a target acceptance rate q per group (sweeping q over a small grid); (iii) **EO**: group-specific thresholds selected by a grid search to (approximately) satisfy an EO tolerance on validation; and (iv) **CAL**: isotonic regression calibration on validation, then mixing calibrated and raw scores with a coefficient λ and thresholding at 0.5.

Each individual i has subjective FP/FN costs (c_i^{FP}, c_i^{FN}). We generate costs using two *orthogonal* heterogeneity knobs: a between-group *scale* parameter γ and a within-group *preference* parameter σ,

$$m_i = \begin{cases} e^{\gamma} & (A_i = 1) \\ e^{-\gamma} & (A_i = 0) \end{cases}, \quad u_i \sim \mathcal{N}(0, \sigma^2), \quad c_i^{FP} = m_i e^{u_i},\ \ c_i^{FN} = m_i e^{-u_i}.$$

We use this family as a *controlled synthetic cost model* to vary heterogeneity along interpretable axes, rather than as a claim about real-world cost structure. It is designed to (a) ensure strictly positive costs, (b) separate group-level burden scaling from within-group FP–FN trade-offs, and (c) yield a transparent cost-stratum axis: $\log(c_i^{FP}/c_i^{FN}) = 2u_i$ controls the relative FP-vs-FN preference, while the common multiplier m_i controls the overall harm scale.[2] Given a policy π producing predictions $\hat{y}$, the realized harm is $H_i(\pi) = c_i^{FP}\mathbf{1}[\mathrm{FP}_i] + c_i^{FN}\mathbf{1}[\mathrm{FN}_i]$.

[2]Note that $c_i^{FP} c_i^{FN} = m_i^2$, so γ changes the overall scale (between groups) without changing the preference ratio, whereas σ changes the preference ratio distribution without changing the scale within each group.

For scalable harm-based evaluation, we summarize envy using the cost-stratum proxy from Section 5. Concretely, we partition individuals into K cost strata by quantiles of $\log(c^{FP}/c^{FN})$ and report the resulting max cost-stratum envy $\widetilde{\text{Envy}}^{\max}$. In this experiment we use a finer partition with $K = 80$ cost strata.

For succinct reporting of harm outcomes, we focus on two regimes: a homogeneous-cost sanity check $(\gamma, \sigma) = (0, 0)$ and a strong-heterogeneity case $(\gamma, \sigma) = (2, 1)$. For each (task, model, seed, regime), we report **BASE** and one representative policy from each family (DP/EO/CAL), chosen as the envy-minimizing candidate among policies that satisfy a small feasibility tolerance on the corresponding statistical objective (DP gap / EO gap / ECE) on the *validation set*; if no candidate satisfies the tolerance, we report the envy-minimizing candidate within that family. All metrics in Table 1 are evaluated on the *test set*.

To quantify mismatch *within* each family under strong heterogeneity, we additionally compute a *discordance rate* at $(\gamma, \sigma) = (2, 1)$: among all pairs of policies inside the same family, the fraction of pairs for which the policy with the *better* statistical score (smaller DP gap / EO gap / ECE) has the *worse* cost-stratum envy (larger $\widetilde{\text{Envy}}^{\max}$). This yields a compact summary of how often "improving the surrogate" moves harm-based fairness in the opposite direction within the same policy family.

Results and analysis. Table 1 reports seed-averaged cost-stratum envy under the homogeneous-cost sanity check and under strong heterogeneity. Under $(\gamma, \sigma) = (0, 0)$, all individuals share the same FP–FN preference ratio, so the cost-stratum partition induced by $\log(c^{FP}/c^{FN})$ collapses to (effectively) a single stratum. Consequently, the cost-stratum envy proxy is mechanically near zero across policies because there is no cross-stratum variation to compare. This sanity check is included only to emphasize that the surrogate-failure effects we study are driven by *value heterogeneity*; it does *not* claim individual-level envy-freeness.

Under $(\gamma, \sigma) = (2, 1)$, harm-based outcomes vary substantially across objectives. DP-selected policies consistently reduce cost-stratum envy relative to **BASE** across all task–model pairs in our setup. In contrast, EO-selected policies often increase cost-stratum envy—sometimes sharply (e.g., Emp-HGB and Inc-LR)—with only limited cases of slight improvement. Calibration-based post-processing shows smaller and mixed effects, frequently remaining close to **BASE** and occasionally reducing envy.

The discordance panel summarizes the core surrogate-failure phenomenon in a single number per family. Discordance measures the fraction of within-family policy pairs for which a *smaller* surrogate error (DP gap / EO gap / ECE) coincides with a *larger* cost-stratum envy. Discordance remains sizable across tasks and models, indicating that, even within a fixed objective family, optimizing the statistical surrogate can frequently move harm-based fairness in the wrong direction under subjective-cost heterogeneity. EO can exhibit particularly high discordance in some settings (e.g., Emp-HGB and Inc-HGB), while the extent of discordance is task-dependent (e.g., Cov-HGB).

8.2 Experiment 2: Bounded rationality and lower bounds under information and policy constraints

Experimental setup. This experiment operationalizes the bounded-rationality claim: even when a mediator *targets* a harm-based axiom, imperfect value information and a restricted policy class can leave a nonzero residual violation.

We reuse the same ACS 2018 (Folktables) benchmark, prediction tasks, score models, and train/validation/test splits as in Experiment 1, and treat the resulting scores $s(x) \in [0, 1]$ as fixed. To isolate the role of informational and policy constraints, we hold the subjective-cost regime fixed throughout, using the same cost family as in Experiment 1 with $(\gamma, \sigma) = (2, 1)$, and evaluate all axiom violations on the test set under these true costs.

On validation, however, the mediator does not observe the true costs (c_i^{FP}, c_i^{FN}). Instead, it observes noisy proxies $(\hat{c}_i^{FP}, \hat{c}_i^{FN})$ with bounded multiplicative error:

$$\hat{c}_i^{FP} = c_i^{FP} e^{z_i^{FP}}, \quad \hat{c}_i^{FN} = c_i^{FN} e^{z_i^{FN}}, \quad z_i^{FP}, z_i^{FN} \sim \text{Unif}[-\epsilon, \epsilon],$$

so that $|\log(\hat{c}/c)| \leq \epsilon$. We sweep $\epsilon \in \{0, 0.25, 0.5, 1.0\}$.

Task	Model	(γ=0, σ=0): envy			
		BASE	DP	EO	CAL
Emp	HGB	0.0000	0.0000	0.0000	0.0000
	LR	0.0000	0.0000	0.0000	0.0000
Inc	HGB	0.0000	0.0000	0.0000	0.0000
	LR	0.0000	0.0000	0.0000	0.0000
Cov	HGB	0.0000	0.0000	0.0000	0.0000
	LR	0.0000	0.0000	0.0000	0.0000

(a)

Task	Model	(γ=2, σ=1): envy			
		BASE	DP	EO	CAL
Emp	HGB	6.2972	3.8361	8.2062	6.2963
	LR	6.4179	3.8513	7.4175	6.4179
Inc	HGB	6.1190	4.6799	7.1064	6.1164
	LR	6.7336	5.1740	9.4393	6.5494
Cov	HGB	6.8917	5.6462	8.1721	6.6435
	LR	8.6135	5.7380	8.4219	8.0447

(b)

Task	Model	(γ=2, σ=1): Δ (BASE)		
		ΔDP	ΔEO	ΔCAL
Emp	HGB	-2.4611	1.9090	-0.0009
	LR	-2.5666	0.9996	0.0000
Inc	HGB	-1.4391	0.9874	-0.0026
	LR	-1.5596	2.7057	-0.1842
Cov	HGB	-1.2455	1.2804	-0.2482
	LR	-2.8755	-0.1916	-0.5688

(c)

Task	Model	(γ=2, σ=1): discordance rate		
		DP	EO	CAL
Emp	HGB	0.4921	0.8889	0.4000
	LR	0.5556	0.6667	0.2333
Inc	HGB	0.4921	0.7778	0.3333
	LR	0.4603	0.6667	0.1000
Cov	HGB	0.3492	0.2222	0.5000
	LR	0.3333	0.6667	0.4000

(d)

Table 1: Summary of statistical-fairness metrics and harm-based fairness outcomes across tasks and models, averaged over three random seeds, under homogeneous and heterogeneous subjective-cost regimes. Envy is reported as max cost-stratum envy $\widetilde{\text{Envy}}^{\max}$ using $K = 80$.

We also study coarse value information by quantizing individuals into k proxy *cost strata* using quantiles of $\log(\hat{c}^{FP}/\hat{c}^{FN})$: $k = 0$ denotes individual-level proxy costs, while $k \in \{2, 5, 10\}$ restricts the mediator to stratum-mean proxy costs.

The mediator then selects a policy from a simple threshold family on validation. We consider: (i) a *single global threshold*; and (ii) *group-specific thresholds* (t_0, t_1). Candidate thresholds are taken from score quantiles (13 grid points). For each (ϵ, k, Π), the mediator chooses the policy minimizing a proxy envy-gap objective on validation, where envy is computed using the same cost-stratum proxy as in Section 5, with strata defined from the proxy costs.

Throughout this experiment, we use a finer-grained envy calculation based on a partition into 80 cost strata, both for proxy-based policy selection on validation and for evaluation on the test set under true costs.

For compact reporting, we summarize four representative information regimes:

- **Oracle (no noise, individual-level):** $\epsilon = 0,\ k = 0$.
- **Noisy (individual-level):** $\epsilon = 1,\ k = 0$.
- **Coarse (no noise, stratum-level):** $\epsilon = 0,\ k = 2$.
- **Noisy+coarse (stratum-level):** $\epsilon = 1,\ k = 2$.

We report the realized test envy-gap δ and $\Delta = \delta - \delta_{\mathbf{BASE}}$, where $\delta_{\mathbf{BASE}}$ is computed under the global threshold at 0.5.

Results and analysis. Tables 2–3 report δ and Δ (averaged over three random seeds) using the fine-grained 80-stratum envy measure.

First, even with oracle information ($\epsilon = 0, k = 0$), δ remains well above zero across tasks and models, indicating that simple threshold policies cannot eliminate worst-stratum envy under strong heterogeneity;

this is the policy-class floor in operational form. Moreover, under this tail-sensitive envy measure, oracle selection does not uniformly dominate the fixed baseline threshold: for global thresholds, oracle selection improves on **BASE** in some settings (negative Δ), while worsening it in others (positive Δ).

Second, imperfect value information can substantially perturb true-cost harm-fairness. Under noisy and/or coarse proxies, the mediator can select a policy that looks favorable under proxy envy on validation yet yields higher true-cost envy on test (positive Δ), and the direction can vary by task and model. This behavior is consistent with bounded rationality under value uncertainty: limited and distorted value information, combined with a restricted policy class and finite-sample selection, can move worst-stratum harm-based fairness away from the baseline even when the mediator explicitly optimizes the harm-based proxy.

Finally, expanding the policy class from a single global threshold to group-specific thresholds does not guarantee better harm-based outcomes under this metric. In several settings, group-threshold selection increases worst-stratum envy relative to **BASE** by a wide margin, reinforcing that additional degrees of freedom can exacerbate tail harms when selection is driven by imperfect proxy information.

Task	Model	δ			
		Oracle	Noisy	Coarse	Noisy+coarse
Emp	HGB	4.3838	6.4273	5.9476	6.2358
	LR	5.0616	5.6175	8.5032	5.7925
Inc	HGB	6.9679	6.3603	7.0954	7.7142
	LR	8.0557	7.1278	7.6640	8.3580
Cov	HGB	13.7500	11.4117	14.4939	15.8237
	LR	13.2415	13.8218	14.8084	15.4885

(a)

Task	Model	Δ			
		Oracle	Noisy	Coarse	Noisy+coarse
Emp	HGB	-0.8429	1.2007	0.7210	1.0092
	LR	-0.3962	0.1596	3.0454	0.3347
Inc	HGB	0.9348	0.3271	1.0623	1.6810
	LR	1.7349	0.8070	1.3432	2.0372
Cov	HGB	1.8620	-0.4762	2.6059	3.9357
	LR	-1.4834	-0.9031	0.0836	0.7636

(b)

Table 2: Realized envy-gap δ and change $\Delta = \delta - \delta_{\mathbf{BASE}}$ under true costs $(\gamma, \sigma) = (2, 1)$ for single global-threshold policies, averaged over three random seeds. Envy is computed as the maximum cost-stratum envy using a fine-grained partition into 80 cost strata.

Task	Model	δ			
		Oracle	Noisy	Coarse	Noisy+coarse
Emp	HGB	7.2337	6.5505	7.3212	4.9492
	LR	5.8432	5.8149	8.7682	6.3713
Inc	HGB	7.4117	5.5227	6.7408	7.5837
	LR	6.6658	5.5445	7.4765	8.1880
Cov	HGB	19.2726	16.7099	18.5463	18.5463
	LR	16.5319	12.6476	15.4844	15.5148

(a)

Task	Model	Δ			
		Oracle	Noisy	Coarse	Noisy+coarse
Emp	HGB	2.0070	1.3238	2.0946	-0.2774
	LR	0.3854	0.3571	3.3104	0.9135
Inc	HGB	1.3786	-0.5105	0.7077	1.5506
	LR	0.3451	-0.7763	1.1557	1.8672
Cov	HGB	7.3846	4.8220	6.6584	6.6584
	LR	1.8070	-2.0773	0.7595	0.7900

(b)

Table 3: Realized envy-gap δ and change $\Delta = \delta - \delta_{\mathbf{BASE}}$ under true costs $(\gamma, \sigma) = (2, 1)$ for group-specific threshold policies, averaged over three random seeds. Envy is computed as the maximum cost-stratum envy using a fine-grained partition into 80 cost strata.

8.3 Experiment 3: Axiomatic profiling of DP/EO/Calibration under fixed cost heterogeneity

Experimental setup. This experiment profiles how standard statistical objectives behave when fairness is evaluated as harm allocation. We reuse the same ACS 2018 (Folktables) benchmark, prediction tasks, score models, and train/validation/test splits as in Experiment 1, and treat the resulting scores $s(x) \in [0, 1]$ as fixed. In tables, we abbreviate the three tasks as Inc, Cov, and Emp.

For harm-based evaluation, we again use the same subjective-cost regime as in Experiment 1, fixing $(\gamma, \sigma) = (2, 1)$ on the test set, and define realized harm under a policy π as $H_i(\pi) = c_i^{FP}\mathbf{1}[\mathrm{FP}_i] + c_i^{FN}\mathbf{1}[\mathrm{FN}_i]$.

