*Journal of Harbin Engineering University*
*ISSN: 1006-7043*

Vol 47 No. 11
Nov 2026

# A Composite Metric for Identifying Algorithmic Bias in AI-Driven Governance Models

**Abubakar Yushau[1*], Aminu Aliyu Abdullahi[2], Abdulhamid Mohammed Awais[3]**

[1, 2, 3] Department of Computer Science, Federal University Dutse, Jigawa State, Nigeria

**Abstract**

Public agencies increasingly use AI systems to allocate benefits, flag suspected fraud, prioritize inspections and score risk. Existing bias metrics (demographic parity, equalized odds and their variants) were designed for classifiers: each capture one statistical property at one point in time, treats all errors as equally costly, and is blind to whether affected people can see, contest and obtain redress for a decision. Known impossibility results mean that these criteria cannot all be satisfied together, leaving auditors without a defensible way to report an overall judgement. To propose and evaluate the Composite Governance Bias Index (CGBI), a bounded, interpretable index built for the governance setting rather than adapted from classifier metrics. CGBI combines four components: Distributive Disparity (D, a harm-weighted error gap), Procedural Asymmetry (P, group differences in explanation, human review, appeal and timely redress), Intersectional Exposure (I, the worst partially pooled intersectional cell) and Dynamic Drift (T, the projected worsening across retraining cycles). Components are normalized by a stakes profile and aggregated by a weighted power mean with a red-line veto, so that a severe failure on one dimension cannot be averaged away. We evaluated CGBI in a Monte Carlo study of 5,300 synthetic governance pipelines (typically six monitoring windows of 8,000 decisions, three protected attributes, 12 intersectional cells) with known, injected bias of five types (distributive, procedural, intersectional, drift and combined) plus bias-free and base-rate-difference controls, and compared it with equalized odds, demographic parity, the four-fifths rule, an unshrunk subgroup metric and an arithmetic-mean composite. Every metric was thresholded at 5% false alarms on bias-free systems. CGBI detected distributive bias in 99% of pipelines, procedural-only bias in 100%, intersectional-only bias in 96%, feedback-driven drift in 83% and combined bias in 100%. Conventional screens (four-fifths rule plus an equalized-odds tolerance of 0.10) passed 99% of procedural-only and 98% of intersectional systems; CGBI flagged 100% and 97% of these, whereas equalized odds flagged only 6% and 11%. The component with the largest excess matched the injected bias type in 88%, 100%, 94% and 68% of pipelines (S1–S4), and rankings were stable across 2,000 random weight and exponent configurations (Kendall's $\tau = 0.92$; 5th percentile 0.86). Parity-based metrics flagged 100% of fair systems whose groups simply had different base rates, against 4% for CGBI. Several expectations were not met: at the same false-alarm rate, non-compensatory aggregation improved detection over an arithmetic-mean composite only for mild bias concentrated in one dimension (procedural-only at severity 0.25: 60% vs 30%) and was otherwise equal or up to 9 points worse, the trend component gave no earlier warning of drift than demographic parity, CGBI was less sensitive than equalized odds for a single distributive bias at $n = 2,000$ (40% vs 78%), and percentile-bootstrap intervals under-covered population values (coverage of a nominal 90% ranged from 0% to 100% depending on component and case). In simulation, CGBI works as an omnibus screening and diagnostic instrument that extends audit coverage to procedural and intersectional harms that standard metrics miss, is robust to base-rate differences and to reasonable weighting choices, and identifies the likely source of bias. Its added value over All findings come from synthetic data; validation on real administrative systems with procedural logs, and stakeholder calibration of the stakes profile, are the necessary next steps.

**Keywords**: Algorithmic bias, AI governance, composite indicator, fairness metrics, procedural justice, intersectionality, algorithmic auditing, public-sector algorithms.

## 1. Introduction

Governments now rely on algorithmic systems to decide who is investigated for benefit fraud, which patients receive additional care, which defendants are labelled high risk and which applications are fast-tracked. When such systems err systematically against particular groups, the harm is not only statistical. It is administrative: people are denied support, wrongly accused or left without a route to challenge the outcome. Well-documented failures in criminal-justice

risk scoring (Angwin et al., 2016), health-care resource allocation (Obermeyer et al., 2019), commercial face analysis (Buolamwini & Gebru, 2018) and welfare automation (Eubanks, 2018) show that these harms are not rare edge cases. Regulatory frameworks such as the European Union Artificial Intelligence Act (European Union, 2024) place duties on providers and deployers of high-risk systems to examine and manage bias, yet they do not prescribe how bias should be quantified in a way that is comparable across systems.

The measurement problem is difficult for three reasons. First, the technical literature offers dozens of group-fairness criteria (Hardt et al., 2016; Chouldechova, 2017; Verma & Rubin, 2018; Mehrabi et al., 2021) that cannot generally be satisfied at once when base rates differ (Kleinberg et al., 2017; Chouldechova, 2017). Many of these metrics are also empirically redundant (Garg et al., 2020; Majumder et al., 2023), so practitioners face a crowded menu with no principled way of turning it into a single audit judgement. Second, almost all metrics evaluate a model output on a test set. In governance, however, harm also arises in the pipeline around the model: who is routed to human review, who receives an explanation, who can afford to appeal and how long redress takes. Third, governance systems operate over time, on small intersectional populations, and in contexts where a wrongful flag can be far more damaging than a missed one.

This paper proposes the Composite Governance Bias Index (CGBI) as a response. CGBI measures bias as a property of the socio-technical decision pipeline along four dimensions (distributive, procedural, intersectional and dynamic), scales them by an explicit stakes profile, and aggregates them with a no compensatory rule. We specify the index formally, establish its basic properties, and evaluate it in a large simulation study in which the ground truth about bias is known. We report not only where CGBI performs well but also where our expectations were not met, because an index intended for regulatory use should be judged on a full account of its behavior.

**Fairness Metrics and their Aggregation**

Group-fairness criteria such as demographic parity, equalized odds and predictive parity (Hardt et al., 2016; Chouldechova, 2017) are mutually incompatible outside special cases (Kleinberg et al., 2017), and comparative analyses show considerable overlap among them (Garg et al., 2020; Majumder et al., 2023). Several authors have therefore proposed ways to summarize unfairness. Speicher et al. (2018) borrow inequality indices from economics to quantify unfairness at the individual and group level and to decompose it into between-group and within-group components. Agarwal et al. (2022) propose a Fairness Score and Bias Index for supervised-learning systems together with a certification procedure. Monnier et al. (2026) show that the choice of meta-metric used to aggregate subgroup disparities into a single score can reverse fairness conclusions and argue that such choices must be explicitly justified.

Outside fairness proper, composite trust indices that combine fairness, transparency, privacy and accountability scores are emerging in applied AI governance (Jan et al., 2026). Composite-indicator practice more generally (OECD/JRC, 2008) stresses that weighting and aggregation embed value judgements, and the multi-criteria literature distinguishes compensatory from non-compensatory aggregation (Munda, 2008). Dwork and Ilvento (2018) add a caution that fairness properties of components do not simply carry over to systems composed of them.

