When Fairness Metrics Disagree: Operational Implications for Face Recognition Systems

Kaïra Neily SANON[1,2], Joël DI MANNO[1,2], Christophe CHARRIER[2], and Christophe ROSENBERGER[2]
[1]FIME EMEA, 14000 Caen, France
[2] Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

Fairness evaluation in face recognition relies on a growing number of metrics, each designed to capture a particular form of disparity and often validated under limited or synthetic conditions. As a result, practitioners still lack clear guidance on which metric to use in a given context. We address this issue through a unified framework that evaluates nineteen fairness metrics on four public face datasets. Three complementary sources of bias are considered: algorithmic variation through four training loss functions, data-related variation through image quality degradation and dataset resampling, and operational variation through decision-threshold changes. All experiments are conducted on real biometric data under a common evaluation protocol. For the controlled data-related and operational variations, metric validity is assessed through monotonicity and stability across demographic groups. This analysis reveals how the mathematical structure of each metric determines the type of disparity to which it is most responsive. It also provides a common basis for comparing score-based and error-rate-based metrics under the same experimental conditions. The results show that metric suitability depends strongly on the source and manifestation of bias, and that no single metric is appropriate for all conditions. Based on these findings, we propose a practical guideline linking bias conditions to the metric structures best suited to detect them, supporting fairness audits and the future development of biometric standards.

Index Terms—**Face recognition, Fairness, Bias sources, Operational guidelines, Performance differentials.**

I. Introduction

BIOMETRICS technologies have reached a level of maturity that has enabled their large-scale deployment in a wide range of real-world applications, including biometric authentication, border control, access management, identity verification, and digital identity services. Despite their impressive performance gains over the past decade, these systems have repeatedly been shown to exhibit performance differentials across demographic groups, raising serious concerns regarding security, user experience, trustworthiness, and societal impact. From a security standpoint, such disparities can be exploited to compromise system integrity by targeting specific demographic groups with degraded matching scores, thereby increasing false acceptance or false rejection rates. From a user experience perspective, unequal performance translates into unequal access and service quality, undermining the reliability and inclusiveness of deployed systems. These performance differentials are commonly referred to as demographic bias and have been reported with respect to attributes such as gender, ethnicity, age, and other soft-biometric attributes [1].

In response to these concerns, the biometric community has proposed a growing catalog of fairness metrics, each grounded in a different theoretical principle: absolute error gaps, ratio-based formulations, inequality indices, distributional divergences, and separation measures. Each new metric is typically introduced in a paper that compares it against a limited subset of existing metrics, often on synthetic score distributions or under a single source of bias. The result is a fragmented comparative landscape: every metric appears justified in its own framework, but the practitioner such as auditor, certifier, or system designer has no principled way to select the appropriate metric for a given deployment context. Three structural limitations make this fragmentation problematic. First, validations relying on synthetic scores do not reproduce the complexity of real biometric distributions, where calibration effects, distribution tails, and interactions with image quality matter. Second, each comparative study typically examines fairness under a single source of bias, leaving open the question of whether a metric reliable under demographic imbalance remains reliable under image quality degradation or operational threshold shifts. Third, no comparative work provides an explicit ground truth against which metric behavior can be assessed: on real datasets, the actual bias level is unknown, and for synthetic, the realism is lost.

This work proposes a unified evaluation framework that addresses these three limitations simultaneously. We evaluate nineteen fairness metrics drawn from the literature and from ISO/IEC 19795-10 on four public face datasets and across three complementary sources of bias: algorithmic, data-related, and operational bias. For the controlled data-related and operational experiments, bias is progressively introduced into real biometric data, and the known severity ordering provides the reference for assessing metric validity through monotonicity and stability. And the algorithmic experiment is examined by comparing four training loss functions under the same downstream evaluation protocol. The contributions of this paper are fourfold:

- A unified comparative framework that evaluates nineteen fairness metrics under identical conditions on real biometric data, across three distinct sources of bias.
- A controlled bias induction protocol that provides an explicit ground truth for metric validation, avoiding both the realism loss of synthetic experiments and the ground-

Corresponding author: Ms. SANON (email: neily.sanon@unicaen.fr).