Starting from each trained score model, we construct notion-targeted *post-processing interventions* and evaluate the induced harm allocations on the test set. For each (task, model, seed), we report one representative policy per family: (i) **BASE**: a global threshold at 0.5; (ii) **DP**: group-specific thresholds induced by sweeping target acceptance rates q, selecting the candidate with *minimum validation error* among those with DP gap ≤ 0.02 (best-effort if infeasible); (iii) **EO**: group-specific thresholds selected by validation grid search, choosing the candidate with *minimum validation error* among those with EO gap ≤ 0.02 (best-effort if infeasible); (iv) **CAL**: isotonic calibration on validation with a mixing coefficient λ (swept on a small grid), selecting the candidate with *minimum validation error* among those with ECE ≤ 0.02 (best-effort if infeasible). This tie-breaking rule yields a concise, practitioner-plausible profile for each notion: we impose a standard feasibility tolerance and then choose the most accurate feasible candidate, *without tuning to harm metrics*.

We report harm-based metrics—cost-stratum envy gap (envy), mean harm (mean_h), maximum harm (max_h), and tail excess over the BASE 90th-percentile threshold (prop_excess)—alongside accuracy and the surrogate metrics (DP gap, EO gap, ECE). Cost-stratum envy is computed using a finer-grained partition into 80 cost strata, formed by quantiles of $\log(c^{FP}/c^{FN})$ under the true test-set costs. All values are averaged over three random seeds.

Results and analysis. Experiment 3 addresses a definition-level question: under fixed strong heterogeneity, what *harm profile* is induced when we target different statistical notions (CAL/DP/EO) via standard post-processing? Tables 4 and 5 summarize these profiles by reporting, for each notion-targeted policy, both the achieved surrogate quantities and the resulting harm diagnostics.

The central pattern is *non-monotonicity across criteria*. Improvements in a surrogate metric do not reliably predict improvements in harm-based criteria because DP/EO/CAL constrain cost-blind, group-aggregated properties of the confusion profile, whereas our harm diagnostics depend on the *cost-weighted* distribution of errors across individuals. With this finer, more tail-sensitive envy measure, differences across notions can be amplified relative to coarser partitions, even when accuracy and other harm summaries (mean harm, tail excess) move only mildly.

DP provides a crisp illustration of "rate alignment without harm alignment." By construction it tends to attain a relatively small DP gap (Table 5), but it does not constrain how FP/FN *lotteries are redistributed across individuals* near the decision boundary. Under heterogeneous costs, even modest reshuffling can amplify cross-evaluated harm differences,

$$H_{i\leftarrow j}(\pi) = c_i^{\mathrm{FP}} \Pr_{\pi}(\mathrm{FP} \mid j) + c_i^{\mathrm{FN}} \Pr_{\pi}(\mathrm{FN} \mid j), \qquad \mathrm{EnvyGap}_{i,j}(\pi) = \big[H_{i\leftarrow i}(\pi) - H_{i\leftarrow j}(\pi)\big]_+,$$

yielding larger relational grievances despite improved rate parity. This mechanism is visible, for example, in Emp-HGB and Inc-HGB: DP achieves smaller DP gaps than **BASE**, yet its envy gap is larger (Table 4), showing that selection-rate equalization can coincide with a cost-sensitive reallocation that worsens worst-stratum envy. At the same time, DP can reduce envy in other settings (e.g., Cov-HGB), illustrating that the harm impact of a surrogate intervention is highly task- and model-dependent.

EO exhibits a different profile. When the EO feasibility tolerance is not well satisfied (best-effort selection), EO-targeted post-processing can display sizable EO gaps and large DP gaps (Table 5), and its harm consequences can be mixed: it can substantially reduce envy in some settings (notably Inc-LR), while increasing envy in others (e.g., Cov-HGB). Taken together, Tables 4 and 5 show that CAL/DP/EO encode different operational emphases, and under subjective-cost heterogeneity these emphases can induce distinct—and sometimes conflicting—harm allocations.

8.4 Experiment 4: Comparing fairness algorithms under harm-based axioms

Experimental setup. Experiment 4 asks an implementation-level robustness question: does the surrogate–axiom misalignment observed in Experiments 1–3 persist once we move beyond simple post-processing and compare a *diverse* set of fairness interventions spanning pre-, in-, and post-processing? Where Experiment 3 varies the *statistical notion* within a fixed post-processing construction, here we vary the *intervention family* itself while evaluating all methods in the same harm-based space.

Task	Model	Envy gap			
		BASE	CAL	DP	EO
Emp	HGB	5.3727	5.3754	5.8840	5.8060
	LR	6.3059	6.1411	6.0789	6.0727
Inc	HGB	6.5272	6.0411	7.1476	6.3352
	LR	7.1331	7.2374	7.8536	5.9515
Cov	HGB	13.0990	12.2058	10.6861	13.3671
	LR	10.7189	10.4490	12.9021	11.0700

(a)

Task	Model	Max harm			
		BASE	CAL	DP	EO
Emp	HGB	212.8618	212.8618	212.8618	212.8618
	LR	212.8618	190.4720	212.8618	212.8618
Inc	HGB	104.4634	104.4634	104.4634	104.4634
	LR	108.1216	108.1216	108.1216	108.1216
Cov	HGB	168.8830	168.8830	150.4242	175.8929
	LR	168.8830	168.8830	168.8830	168.8830

(b)

Task	Model	Mean harm			
		BASE	CAL	DP	EO
Emp	HGB	1.4630	1.4695	1.4760	1.4731
	LR	1.5837	1.5694	1.5591	1.5642
Inc	HGB	1.0894	1.0714	1.0465	1.0205
	LR	1.3280	1.3592	1.3303	1.3254
Cov	HGB	1.9602	1.9455	1.9584	2.0075
	LR	1.9811	1.9791	2.0001	1.9852

(c)

Task	Model	Tail excess			
		BASE	CAL	DP	EO
Emp	HGB	1.2068	1.2128	1.2204	1.2149
	LR	1.2768	1.2623	1.2554	1.2594
Inc	HGB	1.0257	1.0080	0.9827	0.9595
	LR	1.0875	1.1132	1.0908	1.0888
Cov	HGB	1.2661	1.2516	1.2613	1.3028
	LR	1.2568	1.2567	1.2696	1.2605

(d)

Table 4: Harm-based metrics under fixed strong cost heterogeneity $(\gamma, \sigma) = (2, 1)$ for representative policies (BASE, CAL, DP, EO), averaged over three random seeds. Envy gap is computed as the maximum cost-stratum envy using a finer-grained partition into 80 cost strata.

Task	Model	Accuracy			
		BASE	CAL	DP	EO
Emp	HGB	0.8146	0.8142	0.8137	0.8112
	LR	0.8023	0.8027	0.8027	0.8023
Inc	HGB	0.8095	0.8084	0.8050	0.8121
	LR	0.7691	0.7691	0.7674	0.7744
Cov	HGB	0.7090	0.7081	0.7069	0.7064
	LR	0.6949	0.6917	0.6905	0.6915

(a)

Task	Model	DP gap			
		BASE	CAL	DP	EO
Emp	HGB	0.0171	0.0164	0.0103	0.0844
	LR	0.0140	0.0162	0.0138	0.0572
Inc	HGB	0.0582	0.0587	0.0194	0.1376
	LR	0.0123	0.0106	0.0247	0.1678
Cov	HGB	0.0145	0.0193	0.0171	0.1116
	LR	0.0319	0.0271	0.0249	0.1010

(b)

Task	Model	ECE			
		BASE	CAL	DP	EO
Emp	HGB	0.0126	0.0099	0.0126	0.0126
	LR	0.0137	0.0150	0.0137	0.0137
Inc	HGB	0.0158	0.0215	0.0158	0.0158
	LR	0.0206	0.0147	0.0206	0.0206
Cov	HGB	0.0199	0.0237	0.0199	0.0199
	LR	0.0164	0.0242	0.0164	0.0164

(c)

Task	Model	EO gap			
		BASE	CAL	DP	EO
Emp	HGB	0.0949	0.0937	0.0949	0.1384
	LR	0.0870	0.0863	0.0839	0.1071
Inc	HGB	0.0333	0.0311	0.0998	0.0743
	LR	0.0806	0.0836	0.0778	0.1425
Cov	HGB	0.0288	0.0354	0.0572	0.1267
	LR	0.0395	0.0348	0.0645	0.1188

(d)

Table 5: Accuracy and surrogate metrics for the same representative policies as Table 4, averaged over three random seeds.

We reuse the same ACS 2018 (Folktables) benchmark, prediction tasks, and train/validation/test splits as in Experiment 1. All methods share the same base learner: a linear logistic classifier trained with `SGDClassifier` (`log_loss`) on one-hot encoded features with light regularization. This keeps the comparison CPU-friendly and focuses attention on the interventions rather than model capacity.

We compare six representative approaches spanning unconstrained training, pre-/in-processing interventions, and post-processing:

- **ERM** (Shalev-Shwartz & Ben-David, 2014): Unconstrained empirical risk minimization with no explicit disparity control.
- **FairBalance** (Kamiran & Calders, 2012): A pre-processing baseline that assigns sample weights by the four (A, Y) strata.
- **FairBatch** (Roh et al., 2020): An in-processing method that adaptively adjusts mini-batch composition and weights, primarily targeting demographic-parity-style balance.
- **GroupDRO** (Sagawa et al., 2019): Distributionally robust training over the four (A, Y) strata, emphasizing the highest-loss groups via exponentiated-gradient updates.
- **FairMixup** (Chuang & Mroueh, 2021): A data-augmentation method that performs mixup-style interpolation with emphasis on cross-group mixing.
- **FairProjection** (Hardt et al., 2016): A post-processing method that selects group-specific thresholds on validation via a coarse grid, enforcing equalized odds up to a small tolerance (best-effort if infeasible).

These methods do not optimize a single shared surrogate objective; rather, they represent distinct intervention families used to mitigate group disparities. Our goal is not to identify a universally best algorithm, but to test whether *algorithmic diversity* restores a monotone link between surrogate metrics and harm-axiom outcomes.

For harm-based evaluation, we again use the same subjective-cost regime as in Experiment 1, fixing $(\gamma, \sigma) = (2, 1)$ on the test set, and evaluate realized harm as $H_i(\pi) = c_i^{FP}\mathbf{1}[\mathrm{FP}_i] + c_i^{FN}\mathbf{1}[\mathrm{FN}_i]$. We report harm-based metrics (mean harm, max harm, tail excess over the ERM 90th-percentile harm threshold, and a cost-stratum envy gap) alongside accuracy and surrogate metrics (DP gap, EO gap, ECE). Cost-stratum envy is computed as the maximum cost-stratum envy using a finer-grained partition into 80 cost strata, formed by quantiles of $\log(c^{FP}/c^{FN})$ under the true test-set costs. All values are averaged over three random seeds.

Results and analysis. Tables 6 and 7 address the main robustness question: *algorithmic diversity does not remove the surrogate–axiom misalignment under strong cost heterogeneity.* Accuracy differences are modest across methods, yet the induced harm allocations differ markedly, and improvements in any single statistical metric do not uniquely determine the harm profile.

Across tasks, we observe persistent *non-dominance* across surrogate and harm criteria, now *across intervention families.* This is expected: the compared methods optimize different cost-blind objectives (or enforce different rate-level constraints), so they move the confusion profile in distinct ways, redistributing FP/FN outcomes across individuals near the decision boundary. With this finer, more tail-sensitive envy measure, relative differences in envy can be amplified even when mean harm and tail excess move only modestly.

The task-level patterns illustrate these mechanism-level differences. On Emp, FairBatch attains the smallest DP gap, and it also yields the smallest mean harm and tail excess among the compared methods; however, it does not minimize envy. Indeed, FairBalance, FairMixup, and FairProjection achieve lower envy than FairBatch despite having larger DP gaps, highlighting that regulating selection-rate parity does not determine worst-stratum relational grievances.

On Inc, methods exhibit sharp trade-offs: FairProjection achieves the lowest mean harm and tail excess but at the cost of a very large DP gap, while GroupDRO attains the lowest envy with substantially higher mean harm and tail excess than ERM. Meanwhile, FairBalance and FairMixup reduce envy modestly relative to ERM and improve calibration (ECE), but do not dominate jointly on harm criteria.

On Cov, FairBatch reduces envy the most relative to ERM, but it increases EO gap and max harm; FairProjection reduces mean harm and tail excess, whereas FairMixup yields relatively low envy while

producing the worst mean harm and tail excess among the compared methods. These patterns underscore that, under strong heterogeneous valuations, no single fairness algorithm consistently dominates across surrogate metrics and harm-based axioms; surrogate improvements are best interpreted as movements within a constrained attainable region rather than monotone progress toward a single fairness ideal.

Task	Method	Accuracy	DP gap	EO gap	ECE
Emp	ERM	0.7991	0.0146	0.0783	0.0407
	FairBalance	0.7977	0.0177	0.0902	0.0429
	FairBatch	0.7929	0.0053	0.0802	0.0471
	GroupDRO	0.7955	0.0132	0.0791	0.0442
	FairMixup	0.8010	0.0151	0.0806	0.0269
	FairProjection	0.7992	0.0607	0.1055	0.0407
Inc	ERM	0.7503	0.0041	0.0694	0.0807
	FairBalance	0.7656	0.0244	0.0930	0.0448
	FairBatch	0.7519	0.0386	0.0687	0.0801
	GroupDRO	0.7376	0.0246	0.0501	0.1012
	FairMixup	0.7602	0.0215	0.0631	0.0342
	FairProjection	0.7690	0.1505	0.1099	0.0807
Cov	ERM	0.6719	0.0437	0.0532	0.0927
	FairBalance	0.6705	0.0373	0.0579	0.1176
	FairBatch	0.6836	0.0457	0.0786	0.0989
	GroupDRO	0.6448	0.1131	0.1307	0.1218
	FairMixup	0.6342	0.0225	0.0494	0.1110
	FairProjection	0.6856	0.1031	0.1200	0.0927

Table 6: Accuracy and surrogate metrics for six fairness methods, averaged over three random seeds.

Task	Method	Mean harm	Max harm	Tail excess	Envy gap
Emp	ERM	1.5734	212.8618	1.2685	6.5496
	FairBalance	1.5806	190.4720	1.2702	5.9554
	FairBatch	1.5558	190.4720	1.2420	5.6830
	GroupDRO	1.5667	182.2880	1.2518	6.5817
	FairMixup	1.5717	212.8618	1.2675	5.9184
	FairProjection	1.5871	212.8618	1.2810	5.8274
Inc	ERM	1.4685	108.1216	1.1559	6.0324
	FairBalance	1.3811	107.4177	1.0898	5.7819
	FairBatch	1.4879	148.2175	1.1738	7.0952
	GroupDRO	1.5973	149.0032	1.2650	6.0182
	FairMixup	1.4243	107.4177	1.1351	7.4050
	FairProjection	1.3353	108.1216	1.0623	6.7323
Cov	ERM	2.0604	168.8830	1.2435	13.8251
	FairBalance	2.2162	163.3486	1.3606	12.1770
	FairBatch	2.0944	175.8929	1.2998	10.7925
	GroupDRO	2.2189	136.9072	1.3543	12.9062
	FairMixup	2.4023	131.5751	1.4496	11.0125
	FairProjection	2.0119	168.8830	1.2165	11.9692

Table 7: Harm-based metrics under fixed strong subjective-cost heterogeneity $(\gamma, \sigma) = (2, 1)$, averaged over three random seeds. Tail excess is measured relative to the ERM 90th-percentile harm threshold. Envy gap is computed as the maximum cost-stratum envy using a finer-grained partition into 80 cost strata.

