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 constrain observable rates, but do not determine how error burdens are allocated. This paper studies fairness as a conditional harm-allocation problem: given an evaluation target and an adopted false-positive/false-negative cost representation, a policy induces cost-weighted error burdens. We introduce fair-division diagnostics for these allocative bads, characterize when statistical criteria can serve as burden surrogates, and show how heterogeneous adopted costs and restricted policy classes create surrogate failures and approximation floors. Experiments on benchmark prediction tasks provide controlled stress tests showing that statistical improvements can diverge from harm-allocation diagnostics.

1 Introduction

Machine-learning systems increasingly mediate consequential decisions in domains such as credit, employment, insurance, education, and public services. Fairness in these settings is often evaluated through statistical criteria such as Demographic Parity, Equalized Odds, and Calibration. These criteria constrain observable decision patterns, including acceptance rates, conditional error rates, and score semantics. However, observable parity does not determine how error-induced burdens are allocated across individuals. The same false-positive or false-negative event can impose different burdens depending on the person affected, the opportunity at stake, and the institutional context of the decision.

This paper studies fairness as the allocation of cost-weighted error burdens. A decision rule induces false-positive and false-negative events relative to an adopted evaluation label, and these events become burdens when evaluated through an adopted harm representation. This separates error-event accounting from harm accounting: the former depends on the policy and evaluation target, while the latter depends on the cost representation used to interpret the error events. Under this view, statistical fairness criteria are surrogate conditions whose harm-based meaning depends on how error costs are structured.

The resulting allocation problem is naturally connected to fair division. Error burdens are allocative bads distributed by a decision procedure. Envy-freeness captures comparative grievance by asking whether an individual, using her own cost representation, evaluates her assigned error burden as worse than another individual's burden. Proportionality captures excessive-share burden by asking whether an individual bears more than an equal share of the total burden in her own cost units. These diagnostics make explicit forms of unfairness that can remain hidden when evaluation is limited to group-level rates.

Institutional implementation further constrains harm-based fairness. Decision makers often rely on proxy cost information and restricted policy classes, such as global thresholds, group-specific thresholds, or finite post-processing families. As a result, residual fair-division violation can persist even when the decision maker directly targets a harm-based objective. Fairness under bounded information and bounded policy expressiveness is therefore an approximation problem.

The contributions are as follows.

- **Conditional harm-allocation framework.** The paper formalizes the allocation of error-induced burdens under an adopted evaluation label and an adopted false-positive/false-negative cost representation.
- **Fair-division diagnostics for allocative bads.** Envy gap and proportional excess provide diagnostics for comparative grievance and excessive-share burden in policy-induced error allocations.
- **Surrogate boundary conditions and failure.** The analysis identifies cost-structure conditions under which statistical fairness criteria become informative about harm allocation, and shows why heterogeneous adopted costs break this surrogate relationship.
- **Bounded-rationality approximation floors.** Proxy cost information and restricted policy classes are incorporated into the fairness problem, showing why residual violation can remain under implementation limits.
- **Empirical stress tests of surrogate limits.** Controlled evaluations examine how aligned costs, heterogeneous costs, policy families, and fairness interventions shape the relationship between statistical improvements and harm-allocation diagnostics.

2 Related Work

2.1 Statistical Fairness Criteria as Surrogates

Fair machine learning has made fairness operational through statistical relationships among predictions, labels, scores, and protected attributes. Early formulations introduced similarity-based individual fairness, fair representation learning, disparate-impact removal, and misclassification-based constraints (Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Zafar et al., 2017). Subsequent work formalized criteria such as Demographic Parity, Equalized Odds, Equal Opportunity, Calibration, and Multicalibration, while also showing that these criteria can conflict under unequal base rates or when calibrated scores are converted into binary decisions (Hardt et al., 2016; Kleinberg et al., 2016; Pleiss et al., 2017; Hébert-Johnson et al., 2018). Together, these criteria define a family of low-dimensional statistical constraints that regulate distinct aspects of the decision process.

A large algorithmic literature studies how such constraints can be audited, enforced, or approximated. Subgroup auditing extends parity-style reasoning beyond coarse protected groups, while reductions-based, constraint-based, and generalization-aware approaches turn fairness requirements into tractable learning or post-processing problems (Kearns et al., 2018; Zafar et al., 2019; Agarwal et al., 2018; Celis et al., 2019; Cotter et al., 2019; Agarwal et al., 2019). Further work examines post-processing limits, missing or noisy demographic information, inequality-based unfairness measurement, and broader critiques of metric-based evaluation (Canetti et al., 2019; Mozannar et al., 2020; Celis et al., 2021; Speicher et al., 2018; Corbett-Davies et al., 2023; Hsu et al., 2022). In our setting, these criteria remain useful observable restrictions, but their harm-based meaning depends on whether the constrained statistical object tracks the allocation of false-positive and false-negative burdens.

2.2 Welfare, Utility, and Cost Representations

A related body of work evaluates algorithmic decisions through welfare, utility, risk, and cost. Welfare-based analyses connect fair machine learning to social choice and distributive justice by asking how decision rules affect the distribution of benefits and losses across individuals or groups (Heidari et al., 2018; Hu & Chen, 2020; Cousins, 2021; Pardeshi et al., 2024). Cost- and risk-based formulations interpret fairness constraints through cost-sensitive classification, subgroup risk, or utility trade-offs (Menon & Williamson, 2018; Williamson & Menon, 2019; Li & Liu, 2022). These perspectives shift the object of analysis from rate satisfaction alone to the valued consequences induced by a policy.

The same concern appears in sequential decision-making, resource allocation, ranking, and recommendation. In bandit, matching, and reinforcement-learning settings, fairness is often expressed through social welfare,

Nash welfare, utility-maximizing distributions, or constrained welfare optimization under uncertainty (Hossain et al., 2021; Barman et al., 2023; Devic et al., 2023; Lobo et al., 2024; Zhang et al., 2024; Siddique et al., 2020). Ranking and recommendation systems make the allocation problem explicit by distributing exposure, attention, and opportunity among multiple stakeholders (Singh & Joachims, 2018; Do et al., 2021; Wang et al., 2021; Ge et al., 2022; Wu et al., 2022; Greenwood et al., 2024). Algorithmic recourse adds a further burden-based perspective by measuring the difficulty, feasibility, or cost of changing an unfavorable decision (Ustun et al., 2019; Barocas et al., 2020; Karimi et al., 2021; Perello et al., 2025). The present paper uses cost representations in a narrower diagnostic role: false-positive and false-negative costs are adopted evaluative inputs, and the question is whether statistical criteria track the burden allocation induced under those inputs.

2.3 Fair Division and Allocation of Bads

Fair division evaluates fairness as a property of allocations. Classic and algorithmic work has developed axioms such as envy-freeness, proportionality, competitive equilibrium from equal incomes, maximin-share guarantees, Nash social welfare, and ordinal assignment principles (Bogomolnaia & Moulin, 2001; Lipton et al., 2004; Bouveret et al., 2005; Bouveret & Lemaître, 2016; Budish, 2011). For indivisible resources, exact fairness is often unattainable, which motivates approximate notions such as envy-freeness up to one item, envy-freeness up to any item, approximate maximin-share guarantees, and welfare-based fair allocations (Cole & Gkatzelis, 2015; Kurokawa et al., 2016; 2018; Caragiannis et al., 2019; Chaudhury et al., 2024; Amanatidis et al., 2021; Garg & Taki, 2020; Amanatidis et al., 2023). For classification, the key insight is that a decision rule induces an allocation of error events and their associated burdens.

The allocation of bads, chores, mixed resources, and public goods is especially close to error-burden analysis. Classification errors are burdens imposed by a decision procedure, and fair-division work on goods and chores, mixed divisible and indivisible resources, public decisions, and public goods shows how allocation principles can be adapted when items have negative value, when benefits and burdens coexist, or when a collective decision affects many agents at once (Aziz et al., 2022; Bhaskar et al., 2020; Bei et al., 2021a;b; Conitzer et al., 2017; Fain et al., 2018; Garg et al., 2021). Envy gap and proportional excess translate this allocation vocabulary into the present setting: the former captures comparative grievance over assigned error-event burdens, while the latter captures excessive-share burden under the individual's adopted cost representation.