**Intersectional Auditing with Sparse Subgroups**

Intersectionality (Crenshaw, 1989) implies that harms may concentrate where attributes overlap even when each marginal group looks fair, a phenomenon formalized as fairness gerrymandering (Kearns et al., 2018). Because intersectional cells are small, recent work addresses estimation: structured regression for small subgroups (Herlihy et al., 2024), size-adaptive and Bayesian hypothesis tests (Ferrara et al., 2025), bootstrap-based simultaneous audits (Cherian et al., 2023), new evaluation metrics for many subgroups (Wang et al., 2022) and an analysis of the desiderata of an adequate solution (Himmelreich et al., 2024). Popoola et al. (2026) combine empirical-Bayes shrinkage with worst-case weighting in a metric for explanation fairness across intersectional subgroups. These contributions treat intersectional estimation as a standalone problem; none embeds it in a wider index that also covers procedure and time. Procedural justice has long been central to legitimacy (Tyler, 2006). In the algorithmic setting, experimental work shows that explanations, outcome control and contestability shape perceived fairness (Lee et al., 2019; Yurrita et al., 2023), while legal and philosophical analyses identify a "procedural gap" in which fairness metrics speak to

distributive justice but not to voice, consent and the right to be heard (Kinchin, 2024; Decker et al., 2024). This body of work is largely perceptual, conceptual or legal. Quantitative treatments of procedure, such as the explanation-stability metric of Popoola et al. (2026), concern how the model reasons for different groups, not whether people in different groups actually receive review, appeal and timely redress.

**Gaps Addressed**

Table 1 summarizes five gaps that, within the literature we examined, remain open when the object of assessment is an AI-driven governance pipeline. We emphasise that individual ingredients of CGBI (cost-weighted errors, shrinkage, worst-case emphasis and trend monitoring) have precedents. The claim is not that each ingredient is new, but that no source we found specifies and evaluates them jointly as a pipeline-level index with a non-compensatory rule.

## 2. The Composite Governance Bias Index

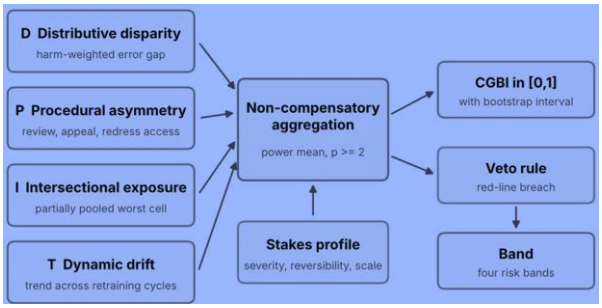


**Figure 1. Structure of CGBI. Four component measures are normalized using a stakes profile, aggregated by a non-compensatory power mean, and reported with a red-line veto and an ordinal band**

**Setting and Notation**

A system takes decisions on individuals $i$ and issues an adverse decision $\hat{Y}_i = 1$ (for example, flagged, denied or escalated) or a favorable decision $\hat{Y}_i = 0$. Where an audit outcome $Y_i$ is available (for example, confirmed fraud), false-positive and false-negative rates can be computed. Each individual has protected attributes $A = (A_1, \ldots, A_k)$; a group $g$ is a level of one attribute and a cell $s$ is a cross-classification of all attributes. For adverse decisions the pipeline records five procedural indicators: *explanation delivered*, *human review conducted*, *appeal filed*, *appeal upheld* (given filing) and *resolved within the service level time* (given filing). Data are organized in $W$ monitoring windows (for example, retraining cycles) indexed by $t$. The audit is conducted at the latest window $W$.

**Distributive Disparity (D)**

Let $c_{FP}$ and $c_{FN}$ be the harm costs of a wrongful adverse decision and a missed case. The cost-weighted error of group $g$ is

$$E_g = (c_{FP} \cdot FPR_g + c_{FN} \cdot FNR_g) / (c_{FP} + c_{FN}) \quad (1)$$

and distributive disparity is the largest within-attribute range of this burden:

$$D = \max_k [\, \max_{g \in Ak} E_g - \min_{g \in Ak} E_g \,] \quad (2)$$

$D$ lies in [0,1] and can be read as the percentage-point gap in expected harm-weighted error between the best- and worst-treated groups of an attribute.

**Procedural Asymmetry (P)**

For each procedural indicator $c$, let $p_{c,g}$ be the rate for group $g$ among the relevant adverse decisions and $p_{c,0}$ the pooled rate. The distance between a group and the pooled experience is the Jensen–Shannon distance (the square root of the base-2 Jensen–Shannon divergence) between the corresponding Bernoulli distributions. Groups with fewer than 10 relevant cases are skipped. Then

$$P_c = \max_{k,g} d_{JS}(\, p_{c,g}, p_{c,0} \,), \quad P = (1/5) \Sigma_c P_c \quad (3)$$

$P$ equals 0 when every group faces identical procedure, lies in [0,1], and is the component for which standard bias metrics have no counterpart.

**Intersectional Exposure (I)**

To read small cells with calibrated caution, cell-level error rates are partially pooled towards the overall rates. With pooled rates $f_0$ (false positive) and $n_0$ (false negative), prior strength $\kappa$, and cell counts $fp_s$, $tn_s$, $fn_s$, $tp_s$, the posterior draws are $FPR_s \sim \text{Beta}(\kappa f_0 + fp_s, \kappa(1 - f_0) + tn_s)$ and $FNR_s \sim \text{Beta}(\kappa n_0 + fn_s, \kappa(1 - n_0) + tp_s)$. Let $E_s^{(m)}$ be the cost-weighted error of draw $m$, $E_0$ the pooled cost-weighted error and $q_s$ the 5th percentile of $E_s^{(m)} - E_0$. Then

$$I = \max\{\, 0, \max_s q_s \,\} \quad (4)$$

$I$ is therefore the excess burden of the worst-off cell that the data support with 95% posterior credibility; it is zero when no cell can be credibly distinguished from the overall rate. We use $\kappa = 20$ and 300 draws.

**Dynamic Drift (T)**

For each $j \in \{D, P, I\}$, let $\text{slope}_j$ be the least-squares slope of the raw component across the $W$ windows, and $\Delta_j = (W - 1) \cdot \text{slope}_j$ the projected change over the

monitoring horizon. With saturation points $\tau_j$ and $\tau_T$ (Section 3.6),

$$T = min\ \{\ 1,\ max_j\ max(0,\ \Delta_j / \tau_j) / \tau_T\ \} \quad (5)$$

Only worsening counts. A feedback diagnostic $F$, the correlation between the adverse-decision rate gap in window $t$ and the change in $D$ in window $t + 1$, is reported alongside but is not part of the index.