9 Discussion

This paper reframes algorithmic fairness as *harm allocation* under heterogeneous subjective error costs and evaluates statistical constraints (DP, EO, and calibration) as *surrogates* for axiomatic fairness over the induced harm vector. The theory establishes (i) representation conditions under which a statistical notion can serve as a valid surrogate (Theorem 7.1), (ii) counterexamples showing that surrogate satisfaction can be arbitrarily misaligned with harm-based axioms under cost heterogeneity (Theorem 7.2), (iii) a reinterpretation of classic incompatibilities as *axiom conflicts* (Proposition 7.3), and (iv) an operational lower bound on attainable axiom satisfaction under informational and policy-class constraints (Theorem 7.4). The experiments (Section 8) complement these results by tracing when surrogate improvement coincides with, diverges from, or even worsens harm-based fairness.

9.1 Surrogate failure is structural under subjective-cost heterogeneity

A central takeaway is that statistical fairness objectives are not merely imperfect proxies for harm-based fairness; under heterogeneous subjective costs, they can be *systematically* misaligned. The representation perspective clarifies why: for a statistical constraint to preserve harm-based axioms, it must implicitly collapse heterogeneous error costs into a common unit (Theorem 7.1). When this collapse fails, one can construct settings in which a policy satisfies the surrogate notion while violating harm-based axioms by a large margin (Theorem 7.2).

Experiment 1 (Section 8.1) provides a concrete empirical manifestation of this phenomenon. Under homogeneous costs, the *cost-stratum envy* proxy is mechanically close to zero because the cost-stratum axis $\log(c^{FP}/c^{FN})$ collapses: all individuals share the same FP–FN preference ratio. Under strong heterogeneity, however, targeting a surrogate metric decouples from harm-based envy-freeness. DP-targeting tends to reduce the envy gap relative to the BASE threshold, whereas EO-targeting more often *increases* cost-stratum envy, sometimes substantially, despite its goal of aligning conditional error rates. Calibration-targeting exhibits smaller and more mixed effects, often remaining close to BASE. This pattern is consistent with the harm-allocation view: equalizing *rates* does not equalize *harm* when individuals attach different costs to the same FP/FN events, and finer, more tail-sensitive envy diagnostics can make this misalignment more visible.

9.2 Bounded rationality turns "fairness" into an approximation-floor problem

A second key insight is that, in realistic mediated settings, the relevant question is not whether a harm-allocation axiom can be satisfied exactly, but what *residual violation* remains once one accounts for (i) imperfect information about subjective costs and (ii) restricted policy expressiveness. The bounded-rationality formulation (Section 6) makes this explicit by casting harm-based fairness as a robust objective over a cost-uncertainty set and a restricted policy family:

$$\delta^\star \;=\; \inf_{\pi\in\Pi_{\text{simple}}} \;\sup_{c\in\mathcal{C}(\hat{c},\epsilon)} V(\pi,c),$$

where $V(\pi,c)$ is an axiom-violation score (e.g., an envy gap or proportionality excess). On this view, fairness becomes an approximation problem: even the best implementable policy typically attains $V > 0$, and the irreducible component becomes the quantity of interest.

What the lower bound means. Theorem 7.4 isolates one mechanism behind such irreducible residuals. In a minimal construction with cost uncertainty of radius ϵ and a restricted policy class (e.g., a coupled threshold rule), the minimax residual is bounded away from zero with tight scaling $\delta^\star \geq f(\epsilon)$ (in our construction, $f(\epsilon) = \epsilon$). The point is not that the toy instance is realistic, but that it cleanly demonstrates how *mediated choice* together with *imperfect value information* can impose a nonzero "price" on axiom satisfaction even when the mediator optimizes.

How Experiment 2 decomposes the floor. Experiment 2 (Section 8.2) exhibits the same phenomenon empirically and clarifies that a single reported δ value can combine two distinct sources of residual unfairness:

- **Policy-class floor (expressiveness).** Even with oracle costs, that is, with no informational error, realized cost-stratum envy remains substantially above zero for simple threshold-based policy classes. This reflects an expressiveness barrier: within Π_{simple}, envy-freeness is generically unattainable under strong heterogeneity, so $\inf_{\pi \in \Pi_{\text{simple}}} V(\pi, c)$ is itself nonzero.
- **Information floor (selection under uncertainty).** When costs are observed only through noisy proxies $\hat{c}$ (and possibly coarse cost strata), the mediator selects a policy using $V(\pi, \hat{c})$ on validation rather than $V(\pi, c)$. As a result, the chosen policy can reduce proxy envy while increasing true-cost envy on test (positive Δ), revealing an additional residual induced by value mismeasurement and robust selection. Coarsening to a small number of cost strata can sometimes stabilize selection by reducing estimation variance, but it cannot remove the expressiveness barrier and therefore does not eliminate the overall floor.

In short, mediated evaluations should interpret observed δ as the combination of a *policy-class floor*, reflecting what is unattainable even with perfect information, and an *information floor*, reflecting the additional residual induced by noisy or coarse value information. This decomposition also helps explain why "better fairness" need not be monotone in surrogate optimization: under bounded rationality, fairness progress is constrained by the attainable region jointly determined by Π_{simple} and the informativeness of the mediator's value observations.

9.3 Statistical notions induce distinct, and sometimes conflicting, harm profiles

Beyond the existence of surrogate failure, our results support a sharper reading: different statistical notions do not merely provide noisy approximations to a single underlying ideal, but instead tend to induce distinct *harm profiles*, that is, systematic patterns in how harm is distributed across individuals and cost strata. In this subsection, we use harm diagnostics to *describe* these patterns and to clarify why disputes over "the right" statistical definition can persist even when parties agree on the empirical facts.

Ethical lenses as an interpretive guide (not a formal equivalence). A useful way to interpret harm diagnostics is through familiar ethical vocabulary. We emphasize that the correspondences below are *heuristic* and serve only as an organizing perspective; they are not claims that our metrics formally represent any complete moral theory. With that caveat, our diagnostics naturally separate three families of concerns:

- **Aggregate burden (mean harm).** Mean harm summarizes the overall burden induced by a policy and tracks an efficiency-oriented concern with total or average harm.
- **Concentrated burden (max harm, tail excess).** Max harm and tail excess emphasize protection against concentrated burdens, aligning with maximin or prioritarian intuitions in the limited sense that they treat the worst-off or heavily harmed individuals as morally salient.[3]
- **Relational grievance (envy gap).** Envy is conceptually defined by pairwise comparisons under each individual's own subjective costs: it asks whether, when evaluated through their own valuation, someone would prefer another person's outcome. In our experiments, we operationalize this idea using the cost-stratum envy proxy (Section 5), which captures *cross-stratum* comparative grievance rather than enumerating all individual pairs. In this sense, envy-style summaries can be read as a concrete diagnostic of *comparative grievance*: who has reason to complain under the specified harm model and the necessarily coarse information used to evaluate it.

Axiomatic profiling: notion-targeted policies exhibit non-monotone movements in harm space. Experiment 3 (Section 8.3) provides an "axiomatic profile" of policies that target CAL, DP, and EO relative to BASE under fixed strong heterogeneity. The qualitative pattern is consistent across tasks and models: improvements in a surrogate metric do not reliably predict improvements in harm-based criteria. A single

[3]We invoke *maximin* in Rawls's sense and Parfit's *Priority View* only as loose ethical lenses that assign moral salience to protecting the worst-off or heavily harmed individuals (e.g., through max harm and tail excess), rather than as a textual claim about Rawls's theory of the basic structure or a commitment to any particular social-welfare functional.

intervention can move aggregate burden (mean harm), concentration protections (max harm or tail excess), and relational grievance (envy) in different directions. Using a finer-grained partition into 80 cost strata, the envy diagnostic becomes more sensitive to tail and worst-stratum behavior, so small changes in which errors fall on which cost strata can translate into visibly different relational-grievance outcomes even when accuracy and aggregate harm summaries change only mildly.

For example, CAL can reduce, or leave largely unchanged, calibration error (ECE) while having small, mixed, or even adverse effects on envy and tail-oriented harm summaries depending on the task–model pair. DP reduces the DP gap by construction, yet it can either increase or decrease envy across settings, reflecting the fact that selection-rate equalization does not control how FP/FN incidents are redistributed across heterogeneous-cost strata near the decision boundary. Conversely, EO can substantially reduce envy in some settings (e.g., Inc-LR) while worsening other surrogate quantities such as the DP gap, and it can also increase envy in other task–model pairs. These patterns follow from the structural mismatch: statistical constraints regulate coarse, group-aggregated properties of the confusion profile, whereas harm diagnostics evaluate the *cost-weighted* distribution of errors across individuals, including who absorbs tail risk and who experiences cross-stratum comparative grievances.

Plural distributive logics: what different statistical notions operationally foreground. This motivates a pluralist profiling perspective: each statistical notion operationally foregrounds some classes of complaints while treating others as secondary when selecting a policy. We do *not* claim a one-to-one correspondence between DP, EO, or calibration and any particular moral theory. Rather, as *operational criteria*, these notions tend to privilege different distributive emphases in harm space.[4]

- **Demographic Parity (outcome-rate parity).** DP centers parity in acceptance or positive-prediction rates, closely related to disparate-impact-style concerns. In harm space, DP tends to privilege equalization of *selection* rates; under base-rate differences and cost heterogeneity, this can reallocate FP/FN incidents in ways that sometimes worsen tail concentration or amplify envy for specific cost strata.
- **Equalized Odds (conditional error-rate parity).** EO centers parity in error exposure conditional on ground truth by equalizing TPR and FPR across groups. In harm space, EO can mitigate certain comparative-unfairness patterns, reducing envy in some settings, even when it does not equalize selection rates; conversely, it can trade off against outcome-rate parity (the DP gap) or aggregate burden (mean harm), depending on the score, label, and cost structure.
- **Calibration (score semantics and epistemic validity).** Calibration prioritizes the requirement that predicted probabilities carry consistent meaning within each group, thereby supporting interpretability and decision-support validity. However, calibration is largely orthogonal to how errors are *allocated*: it constrains the mapping from scores to empirical frequencies within groups, but does not directly constrain selection rates or conditional error rates, nor does it control how FP/FN incidents interact with heterogeneous costs. As a result, a policy can substantially improve ECE while leaving max harm, tail concentration, and envy only weakly constrained.

Why definition debates persist, and what profiling makes explicit. Seen in this way, fairness debates that focus exclusively on choosing a statistical definition can become circular for a structural reason: different notions elevate different classes of complaints—outcome disparities (DP), conditional error disparities (EO), or within-group score semantics (calibration). Adopting DP, EO, or calibration as a default is therefore not a purely technical move; it implicitly prioritizes one set of complaints over others.

Under cost heterogeneity, these priorities need not be mutually compatible, and classic incompatibility phenomena can therefore be interpreted as clashes among harm-allocation priorities. Our profiling results make these implicit emphases visible through concrete harm diagnostics, clarifying why the same intervention

[4]This pluralist reading resonates with Walzer's "spheres of justice," which argues that distinct social goods call for distinct distributive principles and cautions against treating any single principle as universally appropriate (Walzer, 2008). We invoke Walzer only as an organizing intuition about plural normative logics, not as a claim of one-to-one equivalence: in our setting, different *statistical* notions function as distinct operational logics in harm space.

can appear beneficial under one statistical lens while simultaneously worsening other ethically salient aspects of harm allocation. In the next subsection, we show that this non-dominance is not peculiar to a single intervention class: it persists even when one moves to diverse pre-, in-, and post-processing algorithms (Experiment 4).

9.4 Algorithmic diversity does not remove the structural trade-offs

A natural objection is that surrogate failure might be an artifact of the particular intervention family used in Experiments 1–3, namely simple post-processing of a fixed score. Experiment 4 (Section 8.4) addresses this *implementation-level* concern by replacing that setup with a diverse set of fairness algorithms spanning pre-, in-, and post-processing. Despite this algorithmic diversity, the qualitative picture remains unchanged: predictive performance differs only modestly across methods, whereas harm-based outcomes vary substantially, and improvements in any single statistical metric do not uniquely determine the induced harm allocation.

The results exhibit persistent *non-dominance* across surrogate and harm criteria. Using the finer-grained 80-stratum envy measure, envy becomes a more tail- and worst-stratum-sensitive diagnostic, and the separation between rate-level surrogates and relational-grievance outcomes becomes especially vivid. For example, on Emp, FairBatch achieves the smallest DP gap and also yields the smallest mean harm and tail excess, yet it does *not* minimize envy; several other methods (e.g., FairBalance and FairProjection) attain lower envy despite having larger DP gaps. On Inc, FairProjection attains the lowest mean harm and tail excess while incurring a very large DP gap, whereas the lowest envy is achieved by FairBalance; in the same setting, methods with improved calibration (lower ECE) need not exhibit lower envy. On Cov, FairMixup yields comparatively low envy while producing the worst mean harm and tail excess, illustrating a pronounced separation between relational grievance and aggregate or tail burden. Thus, under strong subjective-cost heterogeneity, no single method consistently dominates jointly on surrogate metrics and harm-based axioms.

Taken together, Experiments 3–4 support a structural interpretation. These trade-offs are not peculiar to a particular post-processing rule or algorithmic family, but arise from the mismatch between cost-blind statistical constraints and cost-sensitive harm allocation under heterogeneous valuations. In short, "better" DP, EO, or calibration should be interpreted as movement within a constrained attainable region, rather than as monotone progress toward a single fairness ideal; meaningful justification therefore requires stating which harm-based complaints are being prioritized and under which mediation constraints.

9.5 Implications for evaluation and interpretation of "fair" ML systems

Our results reframe the role of common statistical notions in fairness work. Under heterogeneous subjective costs, DP, EO, and calibration are best understood as *surrogates*: they regulate low-dimensional rate summaries, but they do not, by themselves, determine the resulting harm allocation. Accordingly, a fairness claim is incomplete unless it states (i) a *harm model* (what counts as harm and for whom) and (ii) the *mediation constraints* (what the decision-maker can observe and implement). We translate the paper's theorems and experiments into a practical evaluation interface: what to report, how to interpret "improvements" under bounded rationality, and how to connect diagnosis to a transparent choice rule.