2.4 Fairness under Uncertainty and Bounded Rationality

Fairness evaluation is shaped by uncertainty in the objects being measured. Labels, protected attributes, and decision targets may be selectively observed, measured through proxies, or produced by institutional procedures. Work on selective labels and imperfect prediction shows that available training data can depend on prior decisions (Kilbertus et al., 2020; Wei, 2021). Related work on proxy labels and unobserved or noisy protected attributes shows that fairness assessment becomes statistically ambiguous when the target construct or protected class is missing, coarse, or indirectly measured (Guerdan et al., 2023; Chen et al., 2019; Kallus et al., 2022; Zhu et al., 2023; Mehrotra & Vishnoi, 2022). In a harm-allocation analysis, the same issue appears through the adopted evaluation label: false positives and false negatives are counted relative to an evaluation target whose institutional status should be made explicit.

Institutional mediation also appears when algorithmic predictions enter human or organizational decision pipelines. Learning-to-defer and human-AI collaboration frameworks study settings in which a model can pass decisions to experts or operate alongside downstream decision makers (Madras et al., 2018; Mozannar & Sontag, 2020; Charusaie et al., 2022; Mozannar et al., 2023; Ghoummaid & Shalit, 2024). Dynamic and strategic models show that static fairness constraints can have delayed effects on future populations, group retention, distributional dynamics, and individual adaptation (Liu et al., 2018; Zhang et al., 2019; Brown et al., 2022; Rateike et al., 2024; Dong et al., 2018; Sundaram et al., 2023; Estornell et al., 2023; Von Kügelgen et al., 2022). Bounded rationality enters our framework through the corresponding implementation problem: even after an evaluation label and harm representation are adopted, the intermediary may observe only proxy or coarse cost information and may be limited to a restricted class of policies. The residual

envy or proportionality violation then reflects both the chosen fairness objective and the information-policy constraints under which the decision rule is selected.

3 Theory

3.1 Error Burdens as Allocative Bads

We consider a binary decision problem over a finite evaluation population indexed by $i \in \{1, \ldots, n\}$. Each individual has features $x_i \in \mathcal{X}$, a sensitive attribute $a_i \in \mathcal{A}$, and an evaluation label $y_i \in \{0, 1\}$. A policy is written as

$$\pi : \mathcal{X} \times \mathcal{A} \to [0, 1], \qquad \pi(x_i, a_i) = \Pr_{\pi}(\hat{Y} = 1 \mid x_i, a_i). \tag{3.1}$$

For an observed instance (x_i, a_i, y_i), the realized-label error-event profile is

$$e_i(\pi) = \left(e_i^{\mathrm{FP}}(\pi), e_i^{\mathrm{FN}}(\pi)\right), \tag{3.2}$$

where

$$e_i^{\mathrm{FP}}(\pi) = \mathbf{1}\{y_i = 0\}\pi(x_i, a_i), \qquad e_i^{\mathrm{FN}}(\pi) = \mathbf{1}\{y_i = 1\}\left(1 - \pi(x_i, a_i)\right). \tag{3.3}$$

False positives and false negatives become error burdens through an adopted nonnegative cost representation

$$c_i = \left(c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}}\right), \qquad c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}} \geq 0. \tag{3.4}$$

The harm borne by individual i is

$$H_i(\pi) = c_i^{\mathrm{FP}} e_i^{\mathrm{FP}}(\pi) + c_i^{\mathrm{FN}} e_i^{\mathrm{FN}}(\pi) = \langle c_i, e_i(\pi) \rangle, \tag{3.5}$$

and the induced harm allocation is $H(\pi) = (H_1(\pi), \ldots, H_n(\pi))$. For cross-individual Fair Division comparisons, define

$$H_{i \leftarrow j}(\pi) = c_i^{\mathrm{FP}} e_j^{\mathrm{FP}}(\pi) + c_i^{\mathrm{FN}} e_j^{\mathrm{FN}}(\pi) = \langle c_i, e_j(\pi) \rangle. \tag{3.6}$$

Thus $H_{i \leftarrow i}(\pi) = H_i(\pi)$, while $H_{i \leftarrow j}(\pi)$ evaluates individual j's assigned error-event profile under individual i's cost representation. We also use aggregate summaries such as

$$\overline{H}(\pi) = \frac{1}{n} \sum_{i=1}^{n} H_i(\pi), \qquad H_{\max}(\pi) = \max_{i \in \{1, \ldots, n\}} H_i(\pi). \tag{3.7}$$

Lemma 3.1 (Cost-blindness of error-event profiles)**.** *For a fixed evaluation population and policy π, the profiles $\{e_i(\pi)\}_{i=1}^n$ are independent of the costs $\{c_i\}_{i=1}^n$. Changing costs can change $H(\pi)$ and $H_{i \leftarrow j}(\pi)$, but not the underlying realized-label error-event profiles.*

Proof. The quantities $e_i^{\mathrm{FP}}(\pi)$ and $e_i^{\mathrm{FN}}(\pi)$ depend only on y_i and $\pi(x_i, a_i)$. Costs enter only through inner products such as $H_i(\pi) = \langle c_i, e_i(\pi) \rangle$. □

3.2 Fair Division Diagnostics

The harm allocation $H(\pi)$ describes how a policy distributes error-induced burdens. We use two Fair Division diagnostics. Envy-freeness captures comparative grievance: individual i should not evaluate her own assigned burden as worse than individual j's assigned burden under i's own cost representation. Proportionality captures excessive-share burden: individual i should not bear more than an equal share of the total burden evaluated in i's own cost units. The directed envy gap is

$$\mathrm{EnvyGap}_{i,j}(\pi) = \left[H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi)\right]_+, \qquad [z]_+ = \max\{z, 0\}, \tag{3.8}$$

with aggregate diagnostic

$$\mathrm{EnvyGap}(\pi) = \max_{i,j \in \{1, \ldots, n\}} \mathrm{EnvyGap}_{i,j}(\pi). \tag{3.9}$$

Individual i's proportional benchmark and excess are

$$\mathrm{Share}_i(\pi) = \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi), \qquad \mathrm{PropExcess}_i(\pi) = \left[H_{i\leftarrow i}(\pi) - \mathrm{Share}_i(\pi)\right]_+, \tag{3.10}$$

with aggregate diagnostic

$$\mathrm{ProportionalExcess}(\pi) = \max_{i\in\{1,\dots,n\}} \mathrm{PropExcess}_i(\pi). \tag{3.11}$$

A policy is envy-free if $\mathrm{EnvyGap}(\pi) = 0$ and proportional if $\mathrm{ProportionalExcess}(\pi) = 0$.

Proposition 3.2 (Envy-freeness implies proportionality, but not conversely)**.** *If a policy π is envy-free, then it is proportional. The converse does not hold in general.*

Proof. If π is envy-free, then $H_{i\leftarrow i}(\pi) \le H_{i\leftarrow j}(\pi)$ for all i, j. Averaging over j gives

$$H_{i\leftarrow i}(\pi) \le \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi), \tag{3.12}$$

so $\mathrm{PropExcess}_i(\pi) = 0$ for every i. For the converse, take three individuals and suppose that, from individual 1's perspective,

$$H_{1\leftarrow 1}(\pi) = 5, \qquad H_{1\leftarrow 2}(\pi) = 4, \qquad H_{1\leftarrow 3}(\pi) = 8. \tag{3.13}$$

Individual 1 envies individual 2, but satisfies proportionality because $5 \le (5+4+8)/3$. The remaining individuals can be assigned costs that generate no violation. □

3.3 Statistical Fairness as Surrogate Conditions

Statistical fairness criteria constrain observable rates, scores, or conditional prediction behavior, while Fair Division diagnostics evaluate cost-weighted error-burden allocations. For a group $a \in \mathcal{A}$, write

$$\mathrm{FPR}_a(\pi) = \Pr_\pi(\hat{Y} = 1 \mid Y = 0, A = a), \qquad \mathrm{FNR}_a(\pi) = \Pr_\pi(\hat{Y} = 0 \mid Y = 1, A = a). \tag{3.14}$$