**Normalization and the Stakes Profile**

Each raw component is mapped to [0,1] by a saturation point, the raw value regarded as maximal concern: $x_D = \min(1, D/\tau_D)$, and similarly for $P$ and $I$, with $x_T = T$. A *stakes profile* rates the deployment context on severity, reversibility and scale of harm, using a structured expert and stakeholder (Delphi) process, and fixes the harm costs, saturation points, weights $w = (w_D, w_P, w_I, w_T)$ with $\Sigma w = 1$, and the red-line level $r$. Table 2 gives the defaults used here, which represent a punitive context such as fraud flagging, where a wrongful flag is three times as costly as a missed case. They are placeholders for a real elicitation, and Section 5.5 tests sensitivity to them.

**Table 2. Default CGBI configuration used in the simulation study**

| Parameter | Default | Meaning |
|---|---|---|
| Harm costs $c_{FP} : c_{FN}$ | 3 : 1 | Wrongful adverse decision three times as harmful as a missed case |
| Saturation points $\tau_D$, $\tau_P$, $\tau_I$ | 0.25 each | Raw value treated as maximal concern |
| Saturation point $\tau_T$ | 0.40 | Projected worsening of 0.4 on the normalised scale treated as maximal |
| Weights ($w_D$, $w_P$, $w_I$, $w_T$) | (0.30, 0.25, 0.25, 0.20) | Relative priority of dimensions in this context |
| Exponent p | 3 | Degree of non-compensation (p = 1 is the arithmetic mean) |
| Red-line level r | 0.60 | Any normalised component at or above r triggers the veto |
| Prior strength κ; posterior draws | 20; 300 | Partial pooling of intersectional cells |

**Nnon-Compensatory Aggregation, Veto and Bands**

$$CGBI = ( \Sigma_j\ w_j \cdot x_j^p )^{1/p},\quad x = (x_D, x_P, x_I, x_T) \in [0,1]^4,\quad p \geq 1 \quad (6)$$

$$V = 1[\ max_j\ x_j \geq r\ ] \quad (7)$$

With $p > 1$ the index penalises an uneven profile more heavily than an even one with the same arithmetic mean. The veto $V$ classifies a system as high risk whenever any single component reaches the red line, regardless of the index value. Systems are placed in four bands: Low (CGBI ≤ 0.133), Moderate (≤ 0.35), Elevated (≤ 0.55) and Critical (> 0.55 or $V$ = 1). The Low/Moderate boundary is the 95th percentile of CGBI among bias-free simulated pipelines; the upper cut points are conventional round values chosen before the analysis and have no empirical justification beyond that.

**Formal Properties**

**Proposition 1 (Bounds and Monotonicity).** For $p \geq 1$ and weights summing to one, $\Sigma_j\ w_j x_j \leq$ CGBI $\leq \max_j\ x_j$, CGBI $\in [0,1]$, and CGBI is non-decreasing in each $x_j$.

**Proposition 2 (Non-Compensation Floor).** For every $j$, CGBI $\geq w_j^{1/p}\ x_j$. Hence a component at its saturation point ($x_j = 1$) forces an index of at least $w_j^{1/p}$, however good the other components are, whereas the arithmetic mean guarantees only $w_j$.

**Proposition 3 (Limit).** As $p \to \infty$, CGBI → max over $\{j: w_j > 0\}$ of $x_j$; at $p = 1$ it is the weighted arithmetic mean. The exponent thus indexes a continuum between full compensation and the worstdimension rule. Proofs are in Appendix A, and Table 3 illustrates the properties numerically.

**Table 3. Illustrative component profiles: arithmetic mean versus CGBI at different exponents (default weights).**

| Profile ($x_D$, $x_P$, $x_I$, $x_T$) | Components | Mean | p = 2 | p = 3 | p = 4 | p = 6 | Veto |
|---|---|---|---|---|---|---|---|
| Severe single dimension (D only) | (1, 0, 0, 0) | 0.30 | 0.55 | 0.67 | 0.74 | 0.82 | Yes |
| Severe drift only (T only) | (0, 0, 0, 1) | 0.20 | 0.45 | 0.58 | 0.67 | 0.76 | Yes |
| One serious, others small | (0.6, 0.1, 0.1,0.1) | 0.25 | 0.34 | 0.40 | 0.44 | 0.49 | Yes |
| Uniformly moderate | (0.3, 0.3, 0.3,0.3) | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 | No |
| Uniformly low | (0.1, 0.1, 0.1,0.1) | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | No |

A system with one fully saturated dimension and three perfect ones has the same arithmetic mean (0.30) as a uniformly moderate system, but CGBI at p = 3 is 0.67 against 0.30, and the veto fires.

## 3. Simulation Study Design

### Rationale

In real governance systems the true extent and type of bias are unknown, and procedural records by group are rarely public. A synthetic pipeline with injected bias of known type and magnitude allows an index to be assessed against ground truth: does it detect the bias, name its source, resist reasonable changes in its settings and quantify its own uncertainty? We attempted to retrieve a public criminal justice dataset for an empirical illustration, but the analysis environment refused the connection to the data repository, so no real-data results are reported. The study is therefore a validation on synthetic data only.

### Data-Generating Process

**Table 4. Data-generating process for one monitoring window**

| Element | Specification |
|---|---|
| Population | N = 8,000 decisions per window (also 2,000 and 20,000 in the sample-size experiment) |
| Protected attributes | $A_1$ binary (50/50); $A_2$ three levels (60/28/12%); $A_3$ binary (80/20%); 12 intersectional cells, the smallest about 1.2% of the population (about 96 records) |
| Outcome | Y ~ Bernoulli(0.20), equal across groups except in scenario S0b |
| Score and decision | Score = $1.6 \cdot Y + \varepsilon$, $\varepsilon \sim N(0,1)$; adverse decision if score > 1.0.<br>Without bias: true-positive rate about 0.73, false-positive rate about 0.16 |
| Harm costs | $c_{FP} = 3$, $c_{FN} = 1$ |
| Pipeline (adverse decisions) | Explanation delivered 0.90; human review 0.80; appeal filed 0.50; appeal upheld 0.45 and resolved on time 0.75 (the last two among appeals filed) |
| Monitoring horizon | W = 6 windows; audit at window 6, trend from all windows (W = 3, 4, 5 in the early-audit experiment) |