A harm-allocation reporting interface. Any surrogate-based report should pair the targeted surrogate metric(s) $S(\pi)$ (e.g., DP/EO gaps or calibration error) with a small set of harm-facing summaries derived from the induced harm vector $H(\pi)$:

(i) **Relational exposure:** an envy-style statistic (e.g., a maximal or average envy gap) capturing whether some individuals, evaluated in their own cost units, would prefer others' outcomes.

(ii) **Concentrated burden:** a tail or worst-case statistic (e.g., tail excess or a max-harm bound) capturing whether harm is heavily concentrated on a minority.

(iii) **Overall burden:** an aggregate statistic (e.g., mean harm) capturing the total burden induced by π.

These three quantities provide a compact diagnostic summary of harm allocation. Across our experiments, improvements in a surrogate do not reliably improve all three quantities and can worsen some under strong heterogeneity; reporting them therefore prevents surrogate progress from being read as a blanket moral guarantee.

Bounded rationality as an attainable region. Under bounded rationality, deployment is constrained in two ways: (a) subjective costs are observed only through imperfect or coarse information, and (b) policies are selected from a restricted class Π_{simple}. As a result, interventions are better interpreted as movements within an attainable set rather than as convergence to a single ideal. Concretely, for a surrogate score S, a violation score $V(\pi, c)$ (e.g., an envy gap or tail excess), a policy class Π_{simple}, and a cost-uncertainty set $\mathcal{C}$, the relevant object is

$$\mathcal{A} \;=\; \{(S(\pi, c), V(\pi, c), \mathrm{Perf}(\pi)) : \; \pi \in \Pi_{\text{simple}}, \; c \in \mathcal{C}\}.$$

This viewpoint explains non-monotonicity: different methods can move to different parts of $\mathcal{A}$, improving S while worsening V, or vice versa. Experiment 4 shows that such non-dominance persists across diverse algorithmic families, consistent with a structural source of misalignment rooted in heterogeneous values and mediation constraints.

From diagnosis to choice: making priorities explicit. Diagnostics matter only insofar as they support a deployment decision. We therefore advocate stating an explicit choice rule that makes normative priorities legible. One natural formulation is a constrained minimax objective,

$$\min_{\pi \in \Pi_{\text{simple}}} \; \sup_{c \in \mathcal{C}} V(\pi, c) \quad \text{s.t.} \quad \mathrm{Perf}(\pi) \geq \tau,$$

which selects, among implementable policies, the one that minimizes worst-case axiom violation under cost uncertainty while satisfying a performance floor. Alternatively, stakeholders may prefer a lexicographic rule that first limits concentrated harm (e.g., tail excess) and then optimizes secondary criteria such as mean harm or the surrogate score. In either case, the key benefit is interpretability: the decision-maker commits to a protection statement (e.g., "limit envy exposure" or "avoid extreme harm") rather than treating parity in a rate metric as a self-sufficient moral warrant. This captures the paper's central message: surrogates are useful *constraints* and *instruments*, but fairness justification ultimately requires explicit commitments about harm and the mediation setting.

9.6 Limitations and directions suggested by the harm-allocation view

The present study adopts a simple harm model based on linear FP/FN costs and emphasizes threshold-style policy families in order to make the surrogate-failure and approximation-floor phenomena transparent. These choices are deliberate, but they also impose limitations and point to several concrete directions for future work.

Subjective costs as inputs: from point estimates to uncertainty and coarse stratification. A central practical concern is that individual FP/FN costs are rarely observed directly. However, the harm-allocation framework does not require precise point identification of c_i in order to be meaningful. Instead, costs can be treated as *imperfectly observed inputs* in two increasingly realistic forms: (i) *stratum-level costs*, in which individuals are mapped to coarse stakeholder strata with shared or interval-valued costs, and (ii) *uncertainty sets* $\mathcal{C}(\hat{c}, \epsilon)$ around estimated costs. This perspective directly matches the mediator model studied in our bounded-rationality analysis: even with imperfect cost information, one can, and should, report harm-based diagnostics under plausible cost uncertainty and quantify the resulting approximation floors.

Value elicitation is part of fairness, not an external prerequisite. The results suggest that progress on fairness depends as much on representing and eliciting stakeholder values, and the mediator's uncertainty about them, as on optimizing any single statistical criterion. From this perspective, "measuring costs" is not a separate engineering problem that must be solved before fairness evaluation begins; it is an integral part of fairness specification. Accordingly, a practically grounded evaluation can proceed by (a) stating a small

set of cost scenarios, or bounds, that reflect stakeholder perspectives and (b) reporting whether conclusions are robust across these scenarios. This reframes the common critique of subjectivity: the alternative is not value-free fairness, but rather unreported implicit values.

Richer harm models and broader settings. Future work can extend beyond linear FP/FN costs to incorporate context-dependent costs, multidimensional harms beyond misclassification (e.g., burdens of scrutiny, stigma, or delay), and population-level externalities. Such extensions may sharpen the normative analysis, but they also raise additional identifiability and comparability challenges, which further motivate robust, uncertainty-aware evaluation rather than single-number fairness claims. Likewise, extending beyond binary sensitive attributes and exploring intersectional groupings would provide a more demanding test of surrogate-failure claims and would better reflect real deployment settings.

Overall, the harm-allocation view implies a shift in what counts as evidence for fairness: statistical objectives alone are insufficient to guarantee axiomatic fairness of the resulting harm allocation unless value heterogeneity and mediation constraints are addressed explicitly.

10 Conclusion

We have reframed ML fairness as the allocation of harm under heterogeneous error costs and evaluated policies through fair-division axioms, treating Demographic Parity, Equalized Odds, and calibration as surrogates rather than as fairness itself. We identified the conditions under which these statistical criteria can serve as valid surrogates, showing that such a role requires strong hidden restrictions on costs, and we provided counterexamples together with an axiom-conflict interpretation that explains why systematic failures emerge under cost heterogeneity. We further showed that under bounded rationality and intermediation—when costs are observed only noisily or coarsely and policies are restricted to simple classes—there is an irreducible floor on approximate axiom satisfaction. Our experiments mirror these theoretical results by tracing surrogate failure, approximation floors, and the harm-allocation profiles induced by common fairness objectives and algorithms. Taken together, these findings clarify what rate-based fairness criteria can and cannot guarantee, and they support evaluating deployed systems through more explicit modeling and reporting of harm allocations.

Broader Impact Statement

This work is primarily conceptual and diagnostic. It reframes ML fairness as the allocation of harms under heterogeneous subjective costs and shows that satisfying statistical criteria (e.g., DP, EO, or calibration) need not guarantee fair harm allocations when value heterogeneity and intermediary constraints, such as limited information and restricted policy classes, are present. In high-stakes decision settings, including access to credit, employment, insurance, or public benefits, errors can generate substantial and uneven burdens. Making the allocation target explicit—that is, clarifying who bears which burdens—can improve the transparency of fairness audits and reduce overreliance on any single statistical notion.

In a positive sense, our framework encourages more transparent governance by (i) clarifying what is being distributed, namely harms arising from errors and interventions; (ii) reporting harm-based violations alongside standard statistical metrics; and (iii) treating statistical notions as surrogates whose validity depends on assumptions about values and institutional constraints. This perspective can help practitioners identify cases in which a "fairness fix" in metric space merely shifts burdens in ways that are ethically significant but invisible to group-level parity metrics.

Potential risks include misuse of the framework as a blanket argument against fairness constraints (e.g., "fairness is impossible") and *value laundering*, whereby convenient cost assumptions are selected to rationalize preferred outcomes. A further risk is that eliciting or estimating costs, including preferences, vulnerability, or downstream impacts, may itself introduce harms, such as privacy intrusion, representational bias in whose values are measured and how, or additional burdens on already disadvantaged stakeholders. These risks are especially salient when cost models are treated as authoritative rather than uncertain and contestable.

We mitigate these risks by emphasizing transparency and robustness rather than any single definitive notion of fairness. Conclusions should be evaluated across plausible cost ranges or uncertainty sets and across feasible policy classes, and they should not be presented as uniquely determined by any single specification. When costs are estimated, they should be treated as noisy and contestable inputs, accompanied by sensitivity analyses and clearly documented elicitation or measurement procedures, and complemented by human oversight and stakeholder consultation. Our experiments are lightweight and do not materially increase computational or environmental burden.

Reproducibility

All experiments in Section 8, as well as all tables reported in the paper, can be reproduced by running the single notebook `Price_of_Justice.ipynb` from start to finish. The notebook downloads and preprocesses the public Folktables ACS 2018 1-Year person survey, executes Experiments 1–4 as described in the paper—including model training, policy and algorithm construction, and evaluation—and reproduces the seed-averaged results in publication-style tabular form. It is designed to run both on Google Colab and in a local environment. Using the same random seeds and package versions yields identical tables, although minor numerical differences may arise across platforms or library versions because of nondeterminism in underlying numerical routines.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International conference on machine learning*, pp. 120–129. PMLR, 2019.

Georgios Amanatidis, Georgios Birmpas, Aris Filos-Ratsikas, and Alexandros A Voudouris. Fair division of indivisible goods: A survey. *arXiv preprint arXiv:2202.07551*, 2022.

Georgios Amanatidis, Haris Aziz, Georgios Birmpas, Aris Filos-Ratsikas, Bo Li, Hervé Moulin, Alexandros A Voudouris, and Xiaowei Wu. Fair division of indivisible goods: Recent progress and open questions. *Artificial Intelligence*, 322:103965, 2023.

Haris Aziz, Ioannis Caragiannis, Ayumi Igarashi, and Toby Walsh. Fair allocation of indivisible goods and chores. *Autonomous Agents and Multi-Agent Systems*, 36(1):3, 2022a.

Haris Aziz, Bo Li, Hervé Moulin, and Xiaowei Wu. Algorithmic fair allocation of indivisible items: A survey and new questions. *ACM SIGecom Exchanges*, 20(1):24–40, 2022b.

Yakov Babichenko, Michal Feldman, Ron Holzman, and Vishnu V Narayan. Fair division via quantile shares. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 1235–1246, 2024.

Artem Baklanov, Pranav Garimidi, Vasilis Gkatzelis, and Daniel Schoepflin. Achieving proportionality up to the maximin item with indivisible goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5143–5150, 2021.

Siddharth Barman and Arpita Biswas. Fair division under cardinality constraints. *arXiv preprint arXiv:1804.09521*, 2018.

Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. Fairness and welfare quantification for regret in multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6762–6769, 2023.

Xiaohui Bei, Zihao Li, Jinyan Liu, Shengxin Liu, and Xinhang Lu. Fair division of mixed divisible and indivisible goods. *Artificial Intelligence*, 293:103436, 2021a.

Xiaohui Bei, Shengxin Liu, Xinhang Lu, and Hongao Wang. Maximin fairness with mixed divisible and indivisible goods. *Autonomous Agents and Multi-Agent Systems*, 35(2):34, 2021b.

Henry C Bendekgey and Erik Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. *Advances in Neural Information Processing Systems*, 34:30023–30036, 2021.

Anna Bogomolnaia and Hervé Moulin. A new solution to the random assignment problem. *Journal of Economic theory*, 100(2):295–328, 2001.

Sylvain Bouveret and Michel Lemaître. Characterizing conflicts in fair division of indivisible goods using a scale of criteria. *Autonomous Agents and Multi-Agent Systems*, 30(2):259–290, 2016.

Sylvain Bouveret, Michel Lemaître, Hélene Fargier, and Jérôme Lang. Allocation of indivisible goods: a general model and some complexity results. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 1309–1310, 2005.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International conference on artificial intelligence and statistics*, pp. 6045–6061. PMLR, 2022.

Eric Budish. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy*, 119(6):1061–1103, 2011.

Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 309–318, 2019.

Ioannis Caragiannis, Nick Gravin, and Xin Huang. Envy-freeness up to any item with high nash welfare: The virtue of donating items. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 527–545, 2019a.

Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum nash welfare. *ACM Transactions on Economics and Computation (TEAC)*, 7(3):1–32, 2019b.

Ioannis Caragiannis, Jugal Garg, Nidhi Rathi, Eklavya Sharma, and Giovanna Varricchio. New fairness concepts for allocating indivisible items. International Joint Conferences on Artificial Intelligence Organization, 2023.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1–38, 2024.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pp. 1349–1361. PMLR, 2021.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *International Conference on Machine Learning*, pp. 2972–3005. PMLR, 2022.

Qinyi Chen, Jason Cheuk Nam Liang, Negin Golrezaei, and Djallel Bouneffouf. Interpolating item and user fairness in multi-sided recommendations. *Advances in Neural Information Processing Systems*, 37: 50189–50229, 2024.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on fairness, accountability and transparency*, pp. 134–148. PMLR, 2018.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

Richard Cole and Vasilis Gkatzelis. Approximating the nash social welfare with indivisible items. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 371–380, 2015.

Vincent Conitzer, Rupert Freeman, and Nisarg Shah. Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 629–646, 2017.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pp. 1397–1405. PMLR, 2019.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. *Advances in Neural Information Processing Systems*, 34:16610–16621, 2021.

Zhun Deng, He Sun, Zhiwei Steven Wu, Linjun Zhang, and David C Parkes. Reinforcement learning with stepwise fairness constraints. *arXiv preprint arXiv:2211.03994*, 2022.

Siddartha Devic, David Kempe, Vatsal Sharan, and Aleksandra Korolova. Fairness in matching under uncertainty. In *International Conference on Machine Learning*, pp. 7775–7794. PMLR, 2023.

Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via lorenz dominance. *Advances in Neural Information Processing Systems*, 34:8596–8608, 2021.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 389–399, 2023.

Brandon Fain, Kamesh Munagala, and Nisarg Shah. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 575–592, 2018.

Alireza Farhadi, Mohammad Ghodsi, Mohammad Taghi Hajiaghayi, Sebastien Lahaie, David Pennock, Masoud Seddighin, Saeed Seddighin, and Hadi Yami. Fair allocation of indivisible goods to asymmetric agents. *Journal of Artificial Intelligence Research*, 64:1–20, 2019.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Hidde Fokkema, Damien Garreau, and Tim van Erven. The risks of recourse in binary classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 550–558. PMLR, 2024.

Jugal Garg and Setareh Taki. An improved approximation algorithm for maximin shares. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 379–380, 2020.

Jugal Garg, Pooja Kulkarni, and Aniket Murhekar. On fair and efficient allocations of indivisible public goods. *arXiv preprint arXiv:2107.09871*, 2021.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 445–453, 2021.

Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 316–324, 2022.

Marah Ghoummaid and Uri Shalit. When to act and when to ask: policy learning with deferral under hidden confounding. *Advances in Neural Information Processing Systems*, 37:56108–56135, 2024.

Gianluigi Greco and Francesco Scarcello. The complexity of computing maximin share allocations on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2006–2013, 2020.