Demographic Parity equalizes $\Pr_\pi(\hat{Y} = 1 \mid A = a)$ across groups; Equalized Odds equalizes $\mathrm{FPR}_a(\pi)$ and $\mathrm{FNR}_a(\pi)$ across groups; Calibration constrains $\Pr(Y = 1 \mid S = u, A = a)$ for a score $S = s(X)$. When costs are group-homogeneous, so that $c_i^{\mathrm{FP}} = c_a^{\mathrm{FP}}$ and $c_i^{\mathrm{FN}} = c_a^{\mathrm{FN}}$ whenever $a_i = a$, the group-average burden decomposes as

$$\overline{H}_a(\pi) = c_a^{\mathrm{FP}} \Pr(Y = 0 \mid A = a)\mathrm{FPR}_a(\pi) + c_a^{\mathrm{FN}} \Pr(Y = 1 \mid A = a)\mathrm{FNR}_a(\pi). \tag{3.15}$$

Theorem 3.3 (Conditional surrogate validity and failure)**.** *The following two claims hold.*

1. *Suppose groups a and b have group-homogeneous costs and satisfy*

$$c_a^{\mathrm{FP}} \Pr(Y = 0 \mid A = a) = c_b^{\mathrm{FP}} \Pr(Y = 0 \mid A = b), \qquad c_a^{\mathrm{FN}} \Pr(Y = 1 \mid A = a) = c_b^{\mathrm{FN}} \Pr(Y = 1 \mid A = b). \tag{3.16}$$

   Then Equalized Odds between a and b implies $\overline{H}_a(\pi) = \overline{H}_b(\pi)$.

2. *Let M be any statistical fairness metric whose value depends only on observable quantities such as $(A, Y, S, \hat{Y})$ and not on costs. If there exist individuals i, j with $e_i(\pi) \ne e_j(\pi)$, then there exists a nonnegative cost representation under which M is unchanged and* $\mathrm{EnvyGap}(\pi) > 0$. *If the admissible cost scale is unbounded, then for every $B > 0$ there exists such a representation with* $\mathrm{EnvyGap}(\pi) \ge B$.

Proof. For the first claim, substitute Equalized Odds into the group-average burden decomposition. The coefficient equalities make the false-positive and false-negative burden components equal across the two groups. For the second claim, changing c leaves M unchanged. If, for example, $e_i^{\mathrm{FN}}(\pi) > e_j^{\mathrm{FN}}(\pi)$, set $c_i^{\mathrm{FN}} = C$ and $c_i^{\mathrm{FP}} = 0$. Then

$$H_{i\leftarrow i}(\pi) - H_{i\leftarrow j}(\pi) = C\left(e_i^{\mathrm{FN}}(\pi) - e_j^{\mathrm{FN}}(\pi)\right), \tag{3.17}$$

which is positive for $C > 0$ and arbitrarily large for large enough C. The false-positive case is identical. □

Calibration has a distinct surrogate status because it constrains score semantics before those scores are converted into error allocations. Once scores are converted into decisions, the induced false positives and false negatives enter the same harm-allocation analysis.

3.4 Bounded Rationality and Approximation Floors

Even if a decision maker targets a harm-based diagnostic directly, exact Fair Division may be unattainable under limited cost information and restricted policy classes. Let $\Pi_{\mathrm{bd}} \subseteq \Pi$ denote the implementable policy class. Let $\tilde{c} = \{\tilde{c}_i\}_{i=1}^n$ be the proxy cost representation available at selection time, and define

$$\mathcal{C}(\tilde{c}, \epsilon) = \{c : \|c_i - \tilde{c}_i\| \leq \epsilon \text{ for all } i \in \{1, \ldots, n\}\}. \tag{3.18}$$

For a Fair Division violation score $\mathrm{Viol}(\pi; c)$, the bounded-rationality benchmark is

$$\mathrm{Ach}(\tilde{c}, \epsilon; \Pi_{\mathrm{bd}}) = \inf_{\pi \in \Pi_{\mathrm{bd}}} \sup_{c \in \mathcal{C}(\tilde{c}, \epsilon)} \mathrm{Viol}(\pi; c). \tag{3.19}$$

A positive value is an approximation floor.

Theorem 3.4 (Minimal approximation floor)**.** *Fix $\epsilon \in (0, 1)$. Consider two individuals with $y_1 = 1$ and $y_2 = 0$. Suppose $\tilde{c}_1 = \tilde{c}_2 = (1, 1)$,*

$$\mathcal{C}(\tilde{c}, \epsilon) = \left\{c : c_i^{\mathrm{FP}}, c_i^{\mathrm{FN}} \in [1 - \epsilon, 1 + \epsilon] \text{ for } i \in \{1, 2\}\right\}, \tag{3.20}$$

and

$$\Pi_{\mathrm{bd}} = \left\{\pi : \Pr_{\pi}(\hat{Y} = 1 \mid i = 1) = \Pr_{\pi}(\hat{Y} = 1 \mid i = 2) = \alpha, \alpha \in [0, 1]\right\}. \tag{3.21}$$

For $\mathrm{Viol}(\pi; c) = \mathrm{EnvyGap}(\pi; c)$,

$$\inf_{\pi \in \Pi_{\mathrm{bd}}} \sup_{c \in \mathcal{C}(\tilde{c}, \epsilon)} \mathrm{EnvyGap}(\pi; c) = \epsilon. \tag{3.22}$$

Proof. Under identical treatment, $e_1^{\mathrm{FN}}(\pi) = 1 - \alpha$ and $e_2^{\mathrm{FP}}(\pi) = \alpha$. The two robust directed envy terms are

$$\sup_{c \in \mathcal{C}(\tilde{c}, \epsilon)} \mathrm{EnvyGap}_{1,2}(\pi; c) = [1 + \epsilon - 2\alpha]_+ \tag{3.23}$$

and

$$\sup_{c \in \mathcal{C}(\tilde{c}, \epsilon)} \mathrm{EnvyGap}_{2,1}(\pi; c) = [2\alpha - (1 - \epsilon)]_+. \tag{3.24}$$

The first term decreases in α, and the second increases in α. They are balanced at $\alpha = 1/2$, where both equal ϵ. □

4 Experiments

4.1 Common Experimental Design

All experiments use the ACSIncome, ACSPublicCoverage, and ACSEmployment tasks from Folktables. We use `SEX` as the sensitive attribute for auditing and group-aware post-processing, while excluding it from the predictive feature set. Each task is evaluated over multiple random seeds with stratified train, validation,

and test splits. The first four experiments use logistic regression and histogram gradient boosting as score models. The method-comparison experiment uses a shared logistic base learner to compare existing fairness interventions under a common model class.

The experiments evaluate three families of statistical surrogates: Demographic Parity, Equalized Odds, and Calibration. The DP family uses group-specific thresholds chosen through target acceptance rates. The EO family uses group-specific threshold pairs selected over a validation grid. The CAL family applies isotonic calibration, optionally mixes calibrated and original scores, and thresholds the resulting scores. Unless a subsection states otherwise, candidate policies are selected on validation according to their target surrogate subject to an accuracy-loss constraint relative to the baseline, and are evaluated once on the test set.

Harm is computed from realized false-positive and false-negative events. Each individual has a nonnegative error-cost vector (c_i^{FP}, c_i^{FN}), and a policy-induced error event contributes c_i^{FP} for a false positive and c_i^{FN} for a false negative. The heterogeneous cost family is controlled by (γ, σ), where γ governs between-group cost-scale heterogeneity and σ governs within-group FP/FN trade-off heterogeneity. Costs are normalized within each regime, allowing the analysis to isolate cost heterogeneity from changes in the overall cost scale.

We report statistical surrogate metrics and harm-allocation diagnostics together. The surrogate metrics are Accuracy, DP Gap, EO Gap, and Calibration Error. The harm diagnostics are Average Error, Tail Excess, Worst-Case Error, Envy Gap, and Proportional Excess. For pairwise policy comparisons, surrogate–burden discordance denotes the fraction of cases in which a policy improves a statistical surrogate while worsening a harm diagnostic. Deltas are computed relative to the baseline policy.

4.2 Experiment 1: Limits of Surrogate Validity under Aligned Costs

Experimental setup. The aligned-cost experiment tests a favorable case for statistical surrogates. Each surrogate family is evaluated under a cost regime deliberately aligned with the partition it regulates: DP under group-aligned costs, EO under EO-aligned costs, and Calibration under score-bin-aligned costs. This design asks whether statistical criteria become reliable fair-division surrogates when the adopted cost representation is chosen to support that interpretation.