### Scenarios

**Table 5. Bias scenarios. Severity multiplies every injected magnitude (main results at severity 1).**

| Scenario | Injection at severity 1 | Replications |
|---|---|---|
| S0 Null | No bias | 500 (N = 8,000); 200 at each other N |
| S0b Base-rate gap | No bias, but base rates of 0.15, 0.25 and 0.35 across levels of $A_2$ (a fair system for which parity must fail) | 300 |
| S1 Distributive | Score shifted by +0.5 SD for the smallest level of $A_2$ (12% of cases), in every window | 300; 100 at other severities |

| Scenario | Injection at severity 1 | Replications |
|---|---|---|
| S2 Procedural | Model unbiased; for the same group, explanation −25, review −25, appeal filing −20, appeal upheld −15 and on-time resolution −25 percentage points | 300; 100 at other severities |
| S3 Intersectional | Score shifted by +1.0 SD in a single cell (about 1.2% of the population); marginal groups are barely affected | 300; 100 at other severities |
| S4 Dynamic drift | Feedback loop: shift in window 1 is 0.10 SD; $shift_{t+1}$ = $shift_t$ + 1.0 × max(0, adverse-rate $gap_t$) (median final shift 0.34 SD) | 300; 100 at other severities; 300 at each W in early audit |
| S5 Combined | Feedback loop starting at 0.15 SD, plus S3 at 0.6 SD, plus S2 at half intensity | 300; 100 at other severities |

In total the study comprises 5,300 main pipelines (severities 0 to 2, three sample sizes), 2,850 pipelines in the early-audit experiment, and 160 datasets with 150 bootstrap resamples each in the interval coverage study. All simulations used fixed seeds. The scenario magnitudes were chosen by the author before the analysis to give effects of moderate size. In S4 the feedback loop produced a median final shift of 0.34 SD, a moderate rather than runaway accumulation.

**Comparators and Evaluation Protocol**

Comparators were (i) the equalized-odds gap (the larger of the true-positive-rate and false-positive-rate ranges, maximised over attributes); (ii) the demographic-parity gap (range of adverse-decision rates); (iii) the four-fifths rule applied to favourable-decision rates; (iv) an unshrunk subgroup equalized-odds gap over the 12 cells (cells with at least five positives and five negatives); and (v) an arithmetic-mean composite using the same normalization and weights as CGBI (that is, $p = 1$ and no veto).

The primary evaluation is *null-calibrated detection*. Each metric was thresholded at its own 95th percentile among the 500 bias-free pipelines of the same sample size, so every metric has a 5% false alarm rate by construction, and detection rate is the share of biased pipelines above the threshold (Wilson 95% intervals in Appendix B). This puts metrics measured on different scales on an equal footing. We also report practitioner screens with fixed thresholds (four-fifths ratio ≥ 0.8 and equalized odds gap ≤ 0.10).

Two hypotheses were specified in the research: H1, that CGBI flags pipelines that pass conventional single-metric screens; and H2, that rankings of systems by CGBI are stable under plausible weight perturbations (we use Kendall's τ ≥ 0.85 as the benchmark for "stable", a conventional rather than pre-registered value). Three further questions were added as exploratory: E1, whether non compensatory aggregation improves detection over an arithmetic mean; E2, whether the trend component gives earlier warning of drift; and E3, whether bootstrap intervals achieve nominal coverage.

Robustness was assessed by drawing 2,000 weight vectors from a Dirichlet distribution centred on the default weights (concentration 10) with exponents $p \in \{2, 3, 4, 6\}$, recalibrating thresholds for each configuration, and comparing rankings of 27 scenario-by-severity cells by Kendall's τ; by scaling all saturation points by 0.6 to 1.5; by varying the prior strength $\kappa \in \{5, 20, 50\}$; and by varying the exponent from 1 to 6.

## 4. Results

**Component Recovery and Diagnosis**

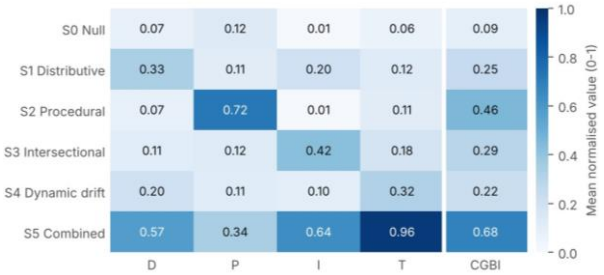


**Figure 2. Mean normalized component values and CGBI by scenario at severity 1 (N = 8,000). Each injected bias type raises its intended component**

In bias-free pipelines the mean normalized components were low ($x_D = 0.07$, $x_P = 0.12$, $x_I = 0.01$, $x_T = 0.06$; CGBI 0.09). Each single-type scenario raised its intended component: $x_D$ to 0.33 in S1, $x_P$ to 0.72 in S2, $x_I$ to 0.42 in S3 and $x_T$ to 0.32 in S4 (Figure 2). Some cross-talk is expected and appropriate: the affected group in S1

contains intersectional cells, so $x_I$ also rose (0.20), and drift in S4 raised the level of $D$ (0.20) as well as its trend.

To test whether the component report identifies the source of bias, we assigned each pipeline to the component with the largest excess over its own bias-free 95th percentile (Table 6). The matching component was chosen in most pipelines for every type, most reliably for procedural bias, and least reliably for drift, where the level of $D$ competes with its trend.

**Table 6. Diagnosis: share of pipelines (severity 1) in which each component had the largest excess over its bias-free 95th percentile**

| Injected bias | D | P | I | T | None above null |
|---|---|---|---|---|---|
| S1 Distributive | **88%** | 0% | 9% | 3% | 0% |
| S2 Procedural | 0% | **100%** | 0% | 0% | 0% |
| S3 Intersectional | 0% | 0% | **94%** | 6% | 0% |
| S4 Dynamic drift | 19% | 0% | 3% | **68%** | 10% |

