HUMAN–AI DISPARITY DISTORTION: A COMPONENT-REFERENCED DIAGNOSTIC FOR HYBRID DECISION SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness audits of hybrid human–AI systems often examine disparity in human judgments, algorithmic predictions, or final hybrid outcomes. Taken alone, these quantities do not distinguish the disparity implied by the component systems from non-additive disparity associated with their integration rule. We study this distinction through a component-referenced diagnostic. For an integration rule f_λ combining human probability p_H and algorithmic probability p_A, additive pooling $L_\lambda = (1-\lambda)p_H + \lambda p_A$ yields group disparity exactly equal to $(1-\lambda)D_H + \lambda D_A$ by linearity of expectation. We define the Human–AI Disparity Distortion Index (HDDI) as the difference between the disparity produced by f_λ and this additive reference. HDDI is identically zero under linear pooling; more generally, nonzero HDDI reflects group-differential departure from additive integration rather than nonlinearity alone. We evaluate HDDI across three complementary settings. On ACSIncome, controlled simulated human–AI pairings establish its behavior across models, groups, and integration weights. On Civil Comments, HDDI extends to identity-sensitive text moderation and is often more pronounced along a non-toxic, false-positive-relevant pathway; finite-sample recovery improves substantially with sample size, with median RMSE decreasing from 0.00222 at $n = 500$ to 0.000214 at $n = 15{,}000$. On COMPAS, nonzero interior HDDI persists when human probabilities are derived from observed judgments rather than simulation, reaching $\max_\lambda |\mathrm{HDDI}_f(\lambda)| = 0.00551$, although its magnitude depends on how human judgments are aggregated. Across all three settings, HDDI complements rather than replaces conventional disparity and fairness measures.

1 INTRODUCTION

Machine-learning systems increasingly inform rather than replace human decision makers. In lending, content moderation, and pretrial risk assessment, final decisions often combine an algorithmic prediction with human judgment that may weigh or override it (Kleinberg et al., 2018). This division of labor is often deliberate: research on human–AI complementarity treats the combined system as the relevant object of optimization (Bansal et al., 2021), making integration an object of analysis in its own right (Vaccaro et al., 2024). Yet selective compliance and other integration policies can yield hybrid disparities that differ from those of either component (Donahue et al., 2022; Gillis et al., 2022). This motivates our central question: *How can we quantify whether, and by how much, group disparity under human–AI probability integration departs from the disparity jointly implied by the human and algorithmic component predictions?*

Conventional group-fairness measures characterize properties of a realized predictor or decision rule, including differences in positive prediction rates, error rates, and calibration across groups (Hardt et al., 2016; Mitchell et al., 2021). Applied separately to human, algorithmic, and hybrid predictions, they quantify the disparity level of each system. The final hybrid disparity alone, however, does not by itself distinguish disparity already implied by the component predictions from additional deviation associated with their integration. The same final disparity can arise from different component disparities and therefore represent different deviations from a component-derived reference. Work on bias amplification similarly demonstrates the value of interpreting model behav-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ior relative to an explicit reference rather than interpreting output disparity in isolation (Zhao et al., 2017; Wang & Russakovsky, 2021). For hybrid systems, this motivates a reference constructed directly from the component predictions being integrated.

We introduce a component-referenced diagnostic based on additive probability pooling. Given the same human and algorithmic probabilities and the same integration weight, additive pooling produces a group disparity determined exactly by the component disparities and their weights. The Human–AI Disparity Distortion Index (HDDI) is the difference between the disparity produced by a specified integration rule and this additive reference. HDDI therefore distinguishes component-implied disparity from non-additive disparity deviation at the group level. Nonzero HDDI reflects a group-differential departure from additive pooling rather than nonlinearity alone. This quantity is descriptive rather than normative: it neither determines whether a system is fair nor identifies the causal effect of algorithmic assistance. Section 3 formalizes the reference and establishes the properties governing HDDI's interpretation.

We evaluate HDDI through a progressive three-stage validation design. ACSIncome (Ding et al., 2021) provides a controlled setting for testing its construct behavior across models, groups, and integration conditions. Civil Comments (Borkan et al., 2019) extends the analysis to identity-sensitive text moderation and examines conditional disparity distortion and finite-sample recoverability as subgroup support varies. Finally, in the COMPAS setting, building on prior work on human and algorithmic recidivism prediction (Dressel & Farid, 2018), we use repeated observed human judgments from Lin et al. (2020). Our contributions are threefold. First, we introduce a component-referenced approach that distinguishes disparity implied by the component predictions from non-additive disparity deviation under their integration. Second, we treat this deviation as an estimand whose empirical usefulness depends on statistical recoverability, and characterize how recovery varies with effect magnitude and effective subgroup support. Third, we evaluate the diagnostic across controlled, identity-sensitive, and observed-human settings, providing evidence of transfer while identifying an important boundary: its magnitude can depend on how human probability judgments are operationalized. In a nutshell, these contributions provide a statistical diagnostic for relating component-level and final-system disparity while keeping that diagnosis distinct from causal attribution and normative fairness assessment.

2 Related Work

Human–AI aggregation. Human–AI decision systems pose an aggregation problem: how should human and algorithmic information be combined? Appropriate reliance depends on perceived system capability and task demands (Lee & See, 2004), while human judgments and machine predictions may exhibit complementary error structures (Kleinberg et al., 2018). Existing methods combine calibrated model probabilities with class-level human predictions using human error patterns (Kerrigan et al., 2021), or account for human and machine confidence within a Bayesian framework (Steyvers et al., 2022). Algorithmic assistance does not necessarily improve human decisions (Green & Chen, 2019), and complementary performance depends on component capabilities, interaction design, and task conditions (Bansal et al., 2021; Vaccaro et al., 2024). This literature primarily examines prediction, decision quality, calibration, and complementarity. We examine how hybrid group disparity relates to the additive component reference determined by the component disparities, conditional on their probabilities and a specified integration rule.

Fairness in human–AI collaboration. Fairness research increasingly examines the combined human–AI system. Donahue et al. (2022) analyze conditions for complementarity and fairness properties of combined classifiers, while Gillis et al. (2022) show that human involvement can alter the relationship between algorithmic predictions and disparities in final decisions. Selective compliance further motivates fairness guarantees that hold across admissible patterns of adherence (Ge et al., 2024). Behavioral studies examine how automation bias, selective adherence, and alignment between human and algorithmic biases affect reliance and perceived fairness (Alon-Barkat & Busuioc, 2023; Zipperling et al., 2025). Gao et al. (2026) study fair decision utility when human groups differ in cognitive capacity. These studies establish that collaborative fairness cannot generally be inferred from the algorithm alone. However, final disparity can combine disparity implied by the component predictions with deviation from a specified component-derived reference. HDDI addresses this

narrower measurement problem by distinguishing the component-derived reference from the hybrid system's departure from it.