Results and analysis. Table 1 shows that alignment improves interpretability but does not remove surrogate–burden mismatch. DP and EO generally improve their target surrogate metrics, yet the fair-division diagnostics can still deteriorate relative to the baseline. EO is the most informative family in this matched setting, with the highest surrogate-improvement rate and the lowest discordance rates. Calibration provides the weakest evidence of harm alignment, reflecting the gap between score-level calibration and realized thresholded error burdens. Aligned costs can therefore make statistical criteria meaningful, but they do not make them complete guarantees of burden fairness.

Table 1: Matched aligned-cost results. Surrogate improvement reports selected-policy surrogate-improvement rates; deltas are computed relative to the baseline. Discordance columns report within-family surrogate–fair-division discordance.

Family	Cost regime	Surrogate	Δ			Discordance	
			Surrogate	Envy	Prop.	Envy	Prop.
CAL	ScoreBinAligned	0.3889	-0.0006	0.0355	0.0316	0.4203	0.4719
DP	GroupAligned	0.7222	-0.0201	0.0766	0.0364	0.4206	0.4868
EO	EOAligned	0.9444	-0.0175	0.0557	0.0492	0.2872	0.3573

4.3 Experiment 2: Surrogate Failure under Heterogeneous Costs

Experimental setup. The heterogeneous-cost experiment moves from aligned costs to heterogeneous subjective error costs. The cost parameters range over $\{(0,0),(0.5,0.5),(1.0,0.75),(2.0,1.0),(3.0,1.25)\}$. The homogeneous setting serves as a sanity check, while the remaining settings introduce between-group

cost-scale heterogeneity and within-group FP/FN trade-off heterogeneity. For each surrogate family, the selected policy is chosen according to validation surrogate performance, with test-set harm used only for final evaluation. A selected-policy failure is recorded when the surrogate-selected policy worsens a harm diagnostic relative to the baseline.

Results and analysis. Table 2 shows that surrogate–burden failures appear as soon as heterogeneous costs are introduced. DP exhibits the most persistent instability: acceptance-rate parity can improve while average, tail, envy, or proportionality burden worsens. EO is more closely connected to FP/FN-based burden because it directly regulates conditional error rates, but it is not protected from heterogeneous-cost failure. Calibration produces smaller selected-policy failures in this setup, yet its pairwise ordering remains discordant with harm diagnostics. Across heterogeneous cost regimes, statistical surrogate ordering and harm-allocation ordering remain separated without following a simple monotone pattern in γ and σ.

Table 2: Heterogeneous-cost selected-policy failure and pairwise discordance rates. EB aggregates Average Error, Tail Excess, and Worst-Case Error; FD aggregates Envy Gap and Proportional Excess.

(γ, σ)	Family	Selected-policy failure				Pairwise discordance	
		Average Error	Tail Excess	Envy Gap	Prop. Excess	Error Burden	Fair Division
$(0, 0)$	CAL	0.2222	0.0000	0.0000	0.0000	0.1111	0.0000
	DP	0.7222	0.0000	0.0000	0.0000	0.1631	0.0000
	EO	0.7222	0.0000	0.0000	0.0000	0.1184	0.0000
$(0.5, 0.5)$	CAL	0.2222	0.2778	0.1111	0.1667	0.2426	0.2667
	DP	0.6667	0.6667	0.3889	0.3333	0.4303	0.4868
	EO	0.7222	0.7222	0.5556	0.3333	0.3539	0.2904
$(1.0, 0.75)$	CAL	0.1667	0.2222	0.1667	0.1667	0.2315	0.3139
	DP	0.6667	0.6667	0.3889	0.3889	0.4277	0.4881
	EO	0.6667	0.7222	0.6667	0.4444	0.3650	0.3087
$(2.0, 1.0)$	CAL	0.1667	0.2778	0.1111	0.2778	0.2185	0.2722
	DP	0.6111	0.5556	0.3889	0.4444	0.4330	0.5291
	EO	0.6667	0.6667	0.4444	0.2778	0.3688	0.3019
$(3.0, 1.25)$	CAL	0.1667	0.2778	0.1111	0.2222	0.2204	0.2972
	DP	0.6111	0.5556	0.5556	0.4444	0.4365	0.4881
	EO	0.6111	0.6667	0.3889	0.2778	0.3690	0.2918

4.4 Experiment 3: Axiomatic Separation of Core Harm Diagnostics

Experimental setup. The diagnostic-separation experiment asks whether the proposed harm diagnostics measure the same underlying object or distinct fairness concerns. We fix the heterogeneous regime at $(\gamma, \sigma) = (2.0, 1.0)$ with negatively correlated FP/FN cost shocks. For each task, model, and seed, we generate a broad candidate-policy set including global thresholds, group thresholds, EO-style threshold pairs, calibration-based policies, and the baseline. Each policy is evaluated by Average Error, Tail Excess, Worst-Case Error, Envy Gap, and Proportional Excess.

Results and analysis. Table 3 shows that the diagnostics are related but not interchangeable. Average Error and Tail Excess tend to rank policies similarly, but they do not always select the same top-ranked policy. Envy Gap and Proportional Excess also form a related fair-division family, yet they capture different complaints: comparative grievance and excessive-share burden. Worst-Case Error is the most clearly separated from the fair-division diagnostics. These results support a multi-diagnostic evaluation of burden fairness.

Table 3: Harm-diagnostic separation across candidate-policy rankings. The table reports Spearman correlations, mean rank disagreement, and Top-1 agreement.

Metric 1	Metric 2	Spearman correlation		Rank comparison	
		Median	Mean	Mean disagreement	Top-1 agreement
Average Error	Tail Excess	0.9761	0.9720	0.0915	0.3333
Envy Gap	Prop. Excess	0.8828	0.8506	0.1482	0.0000
Average Error	Envy Gap	0.6232	0.5936	0.2762	0.0000
Average Error	Prop. Excess	0.4801	0.4505	0.3340	0.0000
Tail Excess	Envy Gap	0.6564	0.6099	0.2717	0.0000
Tail Excess	Prop. Excess	0.5081	0.4722	0.3381	0.0000
Average Error	Worst-Case Error	0.5881	0.3382	0.3334	0.2778
Tail Excess	Worst-Case Error	0.5797	0.3699	0.3158	0.0556
Worst-Case Error	Envy Gap	0.1745	0.1686	0.4230	0.0000
Worst-Case Error	Prop. Excess	0.2032	0.1536	0.4264	0.0000

4.5 Experiment 4: Bounded Rationality and Approximation Floors

Experimental setup. The bounded-rationality experiment examines whether harm-allocation fairness can be achieved when the intermediary directly optimizes a harm-based objective but operates under bounded rationality. The true adopted harm representation is fixed at $(\gamma, \sigma) = (2.0, 1.0)$ with negatively correlated FP/FN cost shocks. Policies are selected on validation using proxy costs and evaluated on test using the true costs. The selection objective is Envy Gap, with Proportional Excess reported as an additional fair-division diagnostic.

We compare four information regimes: Oracle, Noisy, Coarse, and Noisy + Coarse. Oracle uses the true validation costs. Noisy perturbs costs by multiplicative log-noise. Coarse replaces individual costs with a small number of cost types. Noisy + Coarse combines both forms of degradation. We also compare three implementable policy classes: a global-threshold class, a group-threshold class, and a finite candidate-pool class containing the main surrogate-policy candidates.

Results and analysis. Table 4 shows that residual violation remains even when the intermediary targets Envy Gap directly. The test-set floors are positive within every policy class, which indicates an approximation floor created by restricted implementation. Selected policies also remain above these floors, showing additional selection loss from validation-based policy choice and imperfect cost information. Noisy and coarse information do not produce a simple monotone ordering; coarsening can sometimes stabilize noisy costs by averaging individual-level perturbations. The core conclusion is that harm-allocation fairness is an implementable approximation problem, not merely a matter of choosing the right metric.