Sophie Greenwood, Sudalakshmee Chiniah, and Nikhil Garg. User-item fairness tradeoffs in recommendations. *Advances in Neural Information Processing Systems*, 37:114236–114288, 2024.

Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 688–704, 2023.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in neural information processing systems*, 31, 2018.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017, 2021.

Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: Who's the fairest of them all? *Advances in Neural Information Processing Systems*, 35:32749–32761, 2022.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 535–545, 2020.

Taeuk Jang, Hongchang Gao, Pengyi Shi, and Xiaoqian Wang. Achieving fairness through separability: a unified framework for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 28–36. PMLR, 2024.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Yasushi Kawase, Hanna Sumita, and Yu Yokoi. Random assignment of indivisible goods under constraints. *arXiv preprint arXiv:2208.07666*, 2022.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287. PMLR, 2020.

Michael P Kim, Aleksandra Korolova, Guy N Rothblum, and Gal Yona. Preference-informed fairness. *arXiv preprint arXiv:1904.01793*, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

David Kurokawa, Ariel Procaccia, and Junxing Wang. When can the maximin share guarantee be guaranteed? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

David Kurokawa, Ariel D Procaccia, and Junxing Wang. Fair enough: Guaranteeing approximate maximin shares. *Journal of the ACM (JACM)*, 65(2):1–27, 2018.

Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International conference on machine learning*, pp. 12917–12930. PMLR, 2022.

Zhentao Li and Adrian Vetta. The fair division of hereditary set systems. *ACM Transactions on Economics and Computation (TEAC)*, 9(2):1–19, 2021.

Annie Liang and Jay Lu. Algorithmic fairness and social welfare. In *AEA Papers and Proceedings*, volume 114, pp. 628–632. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

Richard J Lipton, Evangelos Markakis, Elchanan Mossel, and Amin Saberi. On approximately fair allocations of indivisible goods. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pp. 125–131, 2004.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Lydia T Liu, Nikhil Garg, and Christian Borgs. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2489–2518. PMLR, 2022.

Elita Lobo, Justin Payan, Cyrus Cousins, and Yair Zick. Fair and welfare-efficient constrained multi-matchings under uncertainty. *Advances in Neural Information Processing Systems*, 37:74579–74616, 2024.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018a.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018b.

Anay Mehrotra and Nisheeth Vishnoi. Fair ranking with noisy protected attributes. *Advances in Neural Information Processing Systems*, 35:31711–31725, 2022.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 230–239, 2019.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*, pp. 10520–10545. PMLR, 2023.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International conference on machine learning*, pp. 7097–7107. PMLR, 2020.

Hoon Oh, Ariel D Procaccia, and Warut Suksompong. Fairly allocating many goods with few queries. *SIAM Journal on Discrete Mathematics*, 35(2):788–813, 2021.

Kanad Pardeshi, Itai Shapira, Ariel D Procaccia, and Aarti Singh. Learning social welfare functions. *Advances in Neural Information Processing Systems*, 37:41733–41766, 2024.

Nicholas Perello, Cyrus Cousins, Yair Zick, and Przemyslaw Grabowicz. Discrimination induced by algorithmic recourse objectives. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1653–1663, 2025.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Miriam Rateike, Isabel Valera, and Patrick Forré. Designing long-term group fair policies in dynamical systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 20–50, 2024.

Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. Modeling assumptions clash with the real world: Transparency, equity, and community challenges for student assignment algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ayush Sawarni, Soumyabrata Pal, and Siddharth Barman. Nash regret guarantees for linear bandits. *Advances in Neural Information Processing Systems*, 36:33288–33318, 2023.

Erel Segal-Halevi. Competitive equilibrium for almost all incomes: existence and fairness. *Autonomous agents and multi-agent systems*, 34(1):26, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pp. 8905–8915. PMLR, 2020.

Rachael Hwee Ling Sim, Yehong Zhang, Bryan Kian Hsiang Low, and Patrick Jaillet. Collaborative bayesian optimization with fair regret. In *International Conference on Machine Learning*, pp. 9691–9701. PMLR, 2021.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2219–2228, 2018.

Seamus Somerstep, Ya'acov Ritov, and Yuekai Sun. Algorithmic fairness in performative policy learning: Escaping the impossibility of group fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 616–630, 2024.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2239–2248, 2018.

Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.

Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023.

Muhammad Faaiz Taufiq, Jean-Francois Ton, and Yang Liu. Achievable fairness on your data with utility guarantees. *Advances in Neural Information Processing Systems*, 37:140405–140450, 2024.

Julius Von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9584–9594, 2022.

Michael Walzer. *Spheres of justice: A defense of pluralism and equality*. Basic books, 2008.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pp. 10686–10696. PMLR, 2021.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International conference on machine learning*, pp. 11035–11046. PMLR, 2021.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International conference on machine learning*, pp. 6786–6797. PMLR, 2019.

Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*, pp. 703–714, 2022a.

Jimmy Wu, Yatong Chen, and Yang Liu. Metric-fair classifier derandomization. In *International Conference on Machine Learning*, pp. 23999–24016. PMLR, 2022b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Mengxiao Zhang, Ramiro Deo-Campo Vuong, and Haipeng Luo. No-regret learning for fair multi-agent social welfare optimization. *Advances in Neural Information Processing Systems*, 37:57671–57700, 2024.

Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in neural information processing systems*, 32, 2019.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019.

A Full Proofs for the Main Results

This appendix collects complete proofs of the main theoretical results. To keep the exposition readable, we first fix notation and isolate reusable technical ingredients as short lemmas; the main proofs then proceed by invoking these lemmas where appropriate.

A.1 Notation Used Throughout the Appendix

For an individual i, write

$$p_i^{FN} := \Pr_\pi(\mathrm{FN} \mid i), \qquad p_i^{FP} := \Pr_\pi(\mathrm{FP} \mid i),$$

and define the (subjective) harm to individual i from receiving ℓ's error lottery as

$$H_{i \leftarrow \ell}(\pi) := c_i^{FP} p_\ell^{FP} + c_i^{FN} p_\ell^{FN}.$$

The directed envy gap is

$$\mathrm{EnvyGap}_{i,j}(\pi) := \big[H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi)\big]_+.$$

Any aggregate envy statistic used in the main paper is obtained by aggregating these pairwise quantities, for example by taking maxima or averages over pairs.

A.2 Theorem 7.1: Representation Conditions

Before proving the theorem, we restate its main message informally. Statistical fairness constraints (DP, EO, and calibration) depend on $(\mathcal{D}, \pi)$ through the induced distribution of $(A, Y, \hat{Y})$, and in the case of calibration, possibly also through S, but they do not depend on the subjective cost profile c. Consequently, if two individuals face different error lotteries under π, then one can alter the cost profile, without changing the statistical metric values, in a way that makes harm-based axiom violations arbitrarily large. Any representation result must therefore impose additional structure on costs (e.g., collapse or alignment conditions) if statistical constraints are to function as reliable surrogates.

We begin with a lemma that captures the amplification mechanism.

Lemma A.1 (Cost-blind metrics and envy amplification)**.** *Fix $(\mathcal{D}, \pi)$. Let $M(\pi)$ be any statistical constraint value defined as a functional of the joint law of $(A, Y, \hat{Y})$ (and possibly (A, Y, S)). Then $M(\pi)$ is invariant to changes in the cost profile c.*

Moreover, if there exist $i \neq j$ such that

$$p_i^{FN} \neq p_j^{FN} \quad \text{or} \quad p_i^{FP} \neq p_j^{FP},$$

then, for any $B > 0$, there exists a cost profile c such that $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$ while $M(\pi)$ remains unchanged.

Proof. The invariance claim is immediate: $M(\pi)$ depends only on the induced distribution of predictions, scores, and labels, and not on subjective costs.

For the amplification claim, assume without loss of generality that $p_i^{FN} - p_j^{FN} = \Delta^{FN} > 0$; the other cases are symmetric. Choose costs

$$c_i^{FN} = C, \quad c_i^{FP} = 0, \qquad \text{and} \qquad c_\ell^{FN} = c_\ell^{FP} = 0 \ \text{ for all } \ell \neq i.$$

Then

$$H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi) = c_i^{FN}(p_i^{FN} - p_j^{FN}) = C\,\Delta^{FN},$$

so $\mathrm{EnvyGap}_{i,j}(\pi) = C\,\Delta^{FN}$. Taking $C := B/\Delta^{FN}$ yields $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$, while, by invariance, $M(\pi)$ is unchanged. □

Proof. The theorem follows by contraposition using Lemma A.1. Suppose the admissible class of instances allows both (i) cost heterogeneity that can scale a coordinate of c_i without affecting $(\mathcal{D}, \pi)$, and (ii) policies π that induce heterogeneous error lotteries across individuals or within groups. Then statistical constraints cannot uniformly control harm-based axiom violations: whenever heterogeneity in (p_i^{FN}, p_i^{FP}) is present, Lemma A.1 yields a cost profile that leaves $M(\pi)$ unchanged while making envy arbitrarily large.

Therefore, if a statistical constraint is to serve as a surrogate for harm-axiom fairness *uniformly* over an instance class, that class must rule out this amplification channel. This requires either extremely strong restrictions on π (e.g., equalizing error lotteries at the relevant granularity) or, more plausibly, structural restrictions that collapse or align subjective costs so that scaling a private value dimension is no longer admissible. □

A.3 Theorem 7.2: Counterexample Under Heterogeneous Costs

We now provide an explicit construction. We choose $(\mathcal{D}, \pi)$ so that a selected statistical metric family is satisfied (in some cases vacuously), and then choose heterogeneous costs that induce arbitrarily large harm-axiom violations.

Proof. Fix a tolerance $\eta \geq 0$ and a metric family $M \in \{\mathrm{DP}, \mathrm{EO/EOp}, \mathrm{Calibration}\}$.

Construction of $(\mathcal{D}, \pi)$. Consider a single-group setting, so that A is constant. Take two individuals $i \neq j$ with $Y_i = Y_j = 1$. Define a randomized policy π such that

$$p_i^{FN} = p, \qquad p_j^{FN} = q, \qquad p > q,$$

and, since they play no role here, set $p_i^{FP} = p_j^{FP} = 0$.

Verification that $M(\pi) \leq \eta$. In a single-group setting, DP and EO-type across-group constraints are vacuous and hence hold exactly. For calibration, take a degenerate score $S \equiv 1$, so that $\Pr(Y = 1 \mid S = 1) = 1$ and groupwise calibration also holds. Thus, in all three cases, $M(\pi) = 0 \leq \eta$.

Choice of costs inducing large envy. Let

$$c_i^{FN} = C, \quad c_i^{FP} = 0, \qquad c_j^{FN} = 1, \quad c_j^{FP} = 0,$$

with $C > 0$ to be chosen. Then

$$\mathrm{EnvyGap}_{i,j}(\pi) = \left[c_i^{FN}(p_i^{FN} - p_j^{FN})\right]_+ = C(p - q).$$

Given any $B > 0$, choosing $C := B/(p - q)$ yields $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$, and hence $\mathrm{Envy}^{\max}(\pi) \geq B$, while $M(\pi) \leq \eta$ continues to hold. □

A.4 Proposition 7.3: Statistical Impossibility as Axiom Conflict

We present the standard incompatibility between calibration and equalized odds under unequal base rates in a form that fits our harm-axiom framework. Once one statistical constraint is enforced, the other must fail, except in degenerate regimes, and any resulting nontrivial gap can be converted into harm-axiom violations under admissible subjective values.

We first isolate two routine facts.

Lemma A.2 (Calibration equalizes PPV across groups at a fixed decision level)**.** *Assume a discretized score S and a decision rule $\hat{Y} = \mathbf{1}\{S = u\}$ for some bin u. If S is calibrated within each group, then*

$$\Pr(Y = 1 \mid \hat{Y} = 1, A = a) = \Pr(Y = 1 \mid \hat{Y} = 1, A = b) = u,$$

so the positive predictive value (PPV) is equal across groups at the positive decision level.

Proof. Since $\hat{Y} = 1$ if and only if $S = u$, groupwise calibration gives

$$\Pr(Y = 1 \mid S = u, A = g) = u$$

for each group g, which is exactly the stated PPV equality. □

Lemma A.3 (Under EO and unequal base rates, PPV differs in the nondegenerate regime)**.** *Let $\mu_g := \Pr(Y = 1 \mid A = g)$ denote the base rate in group g. If EO holds with common* $(\mathrm{TPR}, \mathrm{FPR})$ *across groups and $\mu_a \neq \mu_b$, then, whenever* $\mathrm{TPR} \neq \mathrm{FPR}$ *and* $\mathrm{TPR}, \mathrm{FPR} \in (0, 1)$,

$$\Pr(Y = 1 \mid \hat{Y} = 1, A = a) \neq \Pr(Y = 1 \mid \hat{Y} = 1, A = b).$$

Proof. Under EO, Bayes' rule gives, for each g,

$$\mathrm{PPV}_g = \frac{\mathrm{TPR} \cdot \mu_g}{\mathrm{TPR} \cdot \mu_g + \mathrm{FPR} \cdot (1 - \mu_g)}.$$

In the nondegenerate regime, the right-hand side is strictly increasing in μ_g, so unequal base rates imply unequal PPV. □

Proof. If groups have unequal base rates $\mu_a \neq \mu_b$, Lemma A.2 implies that calibration forces PPV to be equal across groups, whereas Lemma A.3 implies that EO forces PPV to differ in the nondegenerate regime. Therefore, calibration and EO cannot simultaneously hold except in degenerate cases, such as perfect prediction or uninformative prediction.

To connect this incompatibility to harm-axiom conflict, observe that leaving the degenerate regime implies that, when calibration is enforced, at least one of the EO equalities must fail by a nonzero gap Δ, or conversely. If, for example, $\mathrm{FNR}_a - \mathrm{FNR}_b = \Delta > 0$, choose individuals i in group a and j in group b with $Y_i = Y_j = 1$ such that $p_i^{FN} - p_j^{FN} = \Delta$. Allowing group-dependent, or otherwise admissible, scaling of subjective costs and setting $c_i^{FN} = C$ and $c_i^{FP} = 0$ yields

$$\mathrm{EnvyGap}_{i,j}(\pi) = C\Delta,$$

which exceeds any fixed tolerance for sufficiently large C. This shows how statistical incompatibility translates into axiomatic conflict in harm allocation once subjective value is admitted. □

A.5 Theorem 7.4: A Lower Bound Under Bounded Rationality

The proof is constructive. We specify an uncertainty set around proxy costs and a restricted policy class that enforces coupled treatment, and then compute the resulting minimax envy floor. For this instance, the minimax value is in fact *tight* and equal to ϵ.