**Detection Compared with Conventional Metrics**

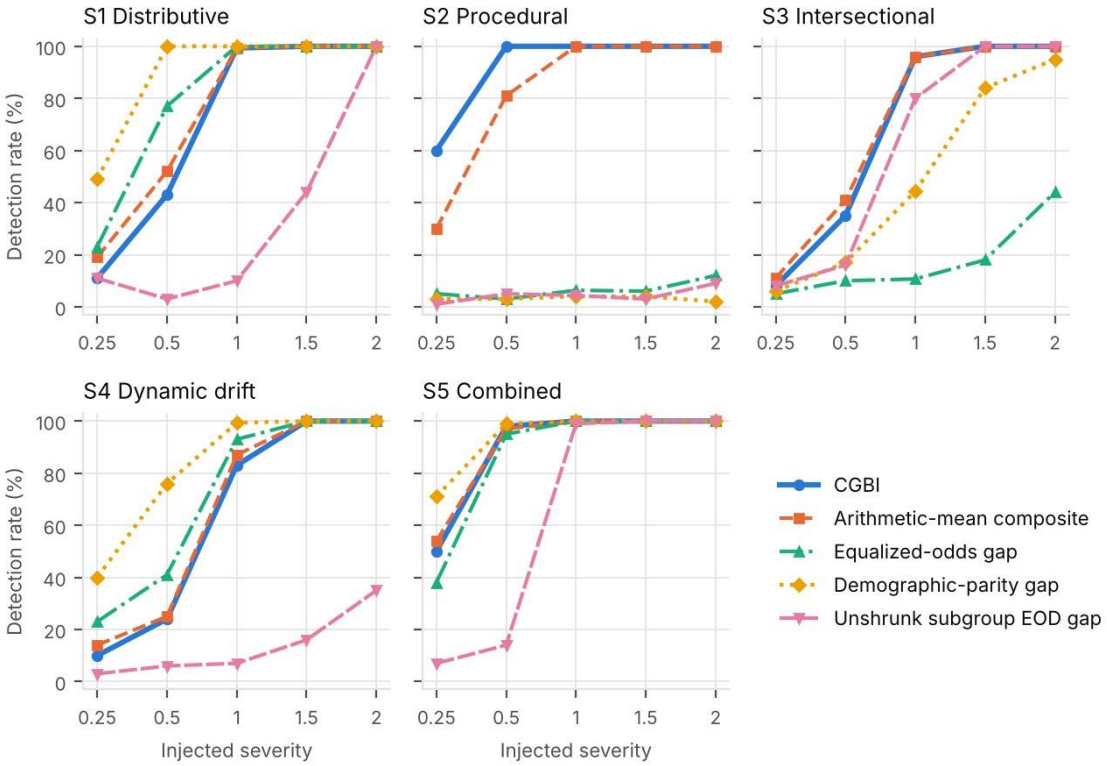


**Figure 3. Null-calibrated detection rate (5% false alarms) against injected severity. Severity 1 uses 300 replications per point; other severities use 100**

**Table 7. Null-calibrated detection rates (%) at severity 1 (N = 8,000; 300 replications per scenario, 500 bias-free pipelines for thresholds). The Null row is 5% by construction**

| Scenario | CGBI | Mean composite | Equalized odds | Parity gap | Four-fifths | Subgroup EOD |
|---|---|---|---|---|---|---|
| S0 Null | 5 | 5 | 5 | 5 | 5 | 5 |
| S1 Distributive | 99 | 99 | 100 | 100 | 100 | 10 |
| S2 Procedural | 100 | 100 | 6 | 4 | 4 | 4 |
| S3 Intersectional | 96 | 96 | 11 | 44 | 48 | 80 |
| S4 Dynamic drift | 83 | 87 | 93 | 99 | 99 | 7 |
| S5 Combined | 100 | 100 | 100 | 100 | 100 | 99 |

**Distributive Bias (S1).** CGBI detected 99% of pipelines, similar to equalized odds (100%) and demographic parity (100%); the unshrunk subgroup metric detected only 10% because its threshold is inflated by noise in small cells. At lower severities CGBI was less sensitive than the parity metrics (severity 0.5: CGBI 43%, equalized odds 77%, parity gap 100%), the price of an omnibus index whose null distribution includes noise from the other components.

**Procedural Bias (S2).** CGBI detected 100%, while every conventional metric stayed at its false-alarm level (equalized odds 6%, parity gap 4%). This result holds by construction: the model is unbiased, so output-based metrics cannot see the injected bias, and *P* measures precisely the pipeline variables in which it was injected. It demonstrates coverage of a blind spot rather than superior accuracy on a shared task. Even at severity 0.25 CGBI detected 60% of cases.

**Intersectional Bias (S3).** CGBI detected 96%, against 11% for equalized odds, 44% for the parity gap, 48% for the four-fifths ratio and 80% for the unshrunk subgroup metric. Marginal metrics miss a harm that concentrates in 1.2% of the population, and partial pooling lets CGBI keep a much lower noise floor than the unshrunk cell-level alternative.

**Drift (S4) and Combined Bias (S5).** For drift CGBI detected 83%, below equalized odds (93%) and the parity gap (99%), because by the final window the accumulated disparity is already visible to static metrics (Section 5.6). For combined bias all metrics except the unshrunk subgroup gap detected essentially every pipeline.

**Conventional Screens with Fixed Thresholds**

Table 8 applies practitioner thresholds. Conventional screens passed 99% of bias-free systems, as they should, but also 99% of procedural-only and 98% of intersectional-only systems. CGBI flagged 100% and 97% of those passing systems respectively, compared with 5% among passing bias-free systems. Hypothesis H1 is therefore supported. The conventional screens did flag most distributive and combined systems, and they passed 22% of drifting systems, of which CGBI flagged 76%.

**Table 8. Conventional screens (four-fifths ratio ≥ 0.8 and equalized-odds gap ≤ 0.10) and CGBI (flag above null-calibrated threshold), severity 1**

| Scenario | Pass conventional screens | CGBI flags among passing | CGBI flags overall | Veto fires |
|---|---|---|---|---|
| S0 Null | 99% | 5% | 5% | 0% |
| S0b Base-rate gap | 99% | 4% | 4% | 0% |
| S1 Distributive | 0% | n/a (none pass) | 99% | 0% |
| S2 Procedural | 99% | 100% | 100% | 97% |
| S3 Intersectional | 98% | 97% | 96% | 11% |
| S4 Dynamic drift | 22% | 76% | 83% | 3% |
| S5 Combined | 0% | n/a (none pass) | 100% | 99% |

**Non-Compensation: Detection Versus Classification**

Exploratory question E1 asked whether the non-compensatory power mean detects more bias than an arithmetic mean of the same normalized components. The answer is mixed. At severity 1 the two differ by at most 4 points in null-calibrated detection in any scenario. Across the full severity grid, CGBI was clearly better in one case, mild procedural-only bias (S2 Procedural, severity 0.25: 60% against 30%; 100% against 81% at severity 0.5), where one component is strongly affected and the others are not. In the other scenarios the arithmetic mean was equal or slightly better at low severity, by up to 9 points (S1 Distributive, severity 0.5), because averaging pools noise less severely than a power mean that amplifies the noisiest component. The advantage of non-compensation shows up elsewhere, in how a given profile is *classified* (Table 9). In S2, where only the procedural dimension is strongly affected, the red-line veto fired in 97% of pipelines and the power mean alone placed 100% in the Elevated or Critical bands, whereas an arithmetic mean with the same cut points placed only 0% there. For S3 the corresponding shares were 21% and 4%. Because the cut points are conventional, the band comparison depends on scale and illustrates the mechanism rather than measuring an accuracy gain.

**Table 9. Share of pipelines (severity 1) classified Elevated or Critical: CGBI without veto, CGBI with veto, and an arithmetic-mean composite using the same cut points (Low/Moderate boundary set at each metric's own null 95th percentile)**

| Scenario | CGBI (no veto) | CGBI with veto | Arithmetic mean | Veto fires |
|---|---|---|---|---|
| S1 Distributive | 1% | 1% | 0% | 0% |
| S2 Procedural | 100% | 100% | 0% | 97% |
| S3 Intersectional | 21% | 21% | 4% | 11% |
| S4 Dynamic drift | 6% | 6% | 1% | 3% |
| S5 Combined | 100% | 100% | 100% | 99% |

The band distribution of CGBI itself (Low/Moderate/Elevated/Critical) was 95% / 5% / 0% / 0% for bias-free pipelines, 1% / 98% / 1% / 0% for S1 and 0% / 0% / 3% / 97% for S2. The latter shows a calibration issue that deserves attention: distributive bias of 0.5 SD (a harm-weighted error gap of about 8 percentage points) is rated Moderate, whereas procedural bias of 15 to 25 points in access is rated Critical. Whether that ordering is appropriate depends on the saturation points, which is exactly the judgement the stakes profile is meant to capture and which cannot be settled by simulation.