Disparity decomposition and reference-based diagnostics. Explicit references have a long history in the interpretation of group differences. Blinder–Oaxaca decompositions partition mean outcome gaps into components defined relative to a counterfactual reference structure (Oaxaca, 1973; Blinder, 1973; Jann, 2008). In algorithmic fairness, Pan et al. (2021) decompose model disparity into contributions associated with causal pathways. Bias-amplification measures instead compare associations in model predictions with those in a reference data distribution (Zhao et al., 2017), with subsequent work emphasizing direction, group base rates, and normative interpretation (Wang & Russakovsky, 2021; Tokas et al., 2025). Although these approaches differ in their estimands and assumptions, they illustrate how a gap's interpretation depends on its reference. HDDI adopts this principle for hybrid integration: it compares an integration operator's disparity with additive pooling of the same human and algorithmic probabilities at the same weight. The resulting accounting is defined by the relationship between the hybrid operator and its component-derived additive reference, rather than by covariate contributions, causal pathways, or prediction–reference-distribution differences. This accounting is noncausal: HDDI does not identify differential treatment, selective reliance, or another behavioral mechanism, nor does it determine whether the system is normatively fair.

Probability pooling and the component reference. Linear opinion pooling forms a convex combination of component probabilities (Stone, 1961), while broader work studies logarithmic and other pooling operators and their properties (Genest et al., 1984; Genest & Zidek, 1986; DeGroot & Mortera, 1991). When group disparity is defined as a difference in conditional expectations, additive pooling yields exactly the weighted combination of component disparities. This identity provides an exact component-derived reference determined by the human and algorithmic disparities and their integration weight. It does not imply that additive pooling is predictively optimal or normatively preferable: nontrivial linear combinations of calibrated forecasts need not preserve calibration or sharpness (Ranjan & Gneiting, 2010). Our empirical analysis uses nonlinear log-odds pooling as the primary integration operator and additive pooling as the reference. HDDI measures whether departure from this reference differs across groups, rather than treating nonlinearity itself as distortion.

Statistical uncertainty and recoverability. A well-defined disparity quantity need not be reliably estimable from finite data. Ethayarajh (2020) develops confidence bounds for classification bias, showing how limited samples can leave apparent disparities statistically inconclusive. Barrainkua et al. (2024) develops an uncertainty-aware framework for comparing performance and fairness under estimation uncertainty. Statistical actionability further distinguishes disparity magnitude from evidentiary reliability and subgroup adequacy when translating fairness evidence into deployment decisions (Fan & Wang, 2026). HDDI poses a related but distinct estimation problem: its contrast between group-conditional quantities must be recovered from finite, potentially imbalanced samples. Recovery depends on both deviation magnitude and available subgroup support, so a nonzero population value alone does not imply that the deviation can be estimated with useful precision in a finite sample. We therefore examine estimation error, sign and magnitude recovery, and effective subgroup support. This analysis distinguishes what HDDI measures from when the available data support a reliable diagnosis.

3 Component-Referenced Disparity in Human–AI Integration

We formalize hybrid group disparity as a component-referenced contrast. Let $p_H(x)$ and $p_A(x)$ denote human and algorithmic probability judgments for the same instance, and let $\lambda \in [0, 1]$ index their integration. Our central comparison holds the component probabilities and integration weight fixed while comparing the disparity produced by a general integration operator with the disparity implied by additive pooling of the same inputs. Figure 1 summarizes the construction. The same p_H, p_A, and λ enter two paths: an additive path defining the component-derived reference and a hybrid path producing the disparity under a specified integration operator. HDDI is the signed difference between these disparities. Equivalently, it is the between-group difference in the mean individual-level departure of the hybrid operator from additive pooling.

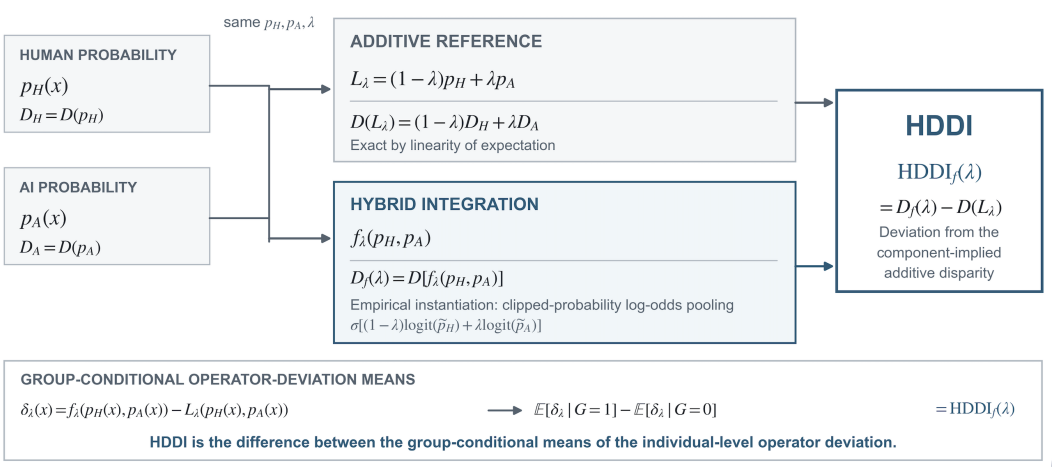


Figure 1: Component-referenced formulation of HDDI. The same human and algorithmic probabilities p_H and p_A and integration weight λ enter an additive reference and a general hybrid integration operator. Additive pooling yields disparity $D(L_\lambda) = (1-\lambda)D_H + \lambda D_A$, while the hybrid operator yields $D_f(\lambda)$. HDDI is their signed difference. Equivalently, HDDI is the difference between the group-conditional means of the individual-level operator deviation $\delta_\lambda = f_\lambda - L_\lambda$. Nonzero HDDI therefore requires the departure from additive pooling to differ on average across groups; nonlinearity alone is insufficient.

3.1 Component Reference and HDDI

Let $G \in \{0, 1\}$ denote the comparison groups, with $\Pr(G = g) > 0$ for both g. For any integrable probability-valued prediction q, define signed group disparity as $D(q) = \mathbb{E}[q(X) \mid G = 1] - \mathbb{E}[q(X) \mid G = 0]$, with $D_H = D(p_H)$ and $D_A = D(p_A)$. A general integration operator $f_\lambda : [0, 1]^2 \to [0, 1]$ produces $p_f^{(\lambda)}(x) = f_\lambda(p_H(x), p_A(x))$, with hybrid disparity $D_f(\lambda) = D(p_f^{(\lambda)})$. Here λ indexes an integration path rather than a causal treatment intensity. We define the additive component reference as $L_\lambda(p_H, p_A) = (1 - \lambda)p_H + \lambda p_A$, corresponding to a two-source linear opinion pool (Stone, 1961; Ranjan & Gneiting, 2010).