Proof. Let $n = 2$ with labels $y_1 = 1$ and $y_2 = 0$. Consider the restricted policy class Π_{simple} that enforces identical treatment probabilities:

$$\Pr_\pi(\hat{Y} = 1 \mid 1) = \Pr_\pi(\hat{Y} = 1 \mid 2) =: \alpha \in [0, 1].$$

Then the induced error probabilities are

$$p_1^{\mathrm{FN}} = 1 - \alpha, \qquad p_2^{\mathrm{FP}} = \alpha.$$

Assume that the mediator observes proxy costs $\hat{c}_1 = \hat{c}_2 = (1, 1)$ and faces the componentwise uncertainty set

$$\mathcal{C}(\hat{c}, \epsilon) = \Big\{ c : \ c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}} \in [1 - \epsilon, 1 + \epsilon] \text{ for } i = 1, 2 \Big\}, \qquad \epsilon \in (0, 1).$$

For a fixed α, consider the two directed envy gaps:

$$\mathrm{EnvyGap}_{1,2}(\pi; c) = \big[c_1^{\mathrm{FN}}(1 - \alpha) - c_1^{\mathrm{FP}}\alpha \big]_+, \qquad \mathrm{EnvyGap}_{2,1}(\pi; c) = \big[c_2^{\mathrm{FP}}\alpha - c_2^{\mathrm{FN}}(1 - \alpha) \big]_+.$$

Step 1: worst-case directed envy gaps. Because c_1^{FN} and c_1^{FP} vary independently over $[1 - \epsilon, 1 + \epsilon]$, the linear form $c_1^{\mathrm{FN}}(1 - \alpha) - c_1^{\mathrm{FP}}\alpha$ is maximized by setting $c_1^{\mathrm{FN}} = 1 + \epsilon$ and $c_1^{\mathrm{FP}} = 1 - \epsilon$. Hence

$$\sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{EnvyGap}_{1,2}(\pi; c) = \Big[(1 + \epsilon)(1 - \alpha) - (1 - \epsilon)\alpha \Big]_+ = \big[1 + \epsilon - 2\alpha \big]_+.$$

Similarly, the linear form $c_2^{\mathrm{FP}}\alpha - c_2^{\mathrm{FN}}(1 - \alpha)$ is maximized by setting $c_2^{\mathrm{FP}} = 1 + \epsilon$ and $c_2^{\mathrm{FN}} = 1 - \epsilon$, yielding

$$\sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{EnvyGap}_{2,1}(\pi; c) = \Big[(1 + \epsilon)\alpha - (1 - \epsilon)(1 - \alpha) \Big]_+ = \big[2\alpha - (1 - \epsilon) \big]_+.$$

Step 2: minimax maximum envy. Define

$$\mathrm{Envy}^{\max}(\pi; c) = \max\{\mathrm{EnvyGap}_{1,2}(\pi; c),\ \mathrm{EnvyGap}_{2,1}(\pi; c)\}.$$

Since the costs of individuals 1 and 2 can be chosen independently within $\mathcal{C}(\hat{c}, \epsilon)$, it follows that

$$\sup_{c \in \mathcal{C}(\hat{c},\epsilon)} \mathrm{Envy}^{\max}(\pi; c) = \max\Big\{ \big[\, 1 + \epsilon - 2\alpha \,\big]_+,\ \big[\, 2\alpha - (1 - \epsilon) \,\big]_+ \Big\}.$$

Minimizing over $\alpha \in [0, 1]$ balances the two terms. At $\alpha = \frac{1}{2}$, they are equal and both attain the value ϵ. Therefore,

$$\inf_{\pi \in \Pi_{\mathrm{simple}}} \sup_{c \in \mathcal{C}(\hat{c},\epsilon)} \mathrm{Envy}^{\max}(\pi; c) = \inf_{\alpha \in [0,1]} \max\Big\{ \big[\, 1 + \epsilon - 2\alpha \,\big]_+,\ \big[\, 2\alpha - (1 - \epsilon) \,\big]_+ \Big\} = \epsilon.$$

In particular, the minimax residual is bounded below by ϵ, establishing the claimed approximation floor and, for this instance, its tightness. □

B Finite-Cost Lower Bounds

Throughout this section, we assume that subjective costs are bounded:

$$c_i^{FP},\ c_i^{FN} \in [0, C_{\max}] \qquad \text{for all } i.$$

The results below quantify how large harm-axiom violations can be *without* taking $C \to \infty$.

B.1 A Bounded Amplification Lemma

The next lemma is the finite-cost analogue of the amplification argument used in the main proofs. It shows that whenever two individuals face different error lotteries, the resulting envy can still be made large, but now with magnitude scaling linearly in $C_{\max}$.

Lemma B.1 (Bounded-cost amplification of envy)**.** *Fix $(\mathcal{D}, \pi)$ and two individuals $i \neq j$. Define*

$$\Delta_{i,j}^{FN} := p_i^{FN} - p_j^{FN}, \qquad \Delta_{i,j}^{FP} := p_i^{FP} - p_j^{FP}.$$

Under the bounded-cost constraint $c_i^{FP}, c_i^{FN} \in [0, C_{\max}]$,

$$\sup_{c_i^{FP},\, c_i^{FN} \in [0, C_{\max}]} \mathrm{EnvyGap}_{i,j}(\pi) \ = \ C_{\max} \cdot \max\{(\Delta_{i,j}^{FN})_+,\ (\Delta_{i,j}^{FP})_+\}.$$

Consequently,

$$\sup_{c \in [0, C_{\max}]^{2n}} \mathrm{Envy}^{\max}(\pi) \ \geq \ C_{\max} \cdot \max_{i,j} \max\{(\Delta_{i,j}^{FN})_+,\ (\Delta_{i,j}^{FP})_+\}.$$

Proof. We can write

$$\mathrm{EnvyGap}_{i,j}(\pi) = \big[c_i^{FP}(p_i^{FP} - p_j^{FP}) + c_i^{FN}(p_i^{FN} - p_j^{FN}) \big]_+ = \big[c_i^{FP} \Delta_{i,j}^{FP} + c_i^{FN} \Delta_{i,j}^{FN} \big]_+.$$

To obtain the upper bound, note that $c_i^{FP}, c_i^{FN} \leq C_{\max}$, so

$$c_i^{FP} \Delta_{i,j}^{FP} + c_i^{FN} \Delta_{i,j}^{FN} \leq C_{\max} \big((\Delta_{i,j}^{FP})_+ + (\Delta_{i,j}^{FN})_+ \big).$$

Maximizing over the coordinate with the larger positive contribution gives

$$\mathrm{EnvyGap}_{i,j}(\pi) \leq C_{\max} \cdot \max\{(\Delta_{i,j}^{FN})_+, (\Delta_{i,j}^{FP})_+\}.$$

For the matching lower bound, if $(\Delta_{i,j}^{FN})_+ \geq (\Delta_{i,j}^{FP})_+$, choose $(c_i^{FN}, c_i^{FP}) = (C_{\max}, 0)$; otherwise choose $(0, C_{\max})$. In either case,

$$\mathrm{EnvyGap}_{i,j}(\pi) = C_{\max} \cdot \max\{(\Delta_{i,j}^{FN})_+, (\Delta_{i,j}^{FP})_+\}.$$

The second claim follows by taking the maximum over ordered pairs (i, j) and over feasible cost assignments for all individuals. □

A direct implication of Lemma B.1 is that if a policy π induces an error-lottery gap of size δ in either FN or FP for some ordered pair, then bounded costs still permit envy violations of order $C_{\max}\delta$.

B.2 A Bounded-Cost Surrogate-Failure Construction

The next proposition mirrors the counterexample logic used in the main text, but constrains all costs to lie in $[0, C_{\max}]$ and yields a *constant* nonvanishing lower bound by fixing the error gap.

Proposition B.2 (Finite-cost counterexample with nontrivial harm-axiom violation)**.** *Fix any $C_{\max} > 0$ and any tolerance $\eta \geq 0$. For each metric family $M \in \{\mathrm{DP}, \mathrm{EO/EOp}, \mathrm{Calibration}\}$, there exists an instance $(\mathcal{D}, c, \pi)$ such that*

$$M(\pi) \leq \eta, \qquad c_i^{FP}, c_i^{FN} \in [0, C_{\max}] \;\; \textit{for all } i,$$

and yet

$$\mathrm{Envy}^{\max}(\pi) \;\geq\; \frac{C_{\max}}{2}, \qquad \mathrm{Prop}^{\max}(\pi) \;\geq\; \frac{C_{\max}}{4}.$$

Proof. Consider a single-group setting, so that any across-group constraint is vacuous. Take two individuals $i \neq j$ with $Y_i = Y_j = 1$ and define a randomized policy π such that

$$p_i^{FN} = \frac{3}{4}, \qquad p_j^{FN} = \frac{1}{4}, \qquad p_i^{FP} = p_j^{FP} = 0.$$

Then the FN gap is $\Delta_{i,j}^{FN} = 1/2$.

To verify metric satisfaction, note that with a single group, the DP and EO/EOp constraints hold trivially. For calibration, take a degenerate score $S \equiv 1$, so that calibration also holds exactly. Hence $M(\pi) = 0 \leq \eta$.

Now choose bounded costs as

$$c_i^{FN} = C_{\max}, \quad c_i^{FP} = 0, \qquad c_j^{FN} = 0, \quad c_j^{FP} = 0,$$

which clearly satisfy the constraint $[0, C_{\max}]$.

For the envy lower bound,

$$\mathrm{EnvyGap}_{i,j}(\pi) = \big[c_i^{FN}(p_i^{FN} - p_j^{FN})\big]_+ = C_{\max} \cdot \frac{1}{2},$$

so $\mathrm{Envy}^{\max}(\pi) \geq C_{\max}/2$.

For the proportionality lower bound, note that in the two-agent case, the proportional-share baseline for agent i is the average of the two bundles evaluated in i's cost units, namely

$$\frac{1}{2}\big(H_{i\leftarrow i}(\pi) + H_{i\leftarrow j}(\pi)\big).$$

Therefore,

$$\mathrm{PropExcess}_i(\pi) = \Big[H_{i\leftarrow i}(\pi) - \frac{1}{2}\big(H_{i\leftarrow i}(\pi) + H_{i\leftarrow j}(\pi)\big)\Big]_+ = \frac{1}{2}\big[H_{i\leftarrow i}(\pi) - H_{i\leftarrow j}(\pi)\big]_+ = \frac{1}{2}\,\mathrm{EnvyGap}_{i,j}(\pi).$$

It follows that $\mathrm{Prop}^{\max}(\pi) \geq \mathrm{PropExcess}_i(\pi) \geq C_{\max}/4$. □

B.3 Interpretation

Proposition B.2 shows that the surrogate-failure phenomenon does not depend on letting costs diverge. Even under the bounded domain $[0, C_{\max}]$, one can satisfy a statistical metric family, up to any prescribed tolerance, while still incurring harm-axiom violations whose magnitude is proportional to $C_{\max}$. Lemma B.1 further shows that, more generally, the relevant scale is given by $C_{\max}$ times the largest induced error-lottery gap.

C Generalized Harm Models

C.1 Bundles of harmful events

Let $\mathcal{K}$ be a finite index set of harmful events. The classical $\{\mathrm{FP}, \mathrm{FN}\}$ setting is recovered by taking $\mathcal{K} = \{\mathrm{FP}, \mathrm{FN}\}$, but $\mathcal{K}$ may also include audits, delays, additional scrutiny, stigma proxies, or any other event whose incidence depends on the policy.

For each individual ℓ, a policy π induces an event-probability bundle

$$p_\ell \in [0,1]^{|\mathcal{K}|}, \qquad p_{\ell,k} := \Pr_\pi(E_k \mid \ell) \quad (k \in \mathcal{K}),$$

and individual i evaluates ℓ's bundle through a harm functional

$$H_{i \leftarrow \ell}(\pi) := u_i(p_\ell),$$

where $u_i : [0,1]^{|\mathcal{K}|} \to \mathbb{R}_+$ is componentwise nondecreasing. The directed envy gap is

$$\mathrm{EnvyGap}_{i,j}(\pi) := \big[u_i(p_i) - u_i(p_j)\big]_+.$$

The point of this section is straightforward: the constructions in the main paper require only the ability to (i) create a gap in at least one coordinate of $p_i - p_j$, and (ii) allow subjective value to place sufficient weight on that coordinate. Neither ingredient is specific to linear FP/FN costs.

C.2 Additive harms over multiple events

A direct generalization of the main model is the additive multi-event form

$$u_i(p) = \sum_{k \in \mathcal{K}} c_{i,k}\, p_k, \qquad c_{i,k} \geq 0.$$

This strictly contains the FP/FN model as a special case.

Proposition C.1 (Multi-event extension of the surrogate-failure mechanism)**.** *Fix $(\mathcal{D}, \pi)$ and two individuals $i \neq j$. Suppose there exists $k \in \mathcal{K}$ such that $p_{i,k} \neq p_{j,k}$. Then, for any $B > 0$, there exists a nonnegative coefficient vector $(c_{i,k})_{k \in \mathcal{K}}$ such that* $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$.

Moreover, any statistical constraint $M(\pi)$ that depends only on the induced distribution of $(A, Y, \hat{Y})$, and possibly a score S, remains unchanged when $(c_{i,k})$ is modified.

Proof. Pick a coordinate k such that $\Delta_k := p_{i,k} - p_{j,k} \neq 0$. If $\Delta_k < 0$, swap the roles of (i,j). Set $c_{i,k} = C$ and $c_{i,k'} = 0$ for all $k' \neq k$. Then

$$u_i(p_i) - u_i(p_j) = C\,(p_{i,k} - p_{j,k}) = C\,\Delta_k,$$

so $\mathrm{EnvyGap}_{i,j}(\pi) = C\,\Delta_k$. Choosing $C := B/\Delta_k$ yields $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$.

The invariance of $M(\pi)$ is immediate: statistical constraints depend on $(\mathcal{D}, \pi)$ through predictions, scores, and labels, not through subjective coefficients. □

This proposition already addresses the most common "beyond FP/FN" objection: even if the harm space is enriched to include audits, delays, and similar events, the same amplification and counterexample logic continues to apply by focusing on a single event coordinate.

C.3 A minimal richness condition beyond linearity

To address nonlinear harms, we do not commit to a specific functional form. Instead, we assume only that the family of admissible harm models is *rich enough* to express concentrated sensitivity to any single event.

Let $\mathcal{U}_i$ denote the admissible family of monotone harm functionals for individual i.

Richness assumption. For each $k \in \mathcal{K}$ and each $\theta \geq 0$, the single-coordinate harm

$$u_{i,k,\theta}(p) := \theta\, p_k$$

belongs to $\mathcal{U}_i$.

This is a weak assumption in practice: most nonlinear families used to model subjective disutility (e.g., monotone transformations, piecewise-linear or convex utilities, max-over-linear forms, or separable sums) contain such linear atoms, at least approximately.