**Convergent and Discriminant Validity**

Across all pipelines with N = 8,000, $x_D$ correlated strongly with the equalized-odds gap (Spearman ρ = 0.83) and the parity gap (ρ = 0.79); $x_I$ correlated most strongly with the unshrunk subgroup metric (ρ = 0.78); and $x_P$ was essentially unrelated to all four conventional metrics (ρ between -0.14 and 0.04). This is the pattern expected if *D* and *I* measure outcome disparity, while *P* measures something conventional metrics do not.

**Table 10. Spearman correlations between CGBI components and conventional metrics (all pipelines, N = 8,000).**

| Component | Equalized odds | Parity gap | Subgroup EOD | Four-fifths (1−ratio) |
|---|---|---|---|---|
| $x_D$ (distributive) | 0.83 | 0.79 | 0.59 | 0.79 |
| $x_P$ (procedural) | -0.03 | -0.13 | 0.04 | -0.14 |
| $x_I$ (intersectional) | 0.68 | 0.66 | 0.78 | 0.67 |
| $x_T$ (dynamic) | 0.53 | 0.49 | 0.41 | 0.49 |

**Robustness**

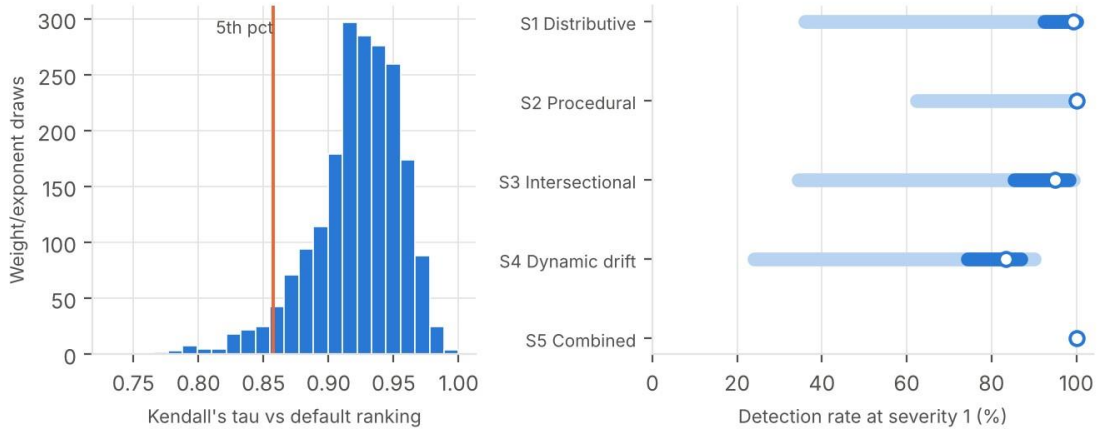


**Figure 4. Left: Kendall's τ between the default ranking of 27 scenario-by-severity cells and the ranking under each of 2,000 random weight and exponent configurations. Right: detection rate at severity 1 across the same configurations (dark bar: 5th to 95th percentile; light bar: minimum to maximum; circle: median)**

**Weights and Exponent:** Rankings were stable: mean τ = 0.92, 5th percentile 0.86, minimum 0.73, and 79% of configurations exceeded 0.90 (H2 supported against the 0.85 benchmark). Detection, however, can degrade under extreme settings. The 5th to 95th percentile ranges were 92%–100% for S1, 85%– 98% for S3 and 74%–87% for S4, with minima of 36%, 34% and 24%, which we attribute to configurations that gave little weight to the dimension in which the bias lay (we did not test this explanation directly). Weights are therefore not innocuous, which is a further reason to elicit them transparently.

**Other Settings:** Table 11 shows that detection was insensitive to the exponent over 1 to 6 (S4 87% to 83%; other scenarios within 4 points), to the scale of the saturation points (S3 92% to 97%), and only mildly to prior strength (S3 97% at $\kappa$ = 5, 96% at 20, 92% at 50, as stronger pooling shrinks a genuine cell effect). Sample size mattered much more.

**Table 11. Detection rate (%) at severity 1 under alternative settings and sample sizes (null-calibrated for each)**

| Setting | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Default (p = 3, τ × 1.0, κ = 20) | 99% | 100% | 96% | 83% | 100% |
| Exponent p = 1 (arithmetic mean) | 99% | 100% | 96% | 87% | 100% |
| Exponent p = 2 | 99% | 100% | 95% | 84% | 100% |
| Exponent p = 6 | 99% | 100% | 96% | 83% | 100% |
| Saturation points × 0.6 | 95% | 100% | 92% | 83% | 100% |
| Saturation points × 1.5 | 100% | 100% | 97% | 83% | 100% |
| Prior strength κ = 5 | 99% | 100% | 97% | 83% | 100% |
| Prior strength κ = 50 | 99% | 100% | 92% | 83% | 100% |
| N = 2,000 per window | 40% | – | 24% | – | 99% |
| N = 20,000 per window | 100% | – | 100% | – | 100% |

**Sample Size:** At N = 2,000 per window CGBI detected 40% of S1 pipelines compared with 78% for equalized odds, so for a single distributive bias in a small system the dedicated metric is more sensitive. For the intersectional scenario CGBI (24%) still exceeded equalized odds (10%) and the unshrunk subgroup metric (10%), though detection was low in absolute terms. At N = 20,000 CGBI detected all S1 and S3 pipelines while equalized odds still detected only 30% of S3.

**Base-Rate Differences:** In S0b the groups have different base rates but the system is fair in the sense of equal error rates. At thresholds calibrated on the equal-base-rate null, the parity gap and four-fifths ratio flagged 100% and 100% of these fair systems, whereas CGBI flagged 4%, equalized odds 1% and the unshrunk subgroup metric 1%. This replicates the known sensitivity of parity metrics to base-rate differences and shows that cost-weighted error components are not affected in the same way.

**Dynamic Drift and Early Warning**

Exploratory question E2 asked whether the trend component gives earlier warning of worsening bias than static metrics. We audited drifting systems after only 3, 4 or 5 windows (Table 12). It did not: demographic parity was the most sensitive indicator at every horizon, and CGBI was at or below equalized odds. The trend component alone ($x_T$) flagged 17%, 31% and 57% of drifting systems at 3, 4 and 5 windows, close to what CGBI itself achieved. In this design the feedback loop acts directly on the adverse-decision rate, which is exactly what the parity gap measures, and few windows give a noisy slope.