4.6 Experiment 5: Comparison with Existing Fairness Methods

Experimental setup. The method-comparison experiment compares existing fairness interventions under standard statistical metrics and rank summaries derived from harm-allocation diagnostics. All methods use a shared logistic base learner, isolating the effect of the intervention mechanism from variation in model class. We compare ERM, FairBalance, FairBatch, GroupDRO, FairMixup, and FairProjection. These methods represent reweighting, batch-based fairness training, distributionally robust optimization, data augmentation, and post-processing families.

All methods are evaluated under the fixed heterogeneous harm regime $(\gamma, \sigma) = (2.0, 1.0)$. We compute an average surrogate rank from Accuracy, DP Gap, EO Gap, and Calibration Error, and an average harm rank from Average Error, Tail Excess, Envy Gap, and Proportional Excess. Rank shift is defined as $\text{RankShift} = \text{HarmRank} - \text{SurrogateRank}$.

Table 4: Bounded-rationality floors and selection loss across information regimes and policy classes.

Information	Policy class	Envy Gap			Proportional Excess		
		Selected	Floor	Excess	Selected	Floor	Excess
Coarse	Candidate pool	8.8897	5.2610	3.6287	2.9763	1.5753	1.4010
	Global threshold	8.5042	5.1822	3.3220	2.8656	1.4169	1.4487
	Group threshold	9.0892	4.5257	4.5636	2.3484	1.1285	1.2199
Noisy	Candidate pool	11.3645	5.2610	6.1034	4.2490	1.5753	2.6737
	Global threshold	10.3512	5.1822	5.1690	3.3374	1.4169	1.9205
	Group threshold	9.0425	4.5257	4.5168	2.7425	1.1285	1.6140
Noisy + Coarse	Candidate pool	8.7463	5.2610	3.4853	2.7937	1.5753	1.2184
	Global threshold	8.7848	5.1822	3.6025	3.1383	1.4169	1.7214
	Group threshold	10.3007	4.5257	5.7750	3.5836	1.1285	2.4551
Oracle	Candidate pool	8.6567	5.2610	3.3956	3.0362	1.5753	1.4609
	Global threshold	8.4751	5.1822	3.2929	3.0174	1.4169	1.6006
	Group threshold	8.5711	4.5257	4.0454	2.7528	1.1285	1.6243

Results and analysis. Table 5 shows that existing methods do not have a stable ordering across statistical and harm-based evaluation. Some methods receive similar surrogate ranks but different harm ranks, while others move sharply between the two ranking perspectives. This pattern is visible across all three ACS tasks, indicating that method comparisons change once fairness is evaluated through the allocation of error-induced burdens.

Table 5: Comparison of existing fairness methods. Surrogate rank is computed from Accuracy, DP Gap, EO Gap, and Calibration Error; harm rank is computed from Average Error, Tail Excess, Envy Gap, and Proportional Excess.

Task	Method	Statistical surrogate metrics				Average rank		
		Accuracy	DP Gap	EO Gap	CAL Error	Surrogate	Harm	Shift
Emp	ERM	0.7991	0.0146	0.0783	0.0407	1.75	2.25	0.50
	FairBalance	0.7977	0.0177	0.0902	0.0429	3.50	5.00	1.50
	FairBatch	0.7929	0.0053	0.0802	0.0471	3.25	3.50	0.25
	GroupDRO	0.7938	0.0167	0.0731	0.0550	3.25	1.50	-1.75
	FairMixup	0.7917	0.0183	0.0994	0.0628	5.50	5.25	-0.25
	FairProjection	0.7992	0.0607	0.1055	0.0407	3.50	3.50	0.00
Inc	ERM	0.7503	0.0041	0.0694	0.0807	3.25	3.75	0.50
	FairBalance	0.7656	0.0244	0.0930	0.0448	2.75	2.75	0.00
	FairBatch	0.7519	0.0386	0.0687	0.0801	3.50	4.50	1.00
	GroupDRO	0.7319	0.0191	0.0562	0.0986	3.75	3.75	0.00
	FairMixup	0.7565	0.0339	0.0765	0.0511	3.25	5.00	1.75
	FairProjection	0.7690	0.1505	0.1099	0.0807	4.25	1.25	-3.00
Cov	ERM	0.6719	0.0437	0.0532	0.0927	2.75	4.75	2.00
	FairBalance	0.6705	0.0373	0.0579	0.1176	3.75	2.50	-1.25
	FairBatch	0.6836	0.0457	0.0786	0.0989	3.50	3.00	-0.50
	GroupDRO	0.6030	0.0581	0.1176	0.1556	5.50	3.75	-1.75
	FairMixup	0.6798	0.0271	0.0414	0.0731	1.50	4.00	2.50
	FairProjection	0.6856	0.1031	0.1200	0.0927	3.75	3.00	-0.75

5 Discussion

5.1 Fairness Metrics as Representation Claims

A statistical fairness metric is not only a rate-level summary; it is also a representation claim. Reporting a smaller DP gap, EO gap, or calibration error assumes that the constrained statistical object is an adequate proxy for the burden distribution that matters in the decision context. This assumption can be reasonable when the adopted harm representation is aligned with the groups, labels, or score bins regulated by the metric. When that alignment fails, statistical improvement need not imply improvement in the allocation of error-induced burdens.

This view does not reject statistical fairness criteria. It clarifies what must be true for them to carry harm-based meaning. Demographic Parity, Equalized Odds, and Calibration compress a high-dimensional allocation problem into low-dimensional observable constraints. The harm-allocation framework makes this compression explicit: once an evaluation label and harm representation are adopted, the relevant question is whether the chosen metric faithfully tracks the resulting burden allocation.

5.2 Hidden Harm Models in Statistical Fairness

This representational role also shows why cost-blind criteria are not value-free. They may appear neutral because they do not explicitly attach weights to false-positive and false-negative events, but their harm-based interpretation still depends on an implicit harm model. Equalized Odds, for example, becomes informative about burden allocation only when the regulated error-rate components correspond to the error burdens that matter under the adopted representation. Demographic Parity and Calibration make analogous commitments through acceptance rates and score semantics.

Making harm weights explicit therefore does not introduce value judgments into an otherwise neutral evaluation. It exposes the value judgments that rate-based criteria already require when they are interpreted as fairness claims. The role of the adopted harm representation is to make this hidden step auditable: if two policies have similar statistical profiles but induce different burden allocations, then their fairness comparison depends on which harm model the evaluation is meant to represent.

5.3 Fairness Disagreement as Complaint Disagreement

Once fairness is viewed through burden allocation, disagreement among metrics becomes more than a technical disagreement about measurement. It can also reflect disagreement about which complaint should count as central. Mean error burden concerns aggregate loss, tail and worst-case burden concern concentrated harm, envy gap concerns comparative grievance, and proportional excess concerns excessive share. These are related but distinct objects, so a policy can improve one form of fairness while leaving another unresolved.

This distinction explains why no single diagnostic should be treated as a complete burden-fairness score. A decision rule that lowers average harm may still leave a small set of individuals with severe burdens, and a rule that reduces tail burden may still generate comparative complaints under individual cost representations. The proposed diagnostics separate the burden-based complaints that can remain under a given policy.

5.4 Bounded Rationality as an Approximation Problem

These complaints are also shaped by institutional constraints. Harm-allocation fairness is not determined only by the choice of a fairness objective; it also depends on the information and policies available to the intermediary. Even when an intermediary targets an envy or proportionality diagnostic directly, the attainable allocation is restricted by the cost information available at selection time and by the policy class that can be implemented. Proxy costs, coarse cost categories, global thresholds, and finite post-processing families limit the set of burden allocations that a decision rule can induce.

This changes how residual violation should be interpreted. A remaining envy gap or proportional excess need not indicate that the wrong fairness concept was chosen. It may reflect an approximation floor created

by institutional limits: the intermediary cannot observe the adopted harm representation at full resolution, cannot implement arbitrary reallocations of error events, or must select policies through validation-based procedures. Under bounded rationality, fairness evaluation is shaped by the target axiom together with the attainable region defined by information and implementation constraints.

5.5 Fairness Interventions as Burden-Allocation Mechanisms

The same perspective applies to existing fairness interventions. Such methods are often described by the statistical criteria they target, such as parity, equalized error rates, calibration, robustness, or reweighting. From the harm-allocation perspective, they also act as burden-allocation mechanisms. By changing scores, thresholds, weights, or post-processing rules, they change where false-positive and false-negative events occur, and therefore change which individuals or cost strata bear the resulting burdens.