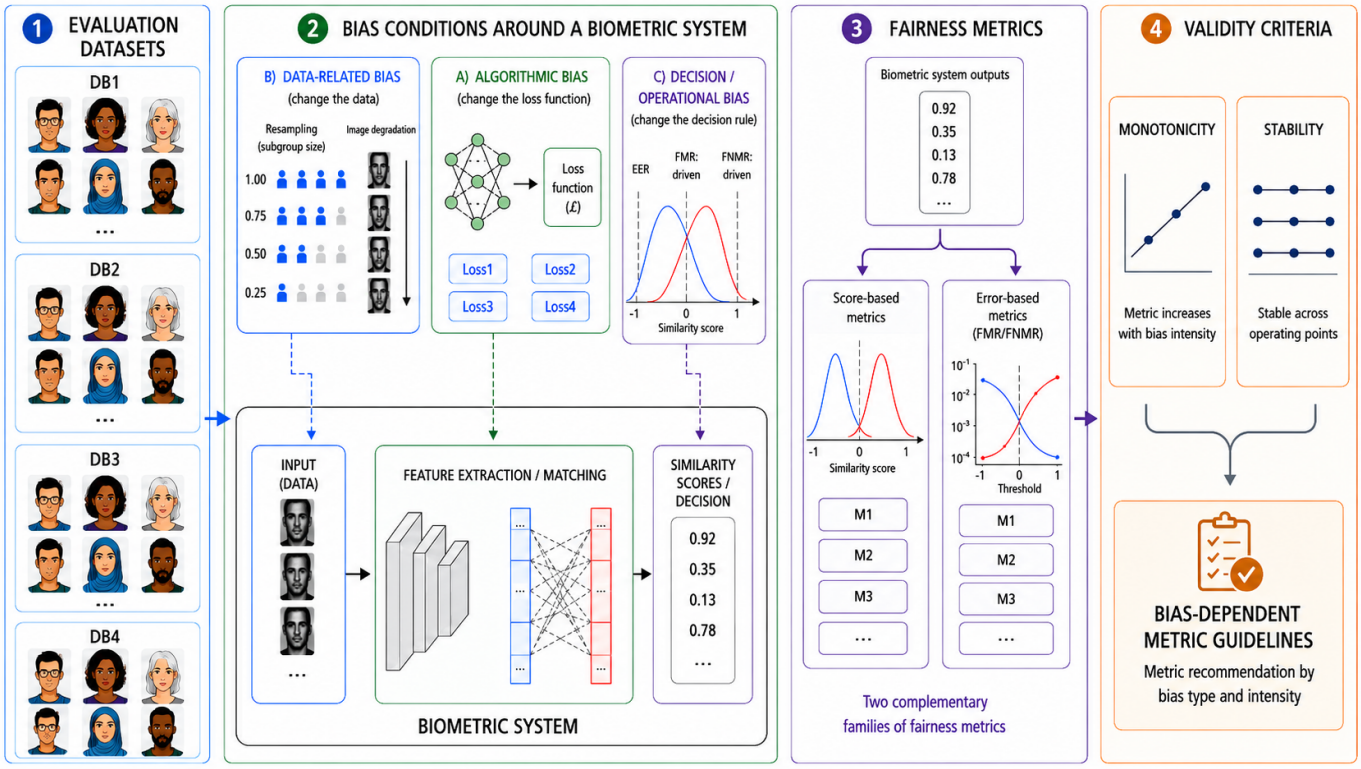

Fig. 1: Unified evaluation pipeline. Algorithmic, data-related, and operational bias-inducing conditions are introduced independently, while the downstream biometric and fairness evaluation protocol remains fixed. The resulting analysis supports the context-dependent metric guidelines reported in Table III.

truth gap of purely observational studies.
- Formal validity criteria with reproducible measurements.
- An operational guideline (Table III) mapping bias sources to appropriate metrics, derived directly from our experimental evidence and intended to support fairness audits and the future development of biometric fairness standards.