Proposition 1 (Additive-reference identity). For any fixed $\lambda \in [0, 1]$,

$$D(L_\lambda) = (1 - \lambda)D_H + \lambda D_A. \tag{1}$$

The identity follows from linearity of conditional expectation and requires no independence assumption between p_H, p_A, and G. It therefore provides an exact component-derived reference without treating additive pooling as predictively optimal or normatively preferable. The HDDI is

$$\mathrm{HDDI}_f(\lambda) = D_f(\lambda) - D(L_\lambda). \tag{2}$$

Equivalently, $D_f(\lambda) = D(L_\lambda) + \mathrm{HDDI}_f(\lambda)$. Thus, HDDI is the signed non-additive departure of hybrid disparity from the component-implied additive reference. This accounting is descriptive rather than causal or normative. Define the individual-level operator deviation as $\delta_\lambda(x) = f_\lambda(p_H(x), p_A(x)) - L_\lambda(p_H(x), p_A(x))$.

Proposition 2 (Operator-deviation representation). For any fixed $\lambda \in [0, 1]$,

$$\mathrm{HDDI}_f(\lambda) = \mathbb{E}[\delta_\lambda(X) \mid G = 1] - \mathbb{E}[\delta_\lambda(X) \mid G = 0]. \tag{3}$$

Equation equation 3 corresponds to the lower panel of Figure 1. Nonlinearity alone is insufficient for nonzero HDDI: a nonlinear operator may depart substantially from additive pooling at the individual level while yielding zero HDDI when the group-conditional mean departures are equal. HDDI is therefore jointly determined by the integration operator and the group-specific distribution of (p_H, p_A). Its sign indicates which group has the larger mean operator deviation; it does not by itself imply harm, unfairness, or an increase in absolute disparity.

3.2 Diagnostic Properties and Empirical Extensions

The component-reference construction extends directly to conditional comparisons. For an event C with positive support in both groups, define $D_C(q) = \mathbb{E}[q \mid G = 1, C] - \mathbb{E}[q \mid G = 0, C]$ and $\mathrm{HDDI}_{f,C}(\lambda) = D_C(f_\lambda) - D_C(L_\lambda)$. Because conditional expectation remains linear, additive pooling continues to provide the exact component-derived reference within C. In Civil Comments, we use $C = \{Y = 0\}$ to characterize HDDI among non-toxic comments. This quantity remains a contrast in mean predicted probabilities and is therefore not itself a thresholded false-positive-rate difference.

In finite samples, let $\widehat{D}_C(q)$ denote the corresponding difference in group-condition sample means, yielding the plug-in estimator $\widehat{\mathrm{HDDI}}_{f,C}(\lambda) = \widehat{D}_C(f_\lambda) - \widehat{D}_C(L_\lambda)$. Its precision depends on the effective support available in the relevant group-condition cells, motivating the finite-sample recovery analysis in Section 4 (Ethayarajh, 2020; Barrainkua et al., 2024).

Two identities provide internal controls for the empirical implementation. If $f_\lambda = L_\lambda$, then $\mathrm{HDDI}_f(\lambda) = 0$ for all λ; for endpoint-consistent operators satisfying $f_0 = p_H$ and $f_1 = p_A$, $\mathrm{HDDI}_f(0) = \mathrm{HDDI}_f(1) = 0$. We separately define absolute-disparity amplification as $A(\lambda) = |D_f(\lambda)| - |D(L_\lambda)|$. HDDI records the signed departure from the additive reference, whereas $A(\lambda)$ records whether the absolute magnitude of disparity increases or decreases relative to that reference. We therefore evaluate the two quantities separately.

3.3 Empirical Integration Operator

The framework is defined for a general f_λ. Across the three studies, we instantiate it using a two-source logarithmic opinion pool, expressed in log-odds form (Genest et al., 1984; Genest & Zidek, 1986). For binary probabilities, this combines weighted component log-odds and maps the result back to the probability scale. For numerical stability, probabilities entering the logit are clipped to $[\epsilon, 1-\epsilon]$:

$$\widetilde{p} = \min\{\max(p, \epsilon), 1-\epsilon\}, \qquad f_\lambda(h, a) = \sigma\Big[(1-\lambda)\,\mathrm{logit}(\widetilde{h}) + \lambda\,\mathrm{logit}(\widetilde{a})\Big], \tag{4}$$

where $\mathrm{logit}(p) = \log[p/(1-p)]$ and $\sigma(z) = (1+e^{-z})^{-1}$.

Log-odds aggregation has also been studied in the probability-forecast combination literature (Satopää et al., 2014). Here, we use the operator as a common nonlinear empirical instantiation rather than as a behavioral model of human reliance. Relative to additive probability pooling, its pointwise departure depends on the component probabilities, but Proposition 2 implies that nonzero HDDI arises only when these departures differ on average across groups. Across studies, we hold the integration operator and weight grid fixed while varying the data domain, source of human probabilities, algorithmic model, and audited groups.

4 Empirical Validation Design

We evaluate HDDI in three complementary settings: ACSIncome provides controlled construct validation, Civil Comments extends the analysis to text moderation and finite-sample recovery, and COMPAS replaces simulated human probabilities with observed repeated judgments. Across all studies, we use the component reference and nonlinear log-odds operator defined in Section 3, evaluated at $\lambda \in \{0, .25, .50, .75, 1\}$ without selecting or optimizing λ. Additive pooling provides the reference, linear pooling serves as the zero control, and endpoint and probability-validity checks are applied before analysis.

ACSIncome. We use the Folktables ACSIncome task on California 2018 ACS 1-Year person records (Ding et al., 2021), with 136,965 training, 29,350 validation, and 29,350 test observations. Splits use seed 42 and joint stratification on the income label, SEX, and RAC1P, with a deterministic fallback for rare joint strata. The algorithmic layer comprises Logistic Regression, HistGradientBoosting, and MLP classifiers, with calibration selected using validation data only. The human layer consists of theory-informed simulated probability scenarios constructed independently of the algorithmic predictions; SEX- and supported RAC1P-based comparisons provide the controlled setting for construct validation.