This interpretation cautions against reading method comparisons only through statistical rankings. A method that improves a surrogate metric may still shift errors toward individuals for whom those errors are more costly, while another method with weaker surrogate performance may induce a more favorable allocation under the adopted representation. Existing fairness interventions should therefore be evaluated through their effects on observable fairness metrics and on the redistribution of decision burdens.

5.6 Limitations and Future Work

The framework is conditional on an adopted evaluation label and an adopted harm representation. The evaluation label defines the target relative to which false-positive and false-negative events are counted, and the harm representation is not an estimate of each individual's true subjective cost. The synthetic cost regimes used in the experiments are controlled stress tests for surrogate validity under heterogeneous burdens, not estimates of real-world damages.

The analysis is therefore a conditional diagnostic framework: given an evaluation target and a harm representation, it asks whether statistical criteria and existing interventions track error-burden allocation. Future work can extend this surrogate-validity analysis to multiclass, ranking, selective, and sequential settings, and develop broader diagnostics, finite-sample guarantees, and scalable tests for surrogate–burden discordance.

6 Conclusion

We studied machine-learning fairness as a conditional harm-allocation problem. Given an evaluation target and an adopted false-positive/false-negative cost representation, statistical fairness criteria can be read as surrogates for the allocation of error burdens. The analysis showed that this surrogate relation can hold under aligned cost structures, but can fail under heterogeneous adopted costs and restricted policy classes. The experiments provide controlled stress tests showing that statistical improvements can diverge from harm-allocation diagnostics. These results support fairness evaluations that make explicit the statistical criterion, harm representation, and implementation limits under which the induced burden allocation is judged.

Broader Impact Statement

This work provides a diagnostic framework for evaluating how machine-learning decision systems allocate error-induced burdens under adopted cost representations. In high-stakes settings such as credit, employment, insurance, education, and public services, pairing statistical fairness criteria with harm-allocation diagnostics can make distributive consequences more explicit.

The framework may help practitioners avoid treating rate-level improvement as a complete fairness guarantee. A policy that improves Demographic Parity, Equalized Odds, or Calibration may still redistribute errors toward individuals or cost strata for whom those errors are more burdensome under the adopted representation.

The framework could also be misused to argue against fairness evaluation itself. The framework clarifies when statistical fairness criteria function as reliable surrogates for burden allocation. Harm-allocation diagnostics should therefore be used alongside existing legal, ethical, and institutional fairness requirements.

The main risks concern false precision, institutional power, and value laundering. Cost representations are contestable inputs, not direct measurements of true individual harm. Responsible use requires documenting these representations, testing plausible alternatives, involving affected stakeholders where appropriate, and interpreting the results within domain-specific oversight.

Reproducibility

We provide all experiments reported in this paper in `Price_of_Justice.ipynb`. The notebook contains the full experimental pipeline, including data preparation, model training, policy construction, evaluation, and table generation. Running the notebook with the provided random seeds and package settings reproduces the results in the main text and appendix, up to minor numerical differences caused by platform- or library-level nondeterminism.

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, Alexandros Hollender, and Alexandros A Voudouris. Maximum nash welfare and other stories about efx. *Theoretical Computer Science*, 863: 69–85, 2021.

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, 2022.

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.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 80–89, 2020.

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.

Umang Bhaskar, AR Sricharan, and Rohit Vaish. On approximate envy-freeness for indivisible chores and mixed resources. *arXiv preprint arXiv:2012.06788*, 2020.

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, 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, 2019.

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.

Bhaskar Ray Chaudhury, Jugal Garg, and Kurt Mehlhorn. Efx exists for three agents. *Journal of the ACM*, 71(1):1–27, 2024.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

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.

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

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.

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.

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.

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, 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.

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

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.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

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.

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.

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.

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

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

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pp. 107–118. PMLR, 2018.

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.

Kanad S 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.

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.

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.

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.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

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.

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, 2022.

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, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

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.

Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In *International Conference on Machine Learning*, pp. 43258–43288. PMLR, 2023.

A Proofs

A.1 Notation

This appendix uses the notation from Section 3. The finite evaluation population is indexed by $i \in \{1, \ldots, n\}$, and each individual has an observed instance (x_i, a_i, y_i) with $y_i \in \{0, 1\}$. A possibly randomized policy is

$$\pi : \mathcal{X} \times \mathcal{A} \to [0, 1], \qquad \pi(x_i, a_i) = \Pr_{\pi}(\widehat{Y} = 1 \mid x_i, a_i).$$

The realized-label error-event profile of individual i is

$$e_i(\pi) = \big(e_i^{FP}(\pi), e_i^{FN}(\pi)\big),$$

where

$$e_i^{FP}(\pi) = \mathbf{1}\{y_i = 0\}\pi(x_i, a_i), \qquad e_i^{FN}(\pi) = \mathbf{1}\{y_i = 1\}\big(1 - \pi(x_i, a_i)\big).$$

When π is clear from context, we write e_i^{FP}, e_i^{FN}, and e_i.

Each individual has an adopted nonnegative error-cost vector

$$c_i = \big(c_i^{FP}, c_i^{FN}\big), \qquad c_i^{FP}, c_i^{FN} \geq 0.$$

The own and cross-evaluated harms are

$$H_i(\pi) = \langle c_i, e_i(\pi) \rangle = c_i^{FP} e_i^{FP}(\pi) + c_i^{FN} e_i^{FN}(\pi),$$

and

$$H_{i \leftarrow j}(\pi) = \langle c_i, e_j(\pi) \rangle = c_i^{FP} e_j^{FP}(\pi) + c_i^{FN} e_j^{FN}(\pi).$$

Thus $H_{i \leftarrow i}(\pi) = H_i(\pi)$.

The directed envy gap is

$$\mathrm{EnvyGap}_{i,j}(\pi) = \big[H_{i \leftarrow i}(\pi) - H_{i \leftarrow j}(\pi)\big]_+, \qquad [z]_+ = \max\{z, 0\},$$

with aggregate diagnostic

$$\mathrm{EnvyGap}(\pi) = \max_{i,j \in \{1,\ldots,n\}} \mathrm{EnvyGap}_{i,j}(\pi).$$

The proportional benchmark and excess are

$$\mathrm{Share}_i(\pi) = \frac{1}{n} \sum_{j=1}^{n} H_{i \leftarrow j}(\pi), \qquad \mathrm{PropExcess}_i(\pi) = \big[H_{i \leftarrow i}(\pi) - \mathrm{Share}_i(\pi)\big]_+,$$

and

$$\mathrm{ProportionalExcess}(\pi) = \max_{i \in \{1,\ldots,n\}} \mathrm{PropExcess}_i(\pi).$$

For surrogate results, $M(\pi)$ denotes a cost-blind statistical metric depending only on observable quantities such as $(A, Y, S, \widehat{Y})$. For bounded-rationality results, $\widetilde{c}$ denotes the proxy cost representation, $C(\widetilde{c}, \epsilon)$ the compatible cost set, and Π_{bd} the implementable policy class.

A.2 Proof of Lemma 3.1

Proof. Fix the evaluation population and policy π. By definition,

$$e_i^{FP}(\pi) = \mathbf{1}\{y_i = 0\}\pi(x_i, a_i), \qquad e_i^{FN}(\pi) = \mathbf{1}\{y_i = 1\}\big(1 - \pi(x_i, a_i)\big).$$

These quantities depend only on y_i and $\pi(x_i, a_i)$, not on c_i. Costs enter only through

$$H_i(\pi) = \langle c_i, e_i(\pi)\rangle \qquad \text{and} \qquad H_{i\leftarrow j}(\pi) = \langle c_i, e_j(\pi)\rangle.$$

Thus changing $c = \{c_i\}_{i=1}^n$ can change harms and fair-division diagnostics, but not the underlying error-event profiles $\{e_i(\pi)\}_{i=1}^n$. □

A.3 Proof of Proposition 3.2

Proof. Suppose π is envy-free. Then $H_{i\leftarrow i}(\pi) \le H_{i\leftarrow j}(\pi)$ for all i, j. Fixing i and averaging over j gives

$$H_{i\leftarrow i}(\pi) \le \frac{1}{n}\sum_{j=1}^{n} H_{i\leftarrow j}(\pi) = \mathrm{Share}_i(\pi).$$

Hence $\mathrm{PropExcess}_i(\pi) = 0$ for every i, so $\mathrm{ProportionalExcess}(\pi) = 0$.