Beyond these contributions, our results reveal a more fundamental observation: the source of bias affects fairness conclusions more strongly than the choice of metric itself. A metric that satisfies the validity criteria under one bias source may fail under another, and this dependence is predictable from the mathematical structure of the metric. Fairness in biometric systems should therefore be treated not as an intrinsic system property measurable by a single metric, but as a property conditional on the operational context in which the evaluation takes place. The remainder of the paper is organized as follows. Section II reviews prior comparative studies of fairness metrics in biometric verification and positions our framework. Section III presents the methodology, including the evaluated metrics, the datasets used, and the controlled bias induction protocol. Section IV reports the experimental results across the three bias sources. Section V discusses the implications for fairness auditing, biometric standards, and operational deployment. Section VI concludes the paper and outlines directions for future work.

II. Related Work

Fairness assessment in biometric verification has received increasing attention in recent years, leading to the introduction of a growing number of dedicated bias metrics. These metrics rely on heterogeneous mathematical principles, including absolute error-rate differences, ratio-based formulations, inequality indices, distributional divergences, and score-separation measures. Each metric is generally motivated by a specific notion of disparity and evaluated against a limited subset of existing alternatives, often under restricted experimental conditions such as synthetic score distributions, a single operating point, or one dominant source of bias. This situation contrasts with the requirements of biometric verification, where system behavior is strongly affected by decision thresholds, demographic subgroups may contain limited numbers of identities, and acquisition quality can vary unevenly across populations. As a result, fairness evaluation in biometrics requires criteria that account for threshold-dependent performance, subgroup-size effects, and quality-induced distributional changes. These specificities have motivated several surveys on biometric fairness and demographic bias [2], [3], [4] and several quantitative audits of demographic accuracy gaps [5], while broader machine-learning fairness taxonomies [6] [7] remain insufficient to fully characterize the operational nature of biometric disparities.

Several representative metrics illustrate the diversity of this landscape. De Freitas Pereira and Marcel [8] introduced and validated the Fairness Discrepancy Rate (FDR) on MOBIO,

MORPH, and MEDS II at a fixed operating point. Howard et al. [9] evaluated the Gini Aggregation Rate (GARBE) and compared it with FDR and the NIST Inequality Rate (IR) using FRVT scores at $FMR = 10^{-5}$. Kotwal and Marcel [10] proposed three score-distribution fairness indexes, namely SFI, CFI, and DFI, and analyzed their behavior on controlled synthetic distributions. Elobaid et al. [11] introduced the Sum of Error Differences by Group (SEDG) and compared it with ratio-based metrics under synthetic error-rate configurations. Sanon et al. [12] proposed the Theil Metric for Differential Performance (TM-DP) and evaluated it on LFW, AgeDB, and DemogPairs using noise injected into feature representations. More recently, ISO/IEC 19795-10:2024 [13] formalized four norm-based aggregated error measures for biometric fairness assessment, but did not provide an empirical comparison between these variants or against previously proposed metrics. Consequently, although the literature now offers a broad set of fairness metrics, it still lacks systematic evidence for selecting an appropriate metric under a given deployment condition. This limitation is particularly important because metric selection directly affects the interpretation of bias and the evaluation of mitigation strategies. Existing bias mitigation approaches, such as group-adaptive classifiers [14], attention-based debiasing [15], and sensitive-attribute-aware loss functions [16], implicitly assume that the bias targeted for correction has been measured in a reliable and meaningful way. If the metric used to quantify bias is unstable, insensitive to the relevant source of disparity, or inconsistent across operating conditions, then the conclusions drawn about mitigation effectiveness may be incomplete or misleading. Therefore, a fragmented metric landscape not only complicates fairness evaluation, but also limits the reliability of downstream debiasing claims.