Civil Comments. We use the identity-annotated Civil Comments corpus (Borkan et al., 2019), retaining 405,130 training, 21,293 validation, and 21,577 test comments with positive identity-annotation counts. Group membership is defined by the content-level identity-reference indicator $G = \mathbb{1}\{\max_j I_j \geq 0.5\}$, where I_j denotes a retained identity-annotation score. These annotations describe identities referenced in comment content rather than demographic attributes of the author. The algorithmic layer contains TF-IDF Logistic Regression and DistilRoBERTa toxicity probabilities, calibrated with isotonic and Platt scaling, respectively, using validation data only. The simulated moderator is a regularized logistic model based on 21 human-visible text features and excludes identity annotations, toxicity annotations and scores, identifiers, and algorithmic predictions. Its 100 frozen moderator conditions span five prespecified families: baseline, uniform identity overflagging, ambiguity-sensitive identity overflagging, oracle-informed non-toxic stress testing, and stochastic moderator inconsistency. Across validation and test splits, the two text models, 100 moderator conditions, and three interior integration weights yield 1,200 interior conditions. We analyze both overall HDDI and $\mathrm{HDDI}_{Y=0}$ for non-toxic comments; the latter is a probability-level conditional disparity diagnostic relevant to false-positive disparities, not a thresholded false-positive-rate gap.

COMPAS. We use the Stanford Policy Lab recidivism-prediction data and repeated human judgments from Lin et al. (2020). Cleaning yields 5,000 participant–defendant judgments from 100 participants, aggregated to 994 defendants. The prespecified African-American versus Caucasian analysis contains 901 defendants (528 African-American and 373 Caucasian). The primary human probability is the mean continuous recidivism rating scaled to $[0, 1]$; the mean binary judgment rate and restrictions to defendants with at least two or three raters are used for sensitivity analysis. Because the COMPAS score is ordinal rather than probabilistic, algorithmic probabilities are obtained from a separate 6,214-record calibration cohort. Logistic score mapping is retained under a prespecified practical-difference rule because the best non-logistic alternative improves on logistic by less than 0.001 in both log loss and Brier score.

Finite-Sample Recovery. We evaluate finite-sample recovery of $\mathrm{HDDI}_{Y=0}$ on Civil Comments. From the 21,577-comment test population, we draw stratified samples without replacement over $G \times Y$ at $n \in \{500, 1000, 2500, 5000, 10000, 15000\}$, with 300 replicates per size. Within each sample-size–replicate pair, the same sampled identifiers are reused across model, moderator, and λ conditions. Recovery is assessed using RMSE, sign recovery, magnitude recovery (absolute error within 20% of the population magnitude for $|\mathrm{HDDI}| > 10^{-4}$), amplification-class recovery, and effective subgroup support; full definitions are reported in the appendix. The interval approximation is audited against exact conditional stratified percentile-bootstrap intervals over 9,600 audited outer samples spanning the eight cases and six sample sizes, with $B = 500$ bootstrap resamples per outer sample.

5 Results

5.1 Controlled Construct Validation

We first test whether HDDI exhibits the behavior implied by its component-referenced definition under controlled variation. Across all three ACS models, absolute HDDI is zero at $\lambda = 0$ and $\lambda = 1$ and becomes nonzero at interior weights (Figure 2A). Across the full ACS evaluation, interior HDDI ranges from -0.001935 to 0.001348, with $\max |\mathrm{HDDI}| = 0.001935$. The linear-pooling control remains zero to numerical tolerance, providing an implementation check on the additive reference.

Structured perturbations produce larger baseline-relative HDDI changes than matched stochastic-noise conditions across the three interior weights (Figure 2B). This comparison is descriptive rather than inferential: it assesses whether HDDI responds more strongly to prespecified structural changes in the human component than to matched random perturbations. HDDI also remains distinct from final hybrid disparity (Figure 2C). Frozen pair comparisons identify both similar final disparities with different HDDI values and similar HDDI values with different final disparities. Thus, $D_f(\lambda)$ describes disparity in the resulting hybrid system, whereas HDDI describes departure from the component-implied additive reference. These comparisons support descriptive construct separation, not statistical independence or causal attribution.

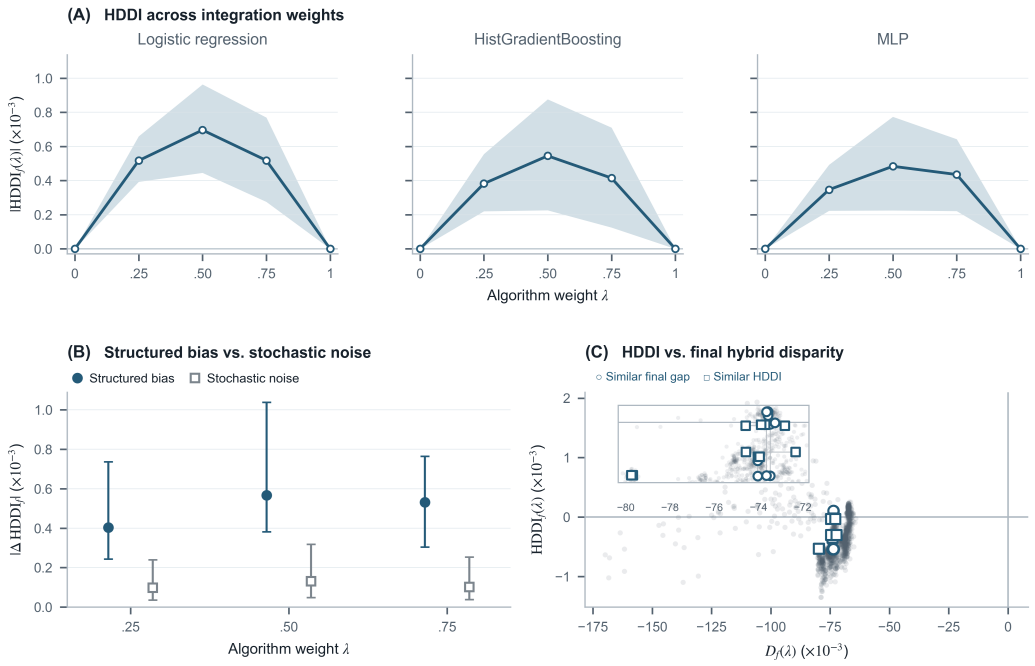


Figure 2: **Controlled construct validation of HDDI on ACSIncome.** (A) Absolute $\mathrm{HDDI}_f(\lambda)$ across integration weights for Logistic Regression, HistGradientBoosting, and MLP; curves and shaded bands summarize variation across prespecified experimental conditions. (B) Baseline-relative absolute HDDI change under structured-bias and matched stochastic-noise conditions; error bars summarize across-condition variation rather than confidence intervals. (C) Final hybrid disparity $D_f(\lambda)$ versus signed $\mathrm{HDDI}_f(\lambda)$. Highlighted circles denote similar final disparities with different HDDI values, whereas squares denote similar HDDI values with different final disparities.