Proposition C.2 (Reduction to a single coordinate)**.** *Assume the richness condition above. Fix $(\mathcal{D}, \pi)$ and suppose there exist $i \neq j$ and $k \in \mathcal{K}$ such that $p_{i,k} \neq p_{j,k}$. Then, for any $B > 0$, there exists a choice of $u_i \in \mathcal{U}_i$ such that* $\mathrm{EnvyGap}_{i,j}(\pi) \geq B$*, while any cost-blind statistical constraint value $M(\pi)$ remains unchanged.*

Proof. Let $\Delta_k := p_{i,k} - p_{j,k}$. If $\Delta_k < 0$, swap (i, j). By the richness assumption, take $u_i = u_{i,k,\theta}$ with $\theta := B/\Delta_k$. Then

$$u_i(p_i) - u_i(p_j) = \theta(p_{i,k} - p_{j,k}) = B,$$

so $\mathrm{EnvyGap}_{i,j}(\pi) = [B]_+ = B$. As in the additive case, $M(\pi)$ depends only on $(\mathcal{D}, \pi)$ and is therefore unaffected by the choice of u_i. □

Proposition C.2 formalizes a simple point: the surrogate-failure constructions in the main paper do not rely on *linearity* as such, but on *directional sensitivity* to at least one harm event. If a harm-model class admits concentrated sensitivity along a coordinate, then any policy that induces a nonzero gap on that coordinate can generate large harm-axiom violations.

C.4 Bounded-rationality floors under generalized harms

The bounded-rationality lower bound in the main text is likewise not tied to linear FP/FN harms. What matters is that the mediator faces an uncertainty set over subjective harms and must choose a policy from a restricted class.

The following proposition is a direct analogue of the main minimax floor, stated for a single event coordinate.

Proposition C.3 (A minimax envy floor with uncertain harm sensitivity)**.** *Fix $k \in \mathcal{K}$ and consider a restricted policy class Π_{simple} that couples treatment so that, for some pair (i, j), the induced event-probability gap*

$$\Delta_k(\pi) := p_{i,k} - p_{j,k}$$

cannot be driven to zero (e.g., because of identical-treatment constraints).

Assume the mediator observes a proxy sensitivity $\hat{\theta}$, while the true sensitivity satisfies $\theta \in [\hat{\theta} - \epsilon, \hat{\theta} + \epsilon]$, and assume that $u_{i,k,\theta}(p) = \theta p_k$ is admissible by richness. Then, for any $\pi \in \Pi_{\text{simple}}$,

$$\sup_{\theta \in [\hat{\theta}-\epsilon, \hat{\theta}+\epsilon]} \mathrm{EnvyGap}_{i,j}(\pi) \;\geq\; \epsilon \cdot |\Delta_k(\pi)|,$$

and consequently,

$$\inf_{\pi \in \Pi_{\text{simple}}} \sup_{\theta \in [\hat{\theta}-\epsilon, \hat{\theta}+\epsilon]} \mathrm{EnvyGap}_{i,j}(\pi) \;\geq\; \epsilon \cdot \inf_{\pi \in \Pi_{\text{simple}}} |\Delta_k(\pi)|.$$

Proof. Under $u_{i,k,\theta}(p) = \theta p_k$,

$$\mathrm{EnvyGap}_{i,j}(\pi) = \big[\theta(p_{i,k} - p_{j,k})\big]_+.$$

If $\Delta_k(\pi) \geq 0$, the worst case over $\theta \in [\hat{\theta} - \epsilon, \hat{\theta} + \epsilon]$ is attained at $\theta = \hat{\theta} + \epsilon$, yielding

$$\mathrm{EnvyGap}_{i,j}(\pi) \geq (\hat{\theta} + \epsilon)\Delta_k(\pi) \geq \epsilon\, \Delta_k(\pi).$$

If $\Delta_k(\pi) < 0$, swapping (i, j) yields the same bound in terms of $|\Delta_k(\pi)|$. Taking the infimum over $\pi \in \Pi_{\mathrm{simple}}$ gives the stated minimax lower bound. □

Proposition C.3 shows that the approximation floor arises from the combination of (i) policy-class restrictions that prevent elimination of event-probability gaps, and (ii) uncertainty about subjective sensitivity. This logic persists even when harm is modeled using richer function classes, provided that the class contains, or can approximate, single-coordinate sensitivity.

D Auxiliary Lemmas and Tightness Notes

This section collects several elementary facts used repeatedly throughout the appendix and records the cases in which the lower bounds proved above are attained, i.e., are tight, on the constructed instances.

D.1 Basic algebra for directed gaps

Directed envy and proportionality expressions typically reduce to signed differences between two bundles evaluated under a fixed individual's valuation. The following elementary observation is used repeatedly.

Lemma D.1 (Orientation of a gap)**.** *For any real number z,*

$$[z]_+ = \max\{z, 0\}.$$

In particular, for any ordered pair (i, j),

$$\mathrm{EnvyGap}_{i,j}(\pi) = [u_i(p_i) - u_i(p_j)]_+.$$

Hence,

$$u_i(p_i) > u_i(p_j) \quad \Longrightarrow \quad \mathrm{EnvyGap}_{i,j}(\pi) > 0.$$

Equivalently, whenever $u_i(p_i) \neq u_i(p_j)$, exactly one of

$$[u_i(p_i) - u_i(p_j)]_+, \qquad [u_i(p_j) - u_i(p_i)]_+$$

is strictly positive.

Proof. This follows immediately from the definition of $[\cdot]_+$. □

D.2 Envy and proportionality for two agents

Several lower-bound arguments in the appendix use two-agent constructions. In this case, proportionality excess can be expressed directly in terms of a directed envy gap under the same valuation.

Lemma D.2 (Two-agent proportionality is half a directed envy gap)**.** *Let $n = 2$. For agent i, define the proportional-share baseline as the average of the two bundles evaluated under i's valuation:*

$$\mathrm{PS}_i(\pi) := \tfrac{1}{2}\big(H_{i\leftarrow i}(\pi) + H_{i\leftarrow j}(\pi)\big), \quad j \neq i,$$

and define

$$\mathrm{PropExcess}_i(\pi) := [H_{i\leftarrow i}(\pi) - \mathrm{PS}_i(\pi)]_+.$$

Then

$$\mathrm{PropExcess}_i(\pi) = \tfrac{1}{2}\,\mathrm{EnvyGap}_{i,j}(\pi).$$

Consequently, on any two-agent instance,

$$\mathrm{Prop}^{\max}(\pi) = \tfrac{1}{2}\,\mathrm{Envy}^{\max}(\pi).$$

Proof. Expanding the definition gives

$$\mathrm{PropExcess}_i(\pi) = \Big[H_{i\leftarrow i}(\pi) - \tfrac{1}{2}\big(H_{i\leftarrow i}(\pi) + H_{i\leftarrow j}(\pi)\big)\Big]_+ = \tfrac{1}{2}\big[H_{i\leftarrow i}(\pi) - H_{i\leftarrow j}(\pi)\big]_+ = \tfrac{1}{2}\,\mathrm{EnvyGap}_{i,j}(\pi).$$

Since for $n = 2$ the only ordered pairs are $(1, 2)$ and $(2, 1)$, taking maxima over i yields

$$\mathrm{Prop}^{\max}(\pi) = \tfrac{1}{2}\,\mathrm{Envy}^{\max}(\pi).$$

□

D.3 Corner maximization under box uncertainty

When the mediator faces a box uncertainty set, i.e., componentwise interval bounds, the worst-case value of any linear form is attained at a corner. This is the only robust-optimization fact needed in the appendix.

Lemma D.3 (Box uncertainty: worst case at a corner)**.** *Let $x \in [\ell, u]^m$ and let $g(x) = a^\top x$ be a linear function. Then*

$$\sup_{x\in[\ell,u]^m} g(x) = \sum_{r=1}^{m} a_r \cdot \begin{cases} u_r & \text{if } a_r \geq 0, \\ \ell_r & \text{if } a_r < 0. \end{cases}$$

In particular, the supremum is attained at a corner point of the box.

Proof. For each coordinate r, the term $a_r x_r$ is maximized by setting $x_r = u_r$ if $a_r \geq 0$ and $x_r = \ell_r$ if $a_r < 0$. Summing over coordinates yields the stated expression. □

Lemma D.3 explains why the robust envy bounds in arguments of the type used for Theorem 7.4 can be derived by explicitly selecting adversarial costs at the endpoints of the uncertainty intervals.

D.4 A simple minimax identity for two affine pieces

Several floor proofs reduce to minimizing the maximum of two affine functions of a scalar decision variable. We record the resulting closed form.

Lemma D.4 (Minimizing the maximum of two lines)**.** *Let*

$$g(\alpha) = \max\{a - b\alpha,\ c + d\alpha\},$$

where $b, d > 0$. Then g is minimized at the unique intersection point $\alpha^\star$ satisfying

$$a - b\alpha^\star = c + d\alpha^\star,$$

namely

$$\alpha^\star = \frac{a - c}{b + d}, \qquad \min_\alpha g(\alpha) = a - b\alpha^\star = c + d\alpha^\star.$$

If α is restricted to an interval, the minimizer is given by clipping $\alpha^\star$ to that interval.

Proof. The function g is the pointwise maximum of two affine functions and is therefore convex and piecewise affine, with a single kink at the intersection point. The maximum is minimized when the two affine pieces are equal; otherwise, moving toward the intersection decreases the larger term. □

D.5 Tightness notes

We now record that the main lower bounds used in the appendix are attained on the constructed instances.

Proposition D.5 (Tightness of the bounded-rationality floor instance)**.** *On the instance used for Theorem 7.4—two individuals, coupled treatment with parameter α, and box uncertainty of radius ϵ around proxy costs—the minimax envy value equals ϵ:*

$$\inf_{\pi\in\Pi_{\text{simple}}} \sup_{c\in\mathcal{C}(\hat{c},\epsilon)} \mathrm{Envy}^{\max}(\pi; c) \;=\; \epsilon.$$

Proof. The lower bound $\geq \epsilon$ follows from the proof of Theorem 7.4, together with Lemma D.3 and Lemma D.4: the bound is minimized at $\alpha = \frac{1}{2}$, where it equals ϵ.

For the matching upper bound, choose the policy $\pi^\star$ corresponding to $\alpha = \frac{1}{2}$. On this instance, the two directed gaps reduce to

$$\mathrm{EnvyGap}_{1,2}(\pi^\star; c) = \left[\tfrac{1}{2}(c_1^{FN} - c_1^{FP})\right]_+, \qquad \mathrm{EnvyGap}_{2,1}(\pi^\star; c) = \left[\tfrac{1}{2}(c_2^{FP} - c_2^{FN})\right]_+.$$

Since $c_i^{FP}, c_i^{FN} \in [1-\epsilon, 1+\epsilon]$, each bracket is bounded above by

$$\tfrac{1}{2} \cdot 2\epsilon = \epsilon,$$

and therefore $\mathrm{Envy}^{\max}(\pi^\star; c) \leq \epsilon$ for all c in the uncertainty set. Hence

$$\sup_{c \in \mathcal{C}(\hat{c}, \epsilon)} \mathrm{Envy}^{\max}(\pi^\star; c) \leq \epsilon,$$

which implies that the minimax value is at most ϵ. Combined with the lower bound, this yields the claimed equality. □

Proposition D.6 (Tightness of bounded-cost amplification for a fixed pair)**.** *Fix a policy π and an ordered pair (i, j). Under bounded costs $c_i^{FP}, c_i^{FN} \in [0, C_{\max}]$,*

$$\sup_{c_i^{FP}, c_i^{FN} \in [0, C_{\max}]} \mathrm{EnvyGap}_{i,j}(\pi) = C_{\max} \cdot \max\{(\Delta_{i,j}^{FN})_+, (\Delta_{i,j}^{FP})_+\},$$

and the supremum is attained by placing the full weight $C_{\max}$ on the coordinate with the larger positive gap.

Proof. This is exactly the optimization carried out in Lemma B.1. The quantity

$$\mathrm{EnvyGap}_{i,j}(\pi) = \left[c_i^{FP} \Delta_{i,j}^{FP} + c_i^{FN} \Delta_{i,j}^{FN}\right]_+$$

is linear in (c_i^{FP}, c_i^{FN}) inside the bracket. Hence, by Lemma D.3, the maximum over the box is attained at an extreme point. The optimal extreme point is either $(C_{\max}, 0)$ or $(0, C_{\max})$, depending on which positive gap is larger. □

These tightness statements justify interpreting the lower bounds as genuine *floors*, rather than as artifacts of a loose analysis: on the constructed instances, the lower bounds coincide with the achievable robust optimum.

E Robustness Analyses for Proxy Envy and Cost Heterogeneity

This appendix provides additional robustness analyses to verify that the paper's main qualitative conclusions are not artifacts of specific implementation choices or extreme parameter settings. First, because the cost-stratum envy proxy depends on the discretization level K used to approximate pairwise comparisons, we sweep K over a wide range and examine whether policy rankings and best-policy selection remain stable relative to the default choice. Second, we vary the heterogeneity-strength parameter γ from mild to extreme values in order to address the concern that the observed misalignment between surrogate fairness satisfaction and harm- or axiom-based criteria is driven only by unrealistically large cost disparities. Instead, we assess whether nontrivial gaps already emerge under mild heterogeneity and then grow as heterogeneity increases.

E.1 K-sweep Stability of the Cost-Stratum Envy Proxy

Experimental setup. We examine the sensitivity of the cost-stratum (type-binned) envy proxy to the discretization level K used to define cost strata. Following the main experimental pipeline, we train two scoring models (LR and HGB) on three ACS tasks (Income, PublicCoverage, Employment) using three random seeds, with `SEX` as the binary group attribute and excluded from the feature set. For each trained

scorer, we generate a collection of post-processing policies within each surrogate family: demographic parity (DP; selection-rate sweep), equalized odds (EO; threshold-pair search under an EO tolerance), and calibration mixing (CAL; isotonic calibration with mixing weights). For each policy, we compute the envy proxy by partitioning individuals into K strata according to a cost-type index derived from their relative FP/FN costs, and then evaluating the resulting maximum type-level envy. We sweep $K \in \{20, 40, 80, 160, 320\}$ and treat $K = 80$ as the reference value. Stability is evaluated under two cost settings: the main heterogeneity configuration (Hetero) and a correlation-sensitivity sweep over ρ (Hetero-ρ). For each setting instance, we exclude degenerate cases in which the reference envy vector at $K = 80$ is constant across policies, since rank-based stability is not meaningful in such cases.