Three main gaps explain why principled metric selection remains an open problem. First, existing comparisons are typically conducted within the evaluation framework introduced by the authors of a newly proposed metric, which may emphasize the specific behavior that the metric was designed to capture. Second, several studies rely on synthetic score distributions, which are useful for controlled analysis but do not fully reproduce calibration effects, distribution tails, subgroup heterogeneity, and acquisition-quality interactions observed in real biometric systems. Third, most evaluations focus on a single source of bias, commonly demographic imbalance or error-rate disparity, leaving unanswered whether a metric that is reliable under one bias mechanism remains valid under others.

Our work addresses these limitations through a unified empirical framework that evaluates nineteen fairness metrics on four real face verification datasets under three distinct sources of bias. Rather than relying on qualitative interpretation alone, we introduce controlled bias induction procedures that provide an explicit experimental ground truth. This enables a systematic assessment of each metric according to two formal validity criteria: monotonicity, which measures whether the metric responds consistently to increasing bias intensity, and stability, which measures whether the metric remains robust under repeated controlled conditions. By applying the same protocol across metrics, datasets, and bias mechanisms, our study provides a comparative basis for understanding when different fairness metrics are appropriate and where their limitations arise.

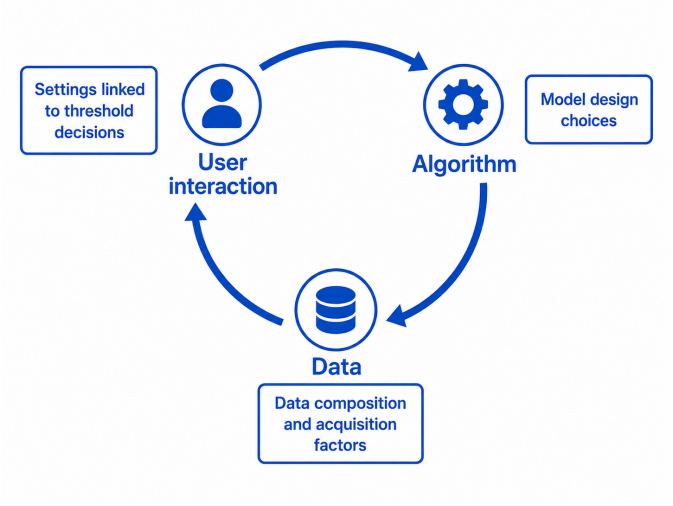

Fig. 2: Types of bias in biometrics (Adapted from [6]).

III. Methodology

This section presents a framework for selecting fairness metrics in biometric verification by measuring their response to controlled, progressively induced bias in real biometric data. We consider three practical sources of bias: algorithmic choices, data-related variations, and operational decision thresholds. By varying one factor at a time while keeping the remaining protocol fixed, the known bias ordering provides a reference for validating each metric. We first describe the evaluation pipeline and validity criteria (Section III-A), followed by the bias-induction protocols (Section III-B) and the nineteen evaluated metrics (Section III-C).

A. Pipeline

The framework varies one bias-inducing factor at controlled severity levels and measures whether each fairness metric responds consistently. Rather than comparing metrics in the abstract, it identifies the bias conditions under which they provide reliable conclusions. The three considered conditions—algorithmic, data-related, and operational bias—are illustrated in Figure 2. After bias induction, all experiments follow the same biometric evaluation chain, shown in Figure 1.

These conditions correspond to distinct mechanisms identified in existing bias taxonomies [2], [6], [3]. Algorithmic bias arises from the recognition model, here through the training loss, which shapes the embedding space and may produce unequal demographic error distributions. Data-related bias results from dataset composition or acquisition quality, including demographic imbalance, illumination, blur, and compression artifacts [17], [18], [19]. Operational bias arises from deployment choices, particularly the decision threshold, which controls the FMR–FNMR trade-off and may affect groups differently [8], [12].

For each controlled data-related and operational condition, we define ordered severity levels while keeping all other factors fixed. Unlike evaluations based solely on uncontrolled real-world disparities or synthetic data, this protocol introduces controlled and progressively increasing variations into real biometric data. The known ordering of these variations provides the reference for metric validation.