5.2 Domain Extension and Statistical Recoverability

We next examine HDDI in identity-sensitive text moderation. Conditioning on non-toxic comments applies the same component-referenced contrast within a pathway relevant to false-positive disparities. Across the 1,200 interior Civil Comments conditions, the conditional magnitude exceeds the corresponding overall magnitude in 974 cases (81.17%; Figure 3A–B). The pattern is consistent across the two text classifiers: conditional HDDI estimates across matched conditions have Pearson correlation $r = 0.966$, with 0.99 sign agreement and 0.99 amplification-class agreement. The conditional excess also varies with the integration weight (Figure 3C). Its median is positive at lower and intermediate algorithmic weights for both models and declines as λ increases. At $\lambda = 0.75$, the TF-IDF median is slightly negative whereas the DistilRoBERTa median remains positive. We therefore interpret this pattern as empirical dependence on the human–AI mixing regime rather than a monotonic property implied by the HDDI definition. The $Y = 0$ quantity remains a probability-level conditional contrast and should not be interpreted as a thresholded false-positive-rate difference.

We then characterize finite-sample recovery of conditional HDDI. Median RMSE decreases from 0.002220 at $n = 500$ to 0.000214 at $n = 15{,}000$ (Figure 4A), while median effective support in the $G = 1, Y = 0$ cell increases from 152 to 4,563 observations. These results characterize recoverability over the studied sampling regime rather than an asymptotic convergence rate or universal sample-size requirement. Recovery differs across diagnostic features (Figure 4B). At $n = 500$, sign and amplification-class recovery are already 0.993 and 0.973, respectively, and both reach 1.000 by $n = 1{,}000$. Magnitude recovery is more sensitive to sample size, increasing from 0.353 at $n = 500$ to 1.000 at $n = 15{,}000$. Within this design, finite-sample uncertainty therefore affects recovery of HDDI magnitude more strongly than recovery of its direction. The interval audit yields the same qualitative recovery picture. Approximate and exact conditional-bootstrap interval widths lie close

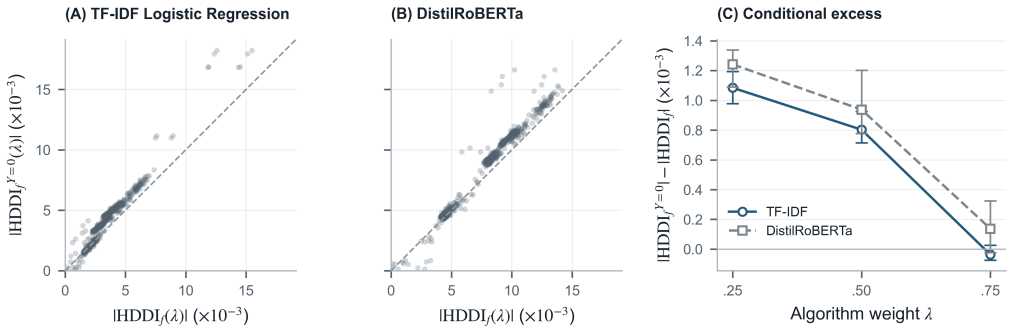


Figure 3: **Conditional HDDI in Civil Comments.** (A) Absolute non-toxic conditional HDDI, $|\mathrm{HDDI}_f^{Y=0}(\lambda)|$, versus absolute overall HDDI for TF-IDF Logistic Regression. (B) The same comparison for DistilRoBERTa; the dashed diagonal denotes equality. (C) Conditional excess, $|\mathrm{HDDI}_f^{Y=0}(\lambda)| - |\mathrm{HDDI}_f(\lambda)|$, across interior integration weights. Points denote medians and vertical bars denote interquartile ranges across prespecified experimental conditions. The $Y = 0$ analysis is relevant to false-positive disparities but is not itself a false-positive rate.

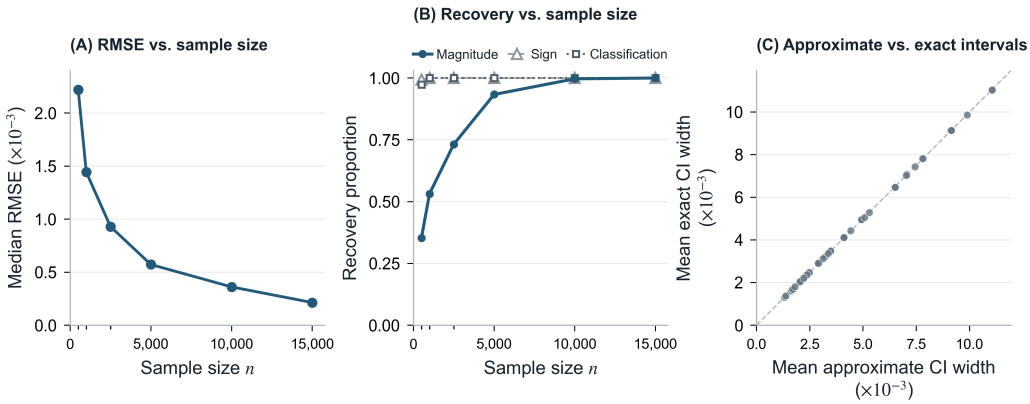


Figure 4: **Finite-sample recoverability of conditional HDDI.** (A) Median RMSE of the non-toxic conditional HDDI estimator across the prespecified sample-size grid. (B) Sign, amplification-class, and magnitude recovery across sample sizes. (C) Mean interval widths from the approximate procedure versus exact conditional stratified percentile-bootstrap widths across the frozen Tier-2 audit cases; the dashed diagonal denotes equality.

to the equality line (Figure 4C), with median relative width differences near 4%. Zero-exclusion agreement remains at least 0.98 across audited case–sample-size combinations, with a median of 1.00 at every sample size; reliability status differs in only 2 of 48 audited combinations. These results support use of the computationally cheaper approximation in the evaluated settings while showing that the two interval procedures are not numerically identical.

5.3 Observed-Human Transfer

Finally, we replace the simulated human layer with probabilities derived from repeated observed judgments. Under the primary continuous human-probability definition, HDDI is 0.0038, 0.0055, and 0.0043 at $\lambda = 0.25$, 0.50, and 0.75, respectively, with $\max|\mathrm{HDDI}| = 0.0055$. Endpoint HDDI and the linear-pooling control remain zero to numerical tolerance. The nonzero interior pattern therefore persists when p_H is derived from observed rather than simulated judgments. Restricting the cohort to defendants with at least two or three raters preserves HDDI sign and amplification

classification while reducing maximum absolute HDDI to 0.0050 and 0.0035, respectively. A larger sensitivity arises from the definition of the human probability: replacing continuous mean ratings with binary judgment rates preserves sign and amplification classification but substantially changes HDDI magnitude. Thus, in this observed-human setting, the component-referenced pattern persists while its magnitude depends on how human judgments are represented probabilistically.