**Table 12. Early-audit detection (%) of the drifting scenario (S4) after W windows; thresholds calibrated on 500 bias-free pipelines with the same W**

| Windows W | Median final shift (SD) | CGBI | Mean composite | Equalized odds | Parity gap | $x_T$ alone |
|---|---|---|---|---|---|---|
| 3 | 0.16 | 17 | 18 | 29 | 75 | 17 |
| 4 | 0.20 | 35 | 36 | 45 | 89 | 31 |
| 5 | 0.26 | 61 | 66 | 66 | 97 | 57 |
| 6 | 0.34 | 83 | 87 | 93 | 99 | 82 |

Two further points weaken the case for the trend component as specified. First, its specificity is imperfect: at severity 1, $x_T$ exceeded its bias-free 95th percentile in 28% of static distributive, 26% of static procedural and 36% of static intersectional pipelines, because the sampling noise of range- and maximum-type components grows with their level. Second, the feedback diagnostic *F* was only weakly informative with six windows: its median was 0.18 in drifting systems against -0.05 in bias-free systems (positive in 59% versus 47% of pipelines), and clearer only in the combined scenario (median 0.48).

**Uncertainty and Interval Coverage**

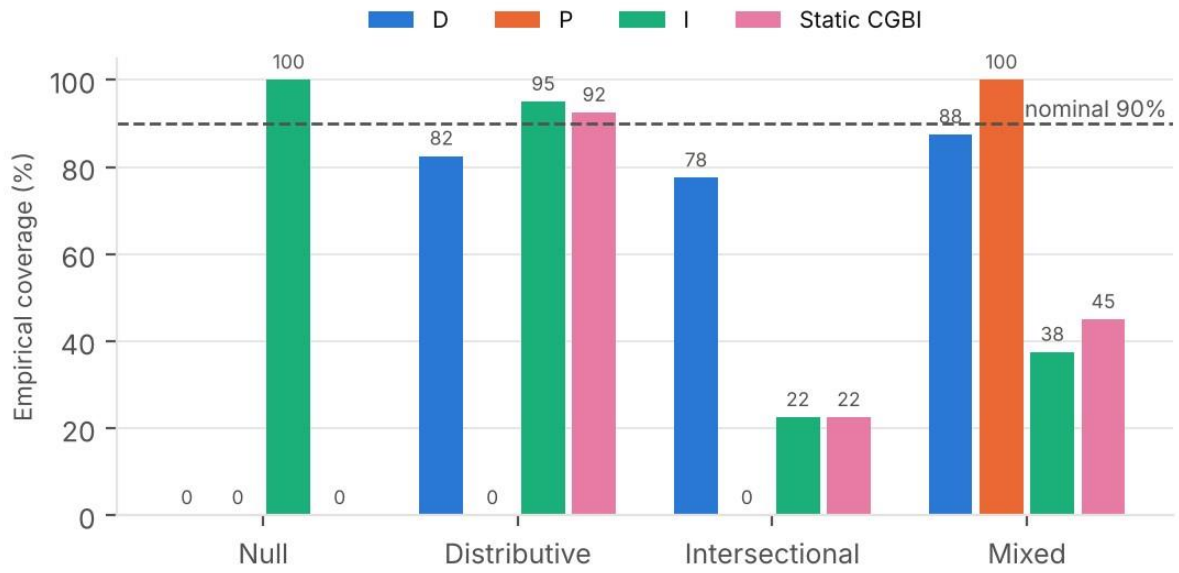


**Figure 5. Empirical coverage of 90% percentile-bootstrap intervals for the population value of each component and of the static CGBI (D, P, I only), 40 datasets per case, 150 resamples each, N = 8,000. Population values were computed from a window of 600,000 records**

Exploratory question E3 asked whether bootstrap percentile intervals cover the population value at the nominal 90% level. They do not in general (Figure 5, Table 13). Intervals for *D* and *P* are range- and maximum-type statistics whose finite-sample estimates are biased upward (mean bias in bias-free data +0.017 and +0.027, against population values near zero), so an interval centred on the estimate cannot reach the truth, and coverage was 0% for both in the null case. The shrunk, one-sided component *I* has the opposite problem: it is biased downward when a genuine cell effect is strong (mean bias -0.081 in S3), and coverage fell to 23%. Coverage of the static CGBI was 93% in S1 but 23% in S3, 45% in the mixed case and 0% in the null case. The intervals do convey sampling variability, but they should not be read as confidence intervals for the true value. A null-referenced reading, comparing the observed value with its distribution in a simulated or permuted bias-free system, is the safer interpretation and is what the null-calibrated results in this paper use.

**Table 13. Coverage (%) of 90% percentile-bootstrap intervals and mean bias of the estimate (estimate minus population value), 40 datasets per case**

| Case | D | P | I | Static CGBI |
|---|---|---|---|---|
| Null | 0% (+0.017) | 0% (+0.027) | 100% (+0.001) | 0% (+0.084) |
| Distributive (S1) | 83% (+0.000) | 0% (+0.024) | 95% (−0.021) | 93% (−0.017) |
| Intersectional (S3) | 83% (+0.000) | 0% (+0.025) | 23% (−0.081) | 23% (−0.214) |
| Mixed | 88% (+0.004) | 100% (+0.002) | 38% (−0.073) | 45% (−0.137) |

**Illustrative Audit Card**

Table 14 shows what a regulator-facing report would contain for one simulated system with mixed bias and for one bias-free system, using the static components. The bias-free example also displays the interval problem described above: the estimate of *P* lies outside its own bootstrap interval.

**Table 14. Example audit card: raw component estimates with 90% bootstrap intervals (static CGBI uses D, P and I with weights renormalized).**

| Quantity | Mixed-bias system | Bias-free system |
|---|---|---|
| Distributive disparity D | 0.087 [0.066, 0.114] | 0.012 [0.009, 0.042] |
| Procedural asymmetry P | 0.084 [0.072, 0.103] | 0.019 [0.024, 0.045] |
| Intersectional exposure I | 0.051 [0.033, 0.099] | 0.000 [0.000, 0.023] |
| Static CGBI | 0.312 [0.273, 0.407] | 0.057 [0.078, 0.146] |

## 5. Discussion

**Table 15. Scorecard of hypotheses and exploratory questions**

| ID | Statement | Evidence | Verdict |
|---|---|---|---|
| H1 | CGBI flags systems that pass conventional screens | 99% of procedural-only and 98% of intersectional-only systems passed conventional screens; CGBI flagged 100% and 97% of them (procedural result is true by construction) | Supported |
| H2 | Rankings stable under weight perturbation ($\tau \geq 0.85$) | Mean $\tau$ 0.92, 5th percentile 0.86; but detection fell to 24%–36% in extreme configurations | Supported, with caveats |
| E1 | Non-compensation improves detection over arithmetic mean | No difference at severity 1 (within 4 points); power mean better for mild concentrated bias (60% vs 30%, procedural, severity 0.25), up to 9 points worse elsewhere; classification of red-line breaches differs | Mixed |
| E2 | Trend component gives earlier warning of drift | Parity gap was more sensitive at every horizon; $x_T$ lacks specificity under static bias | Not supported |
| E3 | Bootstrap intervals achieve nominal coverage | Coverage ranged from 0% to 100% across components and cases; severe under-coverage for D, P in null data and I, CGBI in S3 | Not supported |

**Interpretation and Guidance for Use**

Taken together, the results support a particular role for CGBI: an omnibus *screening and diagnostic* instrument. It extends coverage to harms in pipeline access and in small intersections that standard output metrics do not see, it does not mistake base-rate differences for bias, its component report points to the likely source, and its ranking of systems is stable under reasonable weight choices. It is not a more powerful detector of a single distributive bias than the dedicated metric, especially in small samples; where one hypothesis is of interest, the corresponding single metric remains the better test.