Metric validity is assessed using monotonicity and stability. Monotonicity measures whether the metric changes consistently with increasing bias severity:

$$\rho_m^{(B)} = \mathrm{Corr}_S\big(\ell_k, m(\ell_k)\big), \quad k = 1, \ldots, n, \tag{1}$$

where Corr_S is Spearman's rank correlation [20], $\ell_1 < \cdots < \ell_n$ are the ordered severity levels for bias condition B, and $m(\ell_k)$ is the value of metric m at level ℓ_k. Stability measures whether the resulting conclusions remain consistent across demographic and intersectional groups. A metric is considered valid for a controlled bias condition only when both criteria are satisfied.

B. Bias induction

The framework is evaluated on four public face datasets: DemogPairs [23], Racial Faces in the Wild (RFW) [24], MORPH II [25], and CelebA-HQ [26]. Together, they cover gender, ethnicity, age, and the soft-biometric attributes bangs, heavy makeup, smiling, big nose, and big lips. DemogPairs is balanced by design. For RFW, MORPH II, and CelebA-HQ, we enforce a uniform minimum number of identities and images per identity within each subgroup. The resulting configurations are reported in Table II.

Algorithmic bias is investigated by varying the training objective while fixing the architecture and evaluation protocol. The same Inception-ResNet-SE backbone [27] is trained with Softmax [28], SphereFace [29], CosFace [30], and ArcFace [31]. The four models are trained on MS-Celeb-1M [32] , and their classification accuracies, computed as the proportion of correctly classified identity samples, are 85.24%, 85.02%, 91.75%, and 84.48%, respectively. These values document differences in discriminative performance and are not used as fairness measurements. Score-based metrics are computed directly from each model's score distributions, whereas error-rate-based metrics are evaluated at each model's own Equal Error Rate threshold to reduce the influence of calibration differences on the comparison.

Data-related bias is induced through image degradation and subgroup imbalance. Image quality is modified according to $\tilde{I} = \mathcal{T}(I; t, \ell)$, where t is the degradation type and ℓ its severity. We apply defocus blur, motion blur, sensor noise, JPEG compression, and exposure or illumination variation at four increasing levels. Their effects are verified using OFIQ quality scores [33], which are expected to decrease with severity.

Subgroup imbalance is introduced using four retention levels, $\ell \in \{1.0, 0.75, 0.50, 0.25\}$. For each level $\ell < 1$, resampling is repeated $R = 10$ times using disjoint folds, providing independent repetitions while covering the complete subgroup. Only the target subgroup is resampled; all other groups remain unchanged.

Operational bias is induced by progressively tightening the decision threshold under two complementary regimes. The user-oriented regime uses FNMR targets of 8%, 7%, 6%, and 5%, whereas the security-oriented regime uses FMR targets of 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Metric trajectories across these ordered operating points are used to assess their monotonicity and stability under threshold tightening. The final operational ranking is obtained by averaging the ranks from both regimes.

C. Fairness metrics

We evaluate nineteen fairness metrics under the same controlled protocol. They are divided into score-distribution-based metrics, computed directly from genuine and impostor scores, and error-rate-based metrics, computed after applying a decision threshold.

Let s be the comparison score for a probe–reference pair, with lower values indicating greater similarity. Comparing s with a threshold τ yields the False Match Rate (FMR) and False Non-Match Rate (FNMR). For demographic groups $\mathcal{G} = \{g_1, \ldots, g_K\}$, $\mu_G(g_i)$ and $\sigma_G(g_i)$ denote the mean and standard deviation of genuine scores, while $\mu_I(g_i)$ and $\sigma_I(g_i)$ denote the corresponding impostor statistics. The score distribution of group g_i is denoted by P_i.