6 Discussion and Limitations

Our results position HDDI as a component-referenced diagnostic rather than an additional fairness criterion. Prior work shows that human involvement and reliance can change the fairness properties of combined systems (Donahue et al., 2022; Gillis et al., 2022); HDDI addresses a narrower measurement question by distinguishing disparity already implied by the human and algorithmic component probabilities from the non-additive departure associated with their integration. The ACS experiments establish this distinction under controlled conditions: HDDI exhibits the expected boundary and linear-control behavior while remaining descriptively distinct from final hybrid disparity.

The Civil Comments results further illustrate why the reference matters empirically. Conditional HDDI is often larger within the non-toxic stratum than overall, yet this pattern depends on the integration regime rather than following a monotonic theoretical rule. More broadly, the separation between final disparity and reference-relative deviation parallels the motivation for reference-based bias diagnostics, where interpretation depends on what the observed quantity is compared against (Wang & Russakovsky, 2021). In the present setting, additive pooling provides an explicit component-based reference, while the nonlinear operator determines how the two probability sources are combined. Consequently, HDDI is not a property of the component disparities alone: it characterizes how the specified integration rule departs from what those components jointly imply under the additive reference. This also means that HDDI should be interpreted alongside D_H, D_A, and $D_f(\lambda)$ rather than as a substitute for conventional disparity measurement. This reference-based interpretation also connects HDDI to the broader probability-pooling literature. Linear and logarithmic pooling provide different aggregation rules for combining probabilistic inputs (Stone, 1961; Genest et al., 1984). Here, the additive pool is used as a transparent reference rather than as a claim about how deployed human–AI systems ought to combine predictions. A nonzero HDDI therefore indicates group-differential departure from that reference under the specified operator; its sign or magnitude alone does not establish harm, unfairness, or a preferred intervention.

The finite-sample results add a statistical boundary to this interpretation: a well-defined population contrast need not be equally recoverable at all levels of subgroup support. Magnitude recovery improves substantially with effective sample size even when directional recovery is already high, consistent with broader concerns about uncertainty in fairness estimation (Ethayarajh, 2020). The COMPAS study then extends the diagnostic to repeated observed human judgments. The nonzero interior pattern persists, but its magnitude changes when human judgments are represented probabilistically in different ways. This sensitivity reinforces that HDDI is defined with respect to the probability-valued component inputs and the integration rule, rather than being an invariant property of the underlying decision task.

Several limitations remain. First, although HDDI is defined for a general integration operator, our empirical studies use log-odds pooling as the common nonlinear instantiation; other integration rules may produce different magnitudes or patterns. Second, the human layers in ACSIncome and Civil Comments are simulated, while the COMPAS judgments are observed but do not represent decisions made in an actual deployed human–AI interaction. Third, HDDI characterizes group-mean probability disparity relative to an additive reference and does not identify causal mechanisms or determine normative fairness. Future work should therefore evaluate HDDI under behaviorally estimated or deployed integration policies and broader operator families, and develop inference for richer settings such as multigroup, intersectional, or threshold-dependent disparity diagnostics.

To conclude, these results suggest that auditing hybrid systems only through their final disparity can obscure whether observed differences are already reflected in the component predictions or arise as a non-additive departure under the specified integration rule. HDDI complements conventional disparity measures by making this distinction explicit. It is therefore best interpreted as a diagnostic for locating disparity relative to a component-based reference, rather than as evidence of harm or a prescription for intervention.

AI USE STATEMENT

Generative AI tools were used primarily to assist with code development, debugging, and refinement, and to improve the clarity, grammar, and presentation of manuscript text. AI-assisted writing was reviewed and revised by the authors to ensure that the final text accurately reflects the intended methods, results, and interpretations. The authors take responsibility for all code, analyses, claims, and manuscript content.

ETHICS STATEMENT

This study uses existing public datasets and collects no new personal data. Civil Comments identity annotations refer to identities mentioned in comment content, not commenter demographics, and simulated human probabilities in ACSIncome and Civil Comments are methodological constructs rather than measured human behavior. COMPAS uses previously collected human judgments in a sensitive criminal-justice context; our results are statistical diagnostics, not recommendations for individual risk assessment or deployment.

REPRODUCIBILITY STATEMENT

We designed the empirical pipeline around frozen inputs, deterministic data splits where applicable, prespecified integration weights, validation-only calibration, and prespecified validation checks. The main text reports the quantities needed to interpret the empirical design, while the appendix documents preprocessing and subgroup definitions, calibration procedures, simulated-human scenario construction, finite-sample recovery criteria, bootstrap procedures, and supporting numerical summaries. The anonymized reproducibility package includes analysis scripts, phase manifests, provenance records, validation outputs, and cryptographic checksums for frozen intermediate and final artifacts.

REFERENCES

Saar Alon-Barkat and Madalina Busuioc. Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, 2023. doi: 10.1093/jopart/muac007.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445717.

Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Uncertainty matters: Stable conclusions under unstable assessment of fairness results. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1198–1206. PMLR, 2024. URL https://proceedings.mlr.press/v238/barrainkua24a.html.

Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973. doi: 10.2307/144855.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. Association for Computing Machinery, 2019. doi: 10.1145/3308560.3317593.

Morris H. DeGroot and Julia Mortera. Optimal linear opinion pools. *Management Science*, 37(5): 546–558, 1991. doi: 10.1287/mnsc.37.5.546.

Frances Ding, Moritz Hardt, John P. Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1639–1656. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533221.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018. doi: 10.1126/sciadv.aao5580.

Kawin Ethayarajh. Is your classifier actually biased? measuring fairness under uncertainty with Bernstein bounds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2914–2919. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.262.

Hairu Fan and Shiyuan Wang. What makes a fairness gap actionable? statistical actionability for responsible AI deployment. *arXiv preprint arXiv:2608.16912*, 2026. URL https://arxiv.org/abs/2608.16912.

Jiashi Gao, Kexin Liu, Xinwei Guo, Junlei Zhou, Jiaxin Zhang, Xiangyu Zhao, Guanhua Chen, Xin Yao, and Xuetao Wei. Fair decision utility in human–AI collaboration: Interpretable confidence adjustment for humans with cognitive disparities. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hqq6GyYISN.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Rethinking fairness for human-AI collaboration. In *15th Innovations in Theoretical Computer Science Conference (ITCS 2024)*, volume 287 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 52:1–52:1. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi: 10.4230/LIPIcs.ITCS.2024.52.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986. doi: 10.1214/ss/1177013825.