The converse fails. Consider three individuals with $y_1 = y_2 = y_3 = 0$ and error-event profiles

$$e_1(\pi) = (0.5, 0), \qquad e_2(\pi) = (0.4, 0), \qquad e_3(\pi) = (0.8, 0).$$

Let

$$c_1 = (10, 0), \qquad c_2 = (0, 0), \qquad c_3 = (0, 0).$$

Then

$$H_{1\leftarrow 1}(\pi) = 5, \qquad H_{1\leftarrow 2}(\pi) = 4, \qquad H_{1\leftarrow 3}(\pi) = 8.$$

Therefore

$$\mathrm{EnvyGap}_{1,2}(\pi) = [5 - 4]_+ = 1,$$

so π is not envy-free. However,

$$H_{1\leftarrow 1}(\pi) = 5 \le \frac{5 + 4 + 8}{3} = \mathrm{Share}_1(\pi),$$

and individuals 2 and 3 have zero cross-evaluated harms because $c_2 = c_3 = (0, 0)$. Thus $\mathrm{ProportionalExcess}(\pi) = 0$, so π is proportional. □

A.4 Proof of Theorem 3.3

Proof. For the validity claim, suppose groups a and b have group-homogeneous costs. For $g \in \{a, b\}$,

$$H_g(\pi) = c_g^{FP}\Pr(Y = 0 \mid A = g)FPR_g(\pi) + c_g^{FN}\Pr(Y = 1 \mid A = g)FNR_g(\pi).$$

Equalized Odds gives

$$FPR_a(\pi) = FPR_b(\pi), \qquad FNR_a(\pi) = FNR_b(\pi).$$

The coefficient equalities

$$c_a^{FP}\Pr(Y = 0 \mid A = a) = c_b^{FP}\Pr(Y = 0 \mid A = b),$$

and

$$c_a^{FN}\Pr(Y = 1 \mid A = a) = c_b^{FN}\Pr(Y = 1 \mid A = b)$$

therefore equalize both the false-positive and false-negative burden components. Hence $H_a(\pi) = H_b(\pi)$.

For the failure claim, let $M(\pi)$ be a statistical metric depending only on observable quantities such as $(A, Y, S, \widehat{Y})$, and not on c. Fix the evaluation population and policy π. Changing c leaves $M(\pi)$ unchanged.

Suppose $e_i(\pi) \neq e_j(\pi)$ for some i, j. Then there exist an ordered pair $(u, v) \in \{(i, j), (j, i)\}$ and an error type $r \in \{FP, FN\}$ such that

$$e_u^r(\pi) - e_v^r(\pi) = \Delta > 0.$$

Let r' be the other error type and set

$$c_u^r = C, \qquad c_u^{r'} = 0,$$

with $C > 0$. The remaining costs may be chosen arbitrarily subject to nonnegativity. Then

$$H_{u\leftarrow u}(\pi) - H_{u\leftarrow v}(\pi) = c_u^r\big(e_u^r(\pi) - e_v^r(\pi)\big) + c_u^{r'}\big(e_u^{r'}(\pi) - e_v^{r'}(\pi)\big) = C\Delta.$$

Thus

$$\mathrm{EnvyGap}(\pi) \geq \mathrm{EnvyGap}_{u,v}(\pi) = C\Delta > 0,$$

while $M(\pi)$ is unchanged. If the admissible cost scale is unbounded, choosing $C \geq B/\Delta$ gives $\mathrm{EnvyGap}(\pi) \geq B$ for any $B > 0$. □

A.5 Proof of Theorem 3.4

Proof. Any $\pi \in \Pi_{\mathrm{bd}}$ is represented by some $\alpha \in [0, 1]$ such that

$$\Pr_\pi(\widehat{Y} = 1 \mid i = 1) = \Pr_\pi(\widehat{Y} = 1 \mid i = 2) = \alpha.$$

Since $y_1 = 1$ and $y_2 = 0$,

$$e_1^{FN}(\pi) = 1 - \alpha, \qquad e_2^{FP}(\pi) = \alpha.$$

The two nonzero directed envy terms are

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

and

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

Maximizing the first expression over $C(\widetilde{c}, \epsilon)$ uses $c_1^{FN} = 1 + \epsilon$ and $c_1^{FP} = 1 - \epsilon$, giving

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

Maximizing the second uses $c_2^{FP} = 1 + \epsilon$ and $c_2^{FN} = 1 - \epsilon$, giving

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

The diagonal envy terms are zero, so

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

Let this maximum be $f(\alpha)$. If $\alpha \leq 1/2$, then $1 + \epsilon - 2\alpha \geq \epsilon$, so $f(\alpha) \geq \epsilon$. If $\alpha \geq 1/2$, then $2\alpha - (1 - \epsilon) \geq \epsilon$, so again $f(\alpha) \geq \epsilon$. Hence $f(\alpha) \geq \epsilon$ for all $\alpha \in [0, 1]$.

At $\alpha = 1/2$,

$$\big[1 + \epsilon - 2\alpha\big]_+ = \big[2\alpha - (1 - \epsilon)\big]_+ = \epsilon.$$

Therefore

$$\inf_{\pi\in\Pi_{\mathrm{bd}}} \sup_{c\in C(\widetilde{c},\epsilon)} \mathrm{EnvyGap}(\pi; c) = \inf_{\alpha\in[0,1]} f(\alpha) = \epsilon.$$

□

B Robustness and Sensitivity Analyses

B.1 Common Experimental Design

All sensitivity analyses follow the task, model, seed, split, and candidate-policy generation protocol used in the main surrogate-failure experiment unless otherwise stated. Candidate policies come from the same BASE, DP, EO, and calibration families. DP, EO, and calibration policies are selected on validation data by their corresponding surrogate metrics subject to the applicable feasibility or accuracy rule, and harm-allocation diagnostics are evaluated on the test set under the specified adopted cost representation.

The default heterogeneous cost regime is $(\gamma, \sigma, \rho) = (2.0, 1.0, -1.0)$. Each analysis changes one component of this design: the FP–FN cost-correlation structure, bounded cost scale, cost-stratum resolution, intermediary information regime, implementable policy class, or policy-selection hyperparameters. This isolates whether the surrogate–harm relationship depends on a specific modeling or implementation choice.

For the correlation, bounded-cost, and policy-selection analyses, selected-policy failure rates report whether a surrogate-improving selected policy worsens a harm diagnostic relative to BASE. Pairwise discordance reports whether, within a policy family, better surrogate values induce worse harm values. Error discordance aggregates average, tail, and worst-case error diagnostics, while fair-division discordance aggregates envy and proportionality diagnostics.

B.2 Sensitivity to Cost-Correlation Structure

Experimental setup. We vary the FP–FN cost-correlation parameter while fixing $(\gamma, \sigma) = (2.0, 1.0)$. Negative correlations induce trade-offs between false-positive and false-negative costs, while positive correlations make the two costs move together. The same latent cost shocks are reused across correlation settings, and costs are normalized to keep the average total cost scale comparable.

Results and analysis. Table 6 shows that changing the FP–FN correlation affects the severity of selected-policy failures but does not align surrogate ordering with harm-allocation ordering. Calibration is the least unstable family, yet its discordance remains nonzero. DP and EO exhibit stronger selected-policy failures and persistent ranking disagreement. Disagreement persists across trade-off, independent, and common-scale cost structures, without a simple monotone response to the sign of the correlation.

B.3 Sensitivity to Bounded Cost Scales

Experimental setup. We bound the normalized FP and FN cost components while fixing $(\gamma, \sigma, \rho) = (2.0, 1.0, -1.0)$. For each bound in $\{8, 4, 2, 1.5\}$, costs are clipped to $[1/B, B]$ and then renormalized to preserve comparable average scale. Smaller bounds progressively remove extreme cost values without changing the task, model, policy set, or latent cost-shock structure.

Results and analysis. Table 7 shows that clipping extreme cost values attenuates some failures but does not restore alignment. Calibration remains the most stable family, while DP and EO continue to produce selected-policy failures and ranking discordance under bounded costs. The mismatch therefore does not rely on unbounded or extreme cost magnitudes.