The threshold-independent metrics proposed by Kotwal and Marcel [10] comprise the Separation Fairness Index (SFI), Compactness Fairness Index (CFI), and Distribution Fairness Index (DFI). They respectively measure group disparities in genuine–impostor separation, distribution compactness, and full score distributions through KL divergence. Each is evaluated in normal (N), extremal (E), and weighted (W) variants, representing average, worst-case, and imbalance-aware disparities. Their values lie in $[0, 1]$, with larger values indicating greater fairness.

Error-rate-based metrics include the Fairness Discrepancy Rate (FDR) [8], based on inter-group FMR and FNMR differences; the Inequality Rate (IR) [21], based on the ratio between extreme group-specific FMR values; and GARBE [9], which combines the Gini coefficients of the FMR and FNMR vectors. The two SEDG variants [11] measure the mean and standard deviation of group deviations from global error behavior at the EER threshold. TM-DP [12] applies the Theil inequality index [34] to decompose performance disparities into between-group and within-group components. Finally, four indices derived from ISO/IEC 19795-10 [13] aggregate worst-case FMR and FNMR disparities using linear or geometric combinations. We denote them ISO1, ISO1FACTO, ISO2, and ISO2FACTO. For FDR, GARBE, and the ISO-derived indices, $\alpha = 0.5$, giving equal weight to FMR and FNMR. Table I reports the fairness metrics evaluated in this study, together with their formulations, evaluation protocols, and bias modeling strategies.

TABLE I: Overview of fairness metrics.