Christian Genest, Samaradasa Weerahandi, and James V. Zidek. Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17(1):61–70, 1984. doi: 10.1007/BF00140056.

Talia Gillis, Bryce McLaughlin, and Jann Spiess. On the fairness of machine-assisted human decisions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 890–890. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3533152.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019. doi: 10.1145/3359152.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ben Jann. The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal*, 8 (4):453–479, 2008. doi: 10.1177/1536867X0800800401.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/234b941e88b755b7a72a1c1dd5022f30-Abstract.html.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392.

Zhiyuan Jerry Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652, 2020. doi: 10.1126/sciadv.aaz0652.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021. doi: 10.1146/annurev-statistics-042720-125902.

Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973. doi: 10.2307/2525981.

Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2021. doi: 10.1145/3447548.3467258.

Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B*, 72(1):71–91, 2010. doi: 10.1111/j.1467-9868.2009.00726.x.

Ville A. Satopää, Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356, 2014. doi: 10.1016/j.ijforecast.2013.09.009.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022. doi: 10.1073/pnas.2111547119.

M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32(4):1339–1342, 1961. doi: 10.1214/aoms/1177704873.

Bhanu Tokas, Rahul Nair, and Hannah Kerner. DPA: A one-stop metric to measure bias amplification in classification datasets. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5045.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024. doi: 10.1038/s41562-024-02024-1.

Angelina Wang and Olga Russakovsky. Directional bias amplification. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10882–10893. PMLR, 2021. URL `https://proceedings.mlr.press/v139/wang21t.html`.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1323.

Domenique Zipperling, Luca Deck, Julia Lanzl, and Niklas Kühl. It's only fair when i think it's fair: How gender bias alignment undermines distributive fairness in human-AI collaboration. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025. doi: 10.1145/3715275.3732084.

A Appendix

A Additional Mathematical Details

This appendix provides algebraic details supporting the component-referenced formulation in Section 3. These results do not introduce additional estimands; they make explicit the identities used in the main text.

Additive-reference identity. Recall that

$$D(q) = \mathbb{E}[q \mid G = 1] - \mathbb{E}[q \mid G = 0],$$

and that the additive component reference is

$$L_\lambda(p_H, p_A) = (1 - \lambda)p_H + \lambda p_A.$$

By linearity of conditional expectation,

$$\begin{aligned} D(L_\lambda) &= \mathbb{E}[(1 - \lambda)p_H + \lambda p_A \mid G = 1] - \mathbb{E}[(1 - \lambda)p_H + \lambda p_A \mid G = 0] \\ &= (1 - \lambda)\left\{\mathbb{E}[p_H \mid G = 1] - \mathbb{E}[p_H \mid G = 0]\right\} \\ &\quad + \lambda\left\{\mathbb{E}[p_A \mid G = 1] - \mathbb{E}[p_A \mid G = 0]\right\} \\ &= (1 - \lambda)D_H + \lambda D_A. \end{aligned} \tag{5}$$

No independence assumption between p_H, p_A, and G is required.

Operator-deviation representation. Define the individual-level departure of the integration operator from the additive component reference as

$$\delta_\lambda(x) = f_\lambda(p_H(x), p_A(x)) - L_\lambda(p_H(x), p_A(x)).$$

Because

$$f_\lambda(p_H, p_A) = L_\lambda(p_H, p_A) + \delta_\lambda,$$

the definition

$$\mathrm{HDDI}_f(\lambda) = D(f_\lambda(p_H, p_A)) - D(L_\lambda)$$

gives

$$\begin{aligned} \mathrm{HDDI}_f(\lambda) &= D(\delta_\lambda) \\ &= \mathbb{E}[\delta_\lambda \mid G = 1] - \mathbb{E}[\delta_\lambda \mid G = 0]. \end{aligned} \tag{6}$$

Thus, nonlinearity of f_λ is not sufficient for nonzero HDDI: the mean operator departure must differ between groups.

Conditional extension. For an event C satisfying $\Pr(G = g, C) > 0$ for both groups, define

$$D_C(q) = \mathbb{E}[q \mid G = 1, C] - \mathbb{E}[q \mid G = 0, C].$$

The same linearity argument gives

$$D_C(L_\lambda) = (1 - \lambda)D_C(p_H) + \lambda D_C(p_A),$$

and therefore

$$\mathrm{HDDI}_{f,C}(\lambda) = D_C(f_\lambda) - D_C(L_\lambda) = \mathbb{E}[\delta_\lambda \mid G = 1, C] - \mathbb{E}[\delta_\lambda \mid G = 0, C].$$

The Civil Comments analysis uses $C = \{Y = 0\}$. This remains a contrast in conditional mean probabilities and is not a thresholded false-positive-rate difference.

B Additional Experimental Details

The main text reports the experimental architecture and primary estimands. Here we record implementation details needed to interpret or reproduce those analyses without introducing additional experimental conditions.

ACSIncome. The ACSIncome sample uses California 2018 ACS 1-Year person records. The train/validation/test split uses seed 42 and joint stratification on the income label, SEX, and RAC1P, with deterministic assignment for rare joint strata that cannot support ordinary stratification. The primary feature specification is group-blind and uses AGEP, COW, SCHL, MAR, OCCP, POBP, RELP, and WKHP. The three algorithmic models are Logistic Regression, HistGradientBoostingClassifier, and MLPClassifier. Final calibration is the identity mapping for Logistic Regression and HistGradientBoosting and Platt scaling for the MLP.

The simulated human layer is generated independently of the algorithmic probabilities. Its baseline model is an ℓ_2-regularized logistic regression using the human-visible variables AGEP, SCHL, WKHP, OCCP, and COW. Frozen perturbations include a baseline control, structured group- and qualification-dependent perturbations, and noise-only conditions. These conditions are used to evaluate whether HDDI responds to structured changes in the human component while retaining stochastic perturbations as a descriptive comparison rather than an inferential null model.

Civil Comments. Civil Comments retains rows with positive identity-annotation counts. The analysis group is

$$G = \mathbb{1}\{\max_j I_j \geq 0.5\},$$

where I_j denotes a retained content-level identity annotation. The indicator therefore refers to identity content in the comment rather than the author's demographic membership.

The validation split is divided into 10,646 calibration-fit and 10,647 calibration-selection observations. The selected calibration methods are isotonic calibration for TF-IDF Logistic Regression and Platt scaling for DistilRoBERTa. Before application of the log-odds integration operator, probabilities are clipped using $\epsilon = 10^{-6}$.