The non-compensatory exponent improves detection only for mild bias concentrated in one dimension; its main value is in the reporting policy, because a system with one dimension at its red line should not be rated like one that is uniformly moderate (Table 3, Table 9).

Four practical recommendations follow. First, interpret CGBI against a null reference (simulated or permuted bias-free systems for the same context and sample size) rather than against fixed numeric bands, and treat the bands in Section 3.7 as provisional. Second, always publish the component report alongside the index. Third, elicit the stakes profile (costs, saturation points, weights and red lines) through a documented participatory process, since conclusions about *which* bias is most serious depend on it (Section 5.3). Fourth, do not rely on the current trend component for early warning; use a dedicated monitoring chart on parity and error rates until a better-specified drift statistic is available.

**Implications for Governance and Policy**

The procedural component can only be computed if agencies log, by group, whether explanations were delivered, whether a human reviewed the decision, whether an appeal was filed and with what result, and how long redress took. In many jurisdictions such records either do not exist or are not disaggregated. A regulatory expectation to maintain such logs, within data-protection law, would be a precondition for measuring procedural fairness at all. The stakes profile also offers a structured way to express contextual judgements, such as the higher harm of a wrongful fraud accusation, in a form auditors and affected communities can inspect and contest. CGBI is a measurement aid; a low value is evidence of lower measured bias, not proof of fairness or legal compliance, and it cannot replace qualitative review.

**Relation to Existing Composite Measures**

CGBI differs from the Fairness Score and Bias Index of Agarwal et al. (2022), which summarise fairness of a supervised-learning system across protected attributes and support a certification process, in covering pipeline procedure, intersections with calibrated uncertainty, time and context-specific stakes. It differs from the meta-metric work of Monnier et al. (2026), which analyses how aggregation choices alter conclusions about intersectional disparities, by proposing a concrete aggregation across heterogeneous dimensions and testing its robustness; their warning that aggregation choices must be justified applies equally here. The shrinkage and worst-case ideas in the intersectional component echo Popoola et al. (2026), but we apply them to cost-weighted error rather than explanation stability, and use them as one input to a wider index. Composite trust indices such as that of Jan et al. (2026) aggregate fairness with privacy, transparency and accountability; CGBI is narrower in aim (bias) but treats it as multi-dimensional and non-compensatory.

## 6. Conclusion and Future Work

We proposed the Composite Governance Bias Index, a stakes-weighted, non-compensatory index that treats bias in AI-driven governance as a property of the whole decision pipeline, and evaluated it against known ground truth in a simulation study. CGBI detected distributive, procedural, intersectional, dynamic and combined bias, identified the likely source, stayed stable under reasonable changes of weights and exponent, and did not confuse base-rate differences with bias, and it flagged most systems that conventional screens passed. At the same time, its non-compensatory aggregation improved detection over averaging only for mild, concentrated bias, gave no earlier warning of drift than demographic parity, was less sensitive than equalized odds for a single distributive bias in small samples, and was accompanied by intervals that do not achieve nominal coverage.

**References**

[1] Agarwal, A., et al. (2022). Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics*.*

[2] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*.

[3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91.

[4] Cherian, J. J., et al. (2023). Statistical inference for fairness auditing. *arXiv preprint*.*

[5] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

[6] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist

theory and antiracist politics. *University of Chicago Legal Forum, 1989*(1), 139–167.

[7] D'Amour, A., et al. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[8] Decker, M. C., et al. (2024). Procedural fairness in algorithmic decision-making: The role of public engagement. *Ethics and Information Technology*.*

[9] Dwork, C., & Ilvento, C. (2018). Fairness under composition. *arXiv preprint*.*

[10] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research, 81*, 160–171.

[11] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[12] European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

[13] Ferrara, A., et al. (2025). Size-adaptive hypothesis testing for fairness. *Advances in Neural Information Processing Systems*.*

[14] Garg, P., et al. (2020). Fairness metrics: A comparative analysis. *IEEE International Conference on Big Data*.*

[15] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*.

[16] Hardy, G. H., Littlewood, J. E., & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge University Press.

[17] Herlihy, C. P., et al. (2024). A structured regression approach for evaluating model performance across intersectional subgroups. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.*

[18] Himmelreich, J., et al. (2024). The intersectionality problem for algorithmic fairness. *arXiv preprint*.*

[19] Jan, S., et al. (2026). EAGF: a four-pillar ethical AI governance framework for trustworthy cybersecurity in 5G renewable energy IoT systems. *Scientific Reports*.*

[20] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of the 35th International Conference on Machine Learning*.

[21] Kinchin, N. (2024). "Voiceless": the procedural gap in algorithmic justice. *International Journal of Law and Information Technology*.*

[22] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science (ITCS)*.

[23] Lee, M. K., et al. (2019). Procedural justice in algorithmic fairness. *Proceedings of the ACM on Human-Computer Interaction*.*

[24] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. *Proceedings of the 35th International Conference on Machine Learning*.

[25] Majumder, S., et al. (2023). Fair enough: Searching for sufficient measures of fairness. *ACM*

[26] *Transactions on Software Engineering and Methodology*.*

[27] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6).

[28] Monnier, J., et al. (2026). Beyond base metric: The critical and overlooked role of meta-metrics in intersectional fairness. *ACM Conference on Fairness, Accountability, and Transparency*.* Munda, G. (2008). *Social multi-criteria evaluation for a sustainable economy.* Springer.

[29] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

[30] OECD/JRC. (2008). *Handbook on constructing composite indicators: Methodology and user guide.* OECD Publishing.

[31] Popoola, G., et al. (2026). MESD: A risk-sensitive metric for explanation fairness across intersectional subgroups. *arXiv preprint*.*

[32] Speicher, T., et al. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.*

[33] Tyler, T. R. (2006). *Why people obey the law.* Princeton University Press.

[34] Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness (FairWare)*.

[35] Wang, A., et al. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *ACM Conference on Fairness, Accountability, and Transparency*.*

[36] Yurrita, M., et al. (2023). Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability. *CHI Conference on Human Factors in Computing Systems*.*