B.4 Sensitivity to Cost-Stratum Resolution

Experimental setup. We vary the resolution of the cost-stratum representation used by the stratified fair-division diagnostics. Cost types are constructed from a two-dimensional quantile partition over the FP–FN trade-off, $\log(c_i^{FP}) - \log(c_i^{FN})$, and the overall cost scale, $\frac{1}{2}\{\log(c_i^{FP}) + \log(c_i^{FN})\}$. For each resolution, Stratified Envy Burden Gap and Proportional Burden Excess are recomputed for every candidate policy and compared against the reference ranking using rank-correlation, rank-disagreement, Top-1 agreement, and mean rank displacement.

Table 6: Sensitivity to FP–FN cost correlation.

ρ	Family	Fail		Discordance	
		Average Error	Envy	Error	Fair Division
-1.0	CAL	0.1111	0.0556	0.2222	0.3111
	DP	0.5556	0.2778	0.4065	0.4788
	EO	0.6667	0.6667	0.3737	0.2856
-0.5	CAL	0.1667	0.0556	0.2389	0.4056
	DP	0.6111	0.5556	0.4180	0.4735
	EO	0.7222	0.3889	0.3758	0.2841
0.0	CAL	0.1667	0.1111	0.2333	0.3889
	DP	0.6111	0.3889	0.4030	0.4815
	EO	0.7222	0.6111	0.3699	0.3142
0.5	CAL	0.1111	0.1111	0.2204	0.3389
	DP	0.7222	0.4444	0.3995	0.4762
	EO	0.7222	0.5556	0.3718	0.3311
1.0	CAL	0.1667	0.1111	0.2963	0.2528
	DP	0.7222	0.5000	0.4242	0.5172
	EO	0.6667	0.5000	0.3690	0.4361

Table 7: Sensitivity to bounded cost scales.

Bound	Family	Fail		Discordance	
		Average Error	Envy	Error	Fair Division
8	CAL	0.1111	0.1111	0.2259	0.3167
	DP	0.6667	0.5556	0.3254	0.5344
	EO	0.7222	0.4444	0.2822	0.2642
4	CAL	0.1667	0.1111	0.2167	0.3639
	DP	0.6667	0.4444	0.3245	0.4907
	EO	0.7222	0.6667	0.2789	0.3114
2	CAL	0.2222	0.1111	0.2574	0.3389
	DP	0.6667	0.4444	0.3228	0.4365
	EO	0.7222	0.6111	0.3283	0.3752
1.5	CAL	0.2222	0.1111	0.2407	0.2944
	DP	0.7222	0.5556	0.3325	0.4960
	EO	0.7222	0.6111	0.3336	0.3918

Results and analysis. Table 8 shows that the stratum-based diagnostics preserve broad policy rankings across nearby resolutions. Envy Gap is comparatively stable, while Proportional Excess is more sensitive under finer partitioning. The weaker Top-1 agreement indicates that the identity of the single minimum-violation policy can change even when the broader ordering is preserved. The diagnostics are therefore best read as comparative ranking tools that support policy evaluation without serving as exact policy-selection oracles.

B.5 Sensitivity to Information Regimes and Policy Classes

Experimental setup. We vary the intermediary's cost information and implementable policy class under the same true cost regime. Policies are selected by minimizing validation-set Envy Gap under the information

Table 8: Sensitivity to cost-stratum resolution.

Resolution	Metric	Spearman	Rank Disagree.	Top-1 Agree.	Mean Rank Diff.	Eff. Strata
20	Envy Gap	0.8073	0.1687	0.0000	22.5174	17.2222
	Prop Excess	0.8796	0.1349	0.2222	17.5339	17.2222
40	Envy Gap	0.8233	0.1558	0.1667	21.1024	31.5556
	Prop Excess	0.8934	0.1194	0.1111	15.3121	31.5556
80	Envy Gap	1.0000	0.0000	1.0000	0.0000	55.5556
	Prop Excess	1.0000	0.0000	1.0000	0.0000	55.5556
160	Envy Gap	0.7987	0.1639	0.2778	23.3893	100.0000
	Prop Excess	0.7284	0.1950	0.0556	23.4643	100.0000

available to the intermediary and evaluated on the test set under true costs. The information regimes include oracle costs, noisy costs, coarse cost types, and their combination. The policy classes include global thresholds, group thresholds, and the finite candidate pool formed by BASE, DP, EO, and calibration candidates. Selection loss is measured relative to the best attainable test violation within the same finite class.

Results and analysis. Table 9 shows that harm-based selection remains limited even when cost information is accurate. More expressive policy classes can improve the attainable floor, but they do not guarantee better selected policies. Coarse and noisy cost representations increase selection loss most visibly when the policy class is richer. These patterns match the bounded-rationality view: realized harm allocation depends jointly on information quality, policy-class expressiveness, and validation-based selection.

Table 9: Sensitivity to information regimes and policy classes.

Info	Policy Class	Envy Gap		
		Selected	Best	Excess
Oracle	Global threshold	0.9628	0.6372	0.3257
	Group threshold	0.9608	0.5094	0.4514
	Candidate pool	0.9920	0.5094	0.4826
Noisy	Global threshold	1.0451	0.6372	0.4080
	Group threshold	0.9481	0.5094	0.4387
	Candidate pool	0.9623	0.5094	0.4529
Coarse	Global threshold	1.1293	0.6372	0.4921
	Group threshold	1.0826	0.5094	0.5732
	Candidate pool	1.1128	0.5094	0.6035
Noisy + Coarse	Global threshold	1.1179	0.6372	0.4807
	Group threshold	1.1245	0.5094	0.6151
	Candidate pool	1.1107	0.5094	0.6013

B.6 Sensitivity to Policy-Selection Hyperparameters

Experimental setup. We vary the validation-selection rule used to choose DP, EO, and calibration policies while holding the true cost representation fixed. The default configuration uses the main selection rule with a validation accuracy-loss tolerance of 0.02. StrictAcc imposes a tighter validation accuracy floor, LooseAcc allows a larger accuracy loss, CoarseGrid uses a reduced threshold and calibration grid, and FineGrid uses a denser grid. Selected policies are evaluated on the test set under the same true cost representation.

Results and analysis. Table 10 shows that selection rules affect selected-policy failures but leave the broader ranking mismatch intact. Calibration is comparatively stable across configurations, while DP and EO are more sensitive to accuracy floors and grid resolution. Discordance is less affected than selected-policy failure, indicating that hyperparameters can change the selected policy without removing the surrogate–harm ordering disagreement.

Table 10: Sensitivity to policy-selection hyperparameters.

Config	Family	Fail		Discordance	
		Average Error	Envy	Error	Fair Division
Default	CAL	0.1667	0.0556	0.4042	0.4648
	DP	0.5556	0.3333	0.4374	0.5542
	EO	0.7222	0.2778	0.3846	0.2983
StrictAcc	CAL	0.1667	0.0556	0.4042	0.4648
	DP	0.5000	0.2778	0.4374	0.5542
	EO	0.5000	0.1667	0.3846	0.2983
LooseAcc	CAL	0.1667	0.0556	0.4042	0.4648
	DP	0.4444	0.2778	0.4374	0.5542
	EO	0.7778	0.5556	0.3846	0.2983
CoarseGrid	CAL	0.1667	0.0556	0.3924	0.4654
	DP	0.5556	0.2778	0.4136	0.4907
	EO	0.1111	0.1111	0.3850	0.2932
FineGrid	CAL	0.1667	0.0556	0.4092	0.4695
	DP	0.7778	0.2778	0.4534	0.5519
	EO	0.6667	0.5000	0.3834	0.2997

B.7 Summary of Sensitivity Analyses

The sensitivity analyses separate variation in modeling and implementation choices from the qualitative surrogate–harm pattern. Changes in cost correlation, bounded cost scale, diagnostic resolution, information regime, policy class, and selection rule affect the severity and form of selected-policy failure, but they do not remove the divergence between statistical surrogate improvement and harm-allocation improvement. Alternative cost representations and bounded scales leave ranking discordance in place; stratum-resolution changes preserve broad policy-level comparisons while limiting exact policy-selection claims; and information and hyperparameter variations show that realized violations also depend on bounded implementation.