The simulated moderator is an ℓ_2-regularized logistic regression based on 21 human-visible text-derived features. Identity annotations, toxicity scores and subtype annotations, identifiers, and algorithmic predictions or probabilities are excluded from its feature set. The frozen moderator registry contains five scenario families: H0 baseline control, H1 uniform identity overflagging, H2 ambiguity-sensitive identity overflagging, H3 oracle-informed non-toxic overflagging stress testing, and H4 stochastic moderator inconsistency. H1–H3 use $\gamma \in \{0.25, 0.50, 1.00\}$; H4 uses $\sigma \in \{0.25, 0.50, 1.00\}$ with 30 frozen replicates at each noise level. Together with the baseline, this yields 10 deterministic and 90 stochastic conditions. H3 is used as a controlled stress condition rather than as a behavioral model of human moderation.

COMPAS. The observed-human analysis begins with 9,639 response records and retains 5,000 cleaned participant–defendant judgments from 100 participants, aggregated to 994 defendants. The primary African-American versus Caucasian analysis contains 901 defendants. The primary human probability is the mean continuous recidivism rating divided by 100; the binary sensitivity specification uses the mean binary recidivism judgment.

Because the original COMPAS score is ordinal, its probability mapping is fit on a separate calibration cohort. Of 7,214 source records, 1,000 defendants appearing in the human-judgment vignette set are excluded, leaving 6,214 records. These are divided into 4,349 calibration-fit and 1,865 calibration-selection observations using outcome-by-score stratification and seed 42. Logistic, isotonic, and empirical score mappings are compared. Logistic mapping is retained under the frozen practical-difference rule because the best non-logistic candidate improves on logistic by less than 0.001 in both validation log loss and Brier score.

C Finite-Sample Recovery Details

Let θ_c denote the full-test-population $\mathrm{HDDI}_{Y=0}$ for frozen condition c, and let $\widehat{\theta}_{c,n,r}$ denote its estimate from sample size n and replicate r. The recovery experiment uses

$$n \in \{500, 1000, 2500, 5000, 10000, 15000\}, \qquad R = 300.$$

Samples are drawn without replacement within the four $G \times Y$ strata. Within each sample-size–replicate pair, the same sampled identifiers are reused across model, moderator, and λ conditions.

RMSE. Condition-specific root mean squared error is

$$\mathrm{RMSE}_c(n) = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_{c,n,r} - \theta_c\right)^2\right]^{1/2}.$$

Figure 4Figure 4A reports the median condition-level RMSE at each sample size.

Sign and magnitude recovery. The frozen near-zero threshold is

$$\tau_0 = 10^{-4}.$$

For conditions satisfying $|\theta_c| > \tau_0$, sign recovery for replicate r is

$$S_{c,n,r} = \mathbb{1}\left\{\mathrm{sign}(\widehat{\theta}_{c,n,r}) = \mathrm{sign}(\theta_c)\right\},$$

and magnitude recovery is

$$M_{c,n,r} = \mathbb{1}\left\{\left|\widehat{\theta}_{c,n,r} - \theta_c\right| \leq 0.20|\theta_c|\right\}.$$

Conditions at or below the near-zero threshold are excluded from these two recovery summaries. Condition-level recovery is the mean across the 300 replicates, and Figure 4B reports the median condition-level recovery at each sample size.

Amplification-class recovery and effective support. For each sampled estimate, amplification-class recovery records whether the estimated amplification/attenuation classification matches the corresponding full-population classification under the frozen rule based on

$$A(\lambda) = |D_f(\lambda)| - |D(L_\lambda)|.$$

Effective subgroup support is the realized number of observations in the $G = 1, Y = 0$ cell. The reported support curve summarizes the median effective cell size at each n.

Interval audit. The Tier-2 interval analysis uses eight frozen cases and the same six sample sizes. The frozen approximate procedure uses a conditional stratified bootstrap-style sufficient-statistic interval with 95% percentile limits. Its behavior is audited against exact conditional stratified percentile bootstrap intervals using $B = 500$ bootstrap resamples within each audited outer sample. The audit comprises 200 outer samples for each of the $8 \times 6 = 48$ case–sample-size combinations, or 9,600 outer samples in total. Agreement is assessed through interval width, zero exclusion, and the frozen reliability classification. The main text reports the resulting aggregate agreement rather than treating this audit as a separate inferential analysis.

D Supporting Numerical Summaries

Tables 1 and 2 provide exact numerical support for quantities displayed or summarized in the main text. They do not introduce additional experimental comparisons.

E Reproducibility Details

All analyses use frozen configurations and deterministic identifiers for experimental conditions. Data splitting and the principal model-fitting and calibration procedures use recorded random seeds; the primary split and calibration seed is 42. Civil Comments finite-sample draws use deterministic sample-size–replicate seeds, and the common-sample registry ensures that a given sample-size–replicate draw is reused across the corresponding model, moderator, and λ conditions. The frozen manifests record preprocessing specifications, model hyperparameters, calibration choices, scenario identifiers, software versions, and output provenance. No protected-group annotation, toxicity annotation, row identifier, or algorithmic prediction is included among the 21 human-visible Civil Comments moderator features. The ACS simulated-human construction likewise excludes algorithmic probabilities. COMPAS probability calibration uses only the separate external calibration cohort described above.

All main-text numerical claims and the supporting values reported in this appendix were extracted from the frozen analysis outputs; no additional model training, simulation, subsampling, calibration, or bootstrap analysis was introduced for the appendix.

Table 1: Conditional-excess summaries for Civil Comments. Exact numerical summaries corresponding to Figure 3C. Conditional excess is $|\mathrm{HDDI}_f^{Y=0}| - |\mathrm{HDDI}_f|$. Values are reported in units of 10^{-3}; entries give the median and interquartile range across frozen conditions.

Model	λ	Median	Q1	Q3
TF-IDF	.25	1.0842	0.9773	1.1940
	.50	0.8026	0.7151	0.9256
	.75	-0.0369	-0.0738	0.0256
DistilRoBERTa	.25	1.2415	1.0902	1.3389
	.50	0.9375	0.7777	1.2011
	.75	0.1369	-0.0224	0.3240

Table 2: Observed-human HDDI sensitivity in COMPAS. Interior HDDI values for the primary specification and the rater-coverage and human-probability sensitivities summarized in Section 5.3. N denotes the African-American/Caucasian analysis cohort for the corresponding specification.

Human probability	Rater set	N	$\lambda = .25$	$\lambda = .50$	$\lambda = .75$	$\max \lvert\mathrm{HDDI}\rvert$
Continuous mean	All	901	0.0038	0.0055	0.0043	0.0055
Continuous mean	≥ 2	872	0.0034	0.0050	0.0039	0.0050
Continuous mean	≥ 3	799	0.0023	0.0035	0.0028	0.0035
Binary rate	All	901	0.0184	0.0373	0.0520	0.0520