Results and analysis. Table 8 reports two complementary stability measures relative to $K = 80$: (i) the median Spearman rank correlation between the policy-level envy vector at K and the corresponding vector at $K = 80$, and (ii) the Top-1 agreement rate, i.e., the fraction of settings in which the envy-minimizing policy at K coincides with the envy-minimizing policy at $K = 80$. Overall, EO exhibits the strongest stability, especially under the ρ-sweep, where rank correlations are nearly perfect for $K = 40$ and $K = 160$, and Top-1 agreement remains high across all values of K. DP shows moderate rank stability but weaker Top-1 agreement, particularly at very large K, consistent with the intuition that finer discretization can increase within-stratum estimation noise and thereby perturb the identity of the minimizer even when the overall ranking remains partially preserved. CAL displays a middle-range stability pattern: intermediate values of K (e.g., $K = 40$ or $K = 160$ under Hetero-ρ) yield higher rank stability, whereas extreme discretizations can reduce stability either by over-coarsening the proxy (small K) or by inducing sparse strata (large K). Importantly, across scenarios and surrogate families, stability is generally strongest around the default choice $K = 80$, supporting the use of the cost-stratum envy proxy as a robust comparative signal rather than as a fragile artifact of a single discretization.

Table 8: Stability of the cost-stratum envy proxy with respect to the number of strata K, measured relative to the reference choice $K = 80$. We report (i) the median Spearman rank correlation between the policy-level envy vector at K and that at $K = 80$, and (ii) the Top-1 agreement rate (i.e., whether the envy-minimizing policy matches that selected at $K = 80$). N denotes the number of non-degenerate setting instances included in the aggregation.

Scenario	Policy	N	Spearman rank correlation: K=80				Top-1 agreement: K=80			
			K=20	K=40	K=160	K=320	K=20	K=40	K=160	K=320
Hetero	DP	18	0.6429	0.7679	0.7679	0.4821	0.3333	0.6111	0.3333	0.1667
	EO	18	0.5000	0.5000	1.0000	0.9330	0.5556	0.4444	0.6111	0.6111
	CAL	15	0.8030	0.7456	0.3627	0.2962	0.5333	0.4000	0.4667	0.4667
Hetero-ρ	DP	90	0.5357	0.6099	0.6964	0.6726	0.3556	0.4444	0.3778	0.4000
	EO	90	0.5000	1.0000	1.0000	0.8660	0.5556	0.6222	0.7333	0.6889
	CAL	68	0.3953	0.7906	0.7906	0.4326	0.4265	0.5735	0.4853	0.4118

E.2 γ-sweep Sensitivity to the Strength of Cost Heterogeneity

Experimental setup. We next test whether the surrogate–axiom misalignment observed in the main experiments is an artifact of selecting an unrealistically extreme cost-heterogeneity regime. To this end, we sweep the heterogeneity-strength parameter γ, which controls the magnitude of group-dependent cost scaling, over $\gamma \in \{1.0, 1.25, 1.5, 2.0, 3.0, 5.0\}$, while holding the remaining cost parameters fixed, including the idiosyncratic noise scale and correlation. For each ACS task (Income, PublicCoverage, Employment), model class (LR, HGB), and seed (three random seeds), we generate the same policy candidate sets for each surrogate family (DP, EO, CAL) as in the main pipeline. For each γ, we evaluate an *axiom gap* based on envy: within each surrogate family, we select the *surrogate-optimal* policy, defined as the policy with minimum classification error, and compare its envy with the best attainable envy among *all* candidate

policies, including policies from other families and the baseline. Formally, for each setting instance,

$$\mathrm{Gap}(\gamma) \;=\; E(\pi_{\mathrm{sur}}(\gamma)) \;-\; \min_{\pi\in\Pi} E(\pi;\gamma),$$

where $E(\cdot)$ denotes the cost-stratum envy proxy computed using the default bin count $K = 80$, and Π denotes the full candidate set. We report the median and interquartile range (IQR) of $\mathrm{Gap}(\gamma)$ across all $3 \times 2 \times 3 = 18$ setting instances.

Results and analysis. Table 9 and Figure 1 summarize the γ-sweep robustness analysis. Across all surrogate families (DP, EO, CAL), the axiom gap is already nonzero under mild heterogeneity (e.g., $\gamma \in \{1.0, 1.25, 1.5\}$), indicating that surrogate-optimal choices can be suboptimal from the perspective of envy even when cost disparities are not extreme. As heterogeneity increases ($\gamma \geq 2.0$), the gap generally grows and its dispersion across tasks, models, and seeds also widens, consistent with stronger and more variable misalignment in more heterogeneous regimes. In the most heterogeneous setting considered here ($\gamma = 5.0$), the gap becomes large for all surrogate families, underscoring that surrogate–axiom misalignment can become severe as cost heterogeneity strengthens. Overall, these results support the robustness claim that the observed misalignment is not an artifact of artificially extreme parameter choices; rather, it already emerges under mild heterogeneity and can intensify substantially as heterogeneity increases.

Table 9: Robustness to the strength of cost heterogeneity γ. Top: median axiom gap. Bottom: interquartile range (25th–75th percentiles).

(a) Median axiom gap $E(\pi_{\mathrm{sur}}) - \min_{\pi} E(\pi)$.

	γ=1.0	γ=1.25	γ=1.5	γ=2.0	γ=3.0	γ=5.0
DP	0.3026	0.5478	0.8002	3.2951	3.1256	27.8322
EO	0.4087	0.5630	0.7149	3.1453	3.5498	17.0974
CAL	0.2758	0.5383	0.5223	3.5701	2.7258	16.1681

(b) IQR (25th–75th pct.) of the axiom gap.

	γ=1.0	γ=1.25	γ=1.5	γ=2.0	γ=3.0	γ=5.0
DP	0.1177–0.8750	0.2388–1.2873	0.3675–1.6872	2.8084–4.2489	1.8608–5.7724	15.9506–46.6524
EO	0.1689–0.7194	0.3591–1.1283	0.3649–1.2010	1.6267–5.0086	1.3076–5.7085	5.7565–26.8462
CAL	0.1061–0.4944	0.0941–0.8058	0.2607–1.5470	2.8042–4.2714	0.4738–5.7549	5.0628–38.3248

E.3 Summary and implications

Taken together, these two robustness analyses clarify how the harm-based results should be interpreted in the presence of implementation choices and cost-modeling assumptions. First, the K-sweep indicates that the cost-stratum envy proxy provides a reasonably stable signal for *ranking* policies within a surrogate family around the default discretization, even though the identity of the single best policy may vary when the discretization becomes extremely coarse or extremely fine. This supports using the proxy primarily as a comparative diagnostic, for example in assessing non-dominance and relative trade-offs, rather than as a brittle point estimate tied to a single choice of K. Second, the γ-sweep shows that surrogate–axiom misalignment is not confined to extreme heterogeneity: qualitatively similar gaps are already present under mild heterogeneity and tend to widen as heterogeneity strengthens, with greater dispersion across settings in more heterogeneous regimes. Overall, these checks reinforce the conclusion that the main findings reflect a structural tension between surrogate constraints and harm-based criteria, rather than an artifact of a particular binning choice or an artificially extreme cost scale.

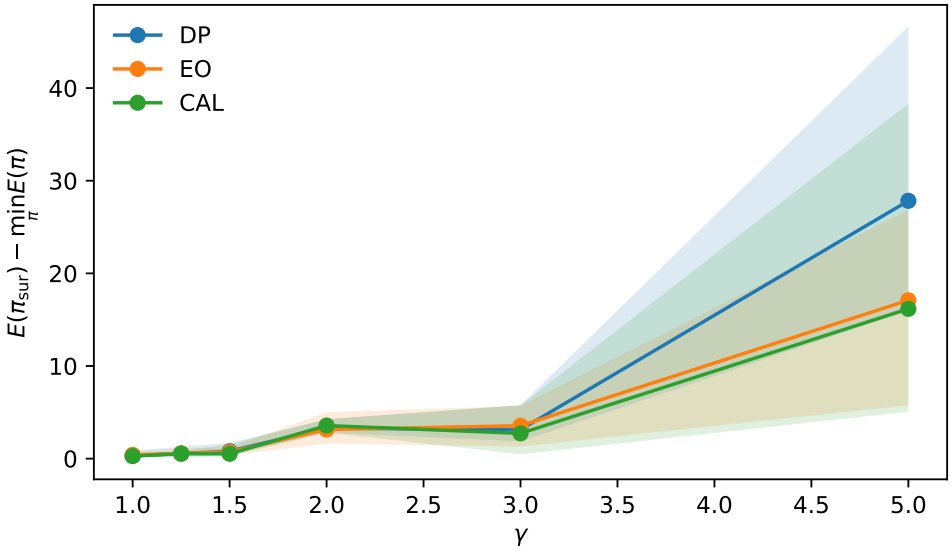

Figure 1: Trend of the axiom gap as γ increases (median with IQR band).

F Discussion of Modeling Scope

This section clarifies what the theoretical results *do* and *do not* cover. Its purpose is not to introduce new claims, but to preempt predictable misunderstandings about the sources of the impossibility results, counterexamples, and approximation floors.

F.1 What drives the surrogate-failure results

At a high level, the surrogate-failure results require only two ingredients.

(i) A policy induces heterogeneous harm-relevant bundles. There exist individuals (or strata) $i \neq j$ such that the policy π induces different bundles $p_i \neq p_j$ of harm-relevant event probabilities. In the main text, this bundle is (p_i^{FP}, p_i^{FN}), but Appendix C shows that the argument extends to richer bundles.

(ii) Subjective value is not collapsed into a single comparable unit. The admissible value or cost class allows individuals to weight at least one coordinate of the harm bundle differently, so that differences in $p_i - p_j$ translate into differences in experienced harm.

The results do *not* depend on any particular learning algorithm, application domain, or optimization procedure; rather, they are policy-level statements about how statistical constraints relate to axioms of harm allocation.

F.2 When statistical fairness can coincide with harm-based axioms

The theory does not claim that statistical constraints *never* align with harm-based axioms. Rather, such alignment requires additional structure that effectively collapses subjective value or eliminates the relevant heterogeneity in induced bundles.

Typical sufficient conditions include:

- **Homogeneous value scales.** All individuals share a common (c^{FP}, c^{FN}) up to a global scaling factor, or, more generally, the admissible costs lie in a low-dimensional aligned family. Under such a collapse, controlling error rates can meaningfully control harm.

- **Bundle equalization at the relevant granularity.** If a policy equalizes the harm bundle itself (e.g., equalizes (p^{FP}, p^{FN}) across the relevant comparison class), then envy violations vanish by definition. This is a strong requirement and is often infeasible without substantial efficiency loss.

In other words, the negative results identify what must hold *for* statistical definitions to function as faithful surrogates; they are not intended to deny that surrogates can succeed in restricted regimes.

F.3 Which assumptions are essential, and which are conveniences

Subjective value heterogeneity. Some form of heterogeneity is essential for the strongest separation results: if all agents share identical values, then the harm allocation is effectively single-objective and many conflicts disappear. However, the results do not depend on any particular parametric form, such as linear FP/FN costs; they require only a class rich enough to express directional sensitivity (Appendix C).

Statistical constraints being cost-blind. This assumption is essential and structural: DP, EO, calibration, and similar criteria are functions of induced prediction or score distributions and labels, not of private costs. The resulting mismatch is therefore not an artifact of optimization, but a representational limitation.

Two-agent or otherwise small constructions. Using two agents, or only a few agents, is a proof convenience. The same mechanisms extend to larger populations by embedding the small instance, for example by holding other agents' bundles fixed or assigning them neutral costs.

Single-group settings in counterexamples. Some constructions use a single group so that across-group statistical constraints become vacuous. This is intentional: it shows that certain harm-axiom violations can arise even in the absence of group structure. When group structure is present, the constructions can be adapted by isolating two individuals within a group or by working across two groups, depending on which notion is being targeted.

F.4 How to interpret the bounded-rationality floors

Our approximation-floor statements quantify an *irreducible residual* in harm-axiom satisfaction that arises when a mediator must choose a policy under both limited information about values and limited policy expressiveness. Formally, the object of interest is the minimax residual

$$\delta^\star \;=\; \inf_{\pi \in \Pi_{\text{simple}}} \; \sup_{c \in \mathcal{C}(\hat{c},\epsilon)} V(\pi; c),$$

where $V(\pi; c)$ is a harm-axiom violation score (e.g., $\text{Envy}^{\max}$ or a proportionality-style excess), Π_{simple} is a restricted policy class, and $\mathcal{C}(\hat{c}, \epsilon)$ encodes uncertainty about the true subjective costs around a proxy $\hat{c}$.

Two ingredients and their roles. The residual $\delta^\star$ is driven by the combination of:

- **Information limits (selection under uncertainty):** the mediator does not observe c directly and must evaluate or select policies using only a proxy $\hat{c}$ (or coarse strata), so Nature can realize some $c \in \mathcal{C}(\hat{c}, \epsilon)$ for which the selected policy violates the axiom.
- **Policy limits (expressiveness):** the mediator chooses from a restricted policy class Π_{simple}, which may be unable to implement the allocations needed to eliminate the relevant bundle gaps for the realized costs.

Why the floor is not merely "one limitation or the other." It is useful to distinguish two sources of residual unfairness. First, even under perfect information, a restricted class can induce a *policy-class floor*

$$\inf_{\pi \in \Pi_{\text{simple}}} V(\pi; c) > 0$$

for some c, simply because the class cannot express the required tailoring. Second, even when the policy class is expressive enough in principle, imperfect information induces an *information floor*: the mediator may choose π using $V(\pi; \hat{c})$, while the realized costs are $c \neq \hat{c}$, so the realized value $V(\pi; c)$ remains bounded away from zero in the worst case.

Importantly, neither limitation alone *universally* implies a nonzero floor: for some instances, perfect information can eliminate the information component, and sufficiently rich policy classes can eliminate the expressiveness component. Our lower-bound results identify instances in which the *combined* setting yields a nontrivial minimax residual.

Interpretation. Appendix D records tightness statements for the canonical two-agent construction, showing that the derived bounds are not artifacts of slack analysis. In that instance, the minimax residual scales linearly with the information radius (e.g., $\delta^\star = \epsilon$), illustrating a direct "price" mechanism: an ϵ-level uncertainty in subjective costs induces an unavoidable ϵ-level residual axiom violation under the restricted policy class.

F.5 Common questions and short answers

Does the theory depend on harms being errors (FP/FN)? No. The arguments require only the existence of a harm-relevant event bundle whose incidence depends on the policy, together with a value class that can weight bundle coordinates. Errors are simply the most standard and measurable examples.

Does allowing bounded costs eliminate the phenomenon? No. Appendix B provides finite-cost lower bounds; the relevant scale becomes $C_{\max}$ times the induced bundle gap rather than being unbounded.

Is this an argument against using statistical fairness metrics in practice? No. The results should instead be read as a caution about interpretation: satisfying a statistical metric is not, by itself, a guarantee of fair harm allocation when subjective value and intermediation constraints are present. In restricted regimes, such as value collapse or bundle equalization, statistical metrics can coincide with harm-based axioms, but those regimes require explicit justification.