Metric	Type	Formula / Definition	Reference	Datasets / Systems	Operating Points	Bias Evaluation Strategy
SFI (N/E/W)	Score-based	$\mathrm{SFI}_N = 1 - \frac{2}{K}\sum_i \lvert z_{S_i} - \bar{z}_S \rvert$, $z_{S_i} = \lvert \mu_G(d_i) - \mu_I(d_i) \rvert$	Kotwal & Marcel, TBIOM 2023 [10]	Synthetic score models; not system-specific	Threshold-independent	Inter-group separation *($[0,1]$; 1 = parity)*
CFI (N/E/W)	Score-based	$\mathrm{CFI}_N = 1 - \frac{2}{K}\sum_i \lvert z_{C_i} - \bar{z}_C \rvert$, $z_{C_i} = \sigma_G(d_i) + \sigma_I(d_i)$	Kotwal & Marcel, TBIOM 2023 [10]	Synthetic score models; not system-specific	Threshold-independent	Inter-group compactness *($[0,1]$; 1 = parity)*
DFI (N/E/W)	Score-based	$\mathrm{DFI}_N = 1 - \frac{1}{K\log_2 K}\sum_i D_{\mathrm{KL}}\left(P_i \,\|\, \bar{P}\right)$, $\bar{P} = \frac{1}{K}\sum_i P_i$	Kotwal & Marcel, TBIOM 2023 [10]	Synthetic score models; not system-specific	Threshold-independent	Distributional divergence *($[0,1]$; 1 = parity)*
FDR	Error-rate	$\mathrm{FDR} = 1 - \left(\alpha\Delta_{\mathrm{FMR}} + (1-\alpha)\Delta_{\mathrm{FNMR}}\right)$, $\Delta_X = \max_{i,j}\left\lvert X^{d_i} - X^{d_j}\right\rvert$	De Freitas Pereira & Marcel, TBIOM 2022 [8]	MOBIO; MORPH; MEDS II. Systems: Inception-ResNet-v2; ArcFace; COTS	$\mathrm{FMR} \in [10^{-1}, 10^{-6}]$	Worst-case inter-group disparity *($[0,1]$; 1 = parity)*
IR	Error-rate	$\mathrm{IR} = \frac{\max_d \mathrm{FMR}^d}{\min_d \mathrm{FMR}^d}$	Grother et al., NIST [21]	FRVT; commercial and academic face-recognition systems	Deployment-specific	Maximum-to-minimum error ratio *($[1,\infty)$; 1 = parity)*
GARBE	Error-rate	$\mathrm{GARBE} = \alpha\,\mathrm{Gini}(\mathrm{FMR}) + (1-\alpha)\,\mathrm{Gini}(\mathrm{FNMR})$	Howard et al., TBIOM 2023 [9]	FRVT; 126 FRVT algorithms	$\mathrm{FMR} = 10^{-5}$	Gini-based dispersion across groups *($[0,1]$; 0 = parity)*
SEDG mean/std	Error-rate	$s^d = \frac{\lvert e^d - e^g \rvert}{e^g}$, $\mathrm{SEDG}_{\mathrm{mean}} = \frac{1}{K}\sum_d s^d$, $\mathrm{SEDG}_{\mathrm{std}} = \mathrm{std}(s^d)$	Elobaid et al., FG 2024 [11]	Synthetic error scenarios; not system-specific	Global EER	Deviation from the global error *($[0,\infty)$; lower values indicate greater parity)*
TM-DP	Error-rate	$\text{TM-DP} = \frac{1}{K}\sum_i \frac{e^{d_i}}{\bar{e}}\ln\left(\frac{e^{d_i}}{\bar{e}}\right)$, $\bar{e} = \frac{1}{K}\sum_i e^{d_i}$	Sanon et al., ICCST 2025 [12]	LFW; AgeDB; DemogPairs. Systems: Inception-ResNet-V1; ArcFace	$\mathrm{FMR} = 10^{-2}, 10^{-3}$	Inter- and intra-group inequality; decomposable into between- and within-group components [22] *($[0,\infty)$; 0 = parity)*
ISO1	Error-rate	$\alpha U_{\mathrm{FMR}} + (1-\alpha)U_{\mathrm{FNMR}}$, $U_X = \max_d X^d - \min_d X^d$	ISO/IEC 19795-10 [13]	Standard specification; N/A	Application-dependent	Linear AEM based on absolute difference *($[0,1]$; 0 = parity)*
ISO1FACTO	Error-rate	$\alpha U_{\mathrm{FMR}} + (1-\alpha)U_{\mathrm{FNMR}}$, $U_X = \frac{\max_d X^d}{\min_d X^d}$	ISO/IEC 19795-10 [13]	Standard specification; N/A	Application-dependent	Linear AEM based on ratio *($[1,\infty)$; 1 = parity)*
ISO2	Error-rate	$U_{\mathrm{FMR}}^{\alpha} U_{\mathrm{FNMR}}^{1-\alpha}$, $U_X = \max_d X^d - \min_d X^d$	ISO/IEC 19795-10 [13]	Standard specification; N/A	Application-dependent	Geometric AEM based on absolute difference *($[0,1]$; lower values indicate greater parity)*
ISO2FACTO	Error-rate	$U_{\mathrm{FMR}}^{\alpha} U_{\mathrm{FNMR}}^{1-\alpha}$, $U_X = \frac{\max_d X^d}{\min_d X^d}$	ISO/IEC 19795-10 [13]	Standard specification; N/A	Application-dependent	Geometric AEM based on ratio *($[1,\infty)$; 1 = parity)*

Datasets	Attributes	Statistics
DEMOGPAIRS GENDER	Men	18×300
	Women	18×300
DEMOGPAIRS ETHNICITY	Asian	18×200
	Black	18×200
	White	18×200
RFW	African	6×56
	Asian	6×56
	Caucasian	6×56
	Indian	6×56
MORPH GENDER	Men	6×92
	Women	6×92
MORPH AGE	18–30	6×92
	31–50	6×92
	51+	6×92
CELEBA-HQ SOFT ATTRIBUTES	Bangs	6×150
	Heavy makeup	6×150
	Smiling	6×150
	Big nose	6×150
	Big lips	6×150

TABLE II: Statistics of the datasets used for experimentation.

IV. Experiments and Results

The four experiments below confront the nineteen fairness metrics with each of the bias sources defined in Section III.

A. Algorithmic bias

This experiment documents how sensitive fairness conclusions are to the training objective before any controlled bias is induced. Figure 4 presents, for each of the twelve demographic and soft-attribute groups, the value of the nineteen metrics under each of the four losses introduced in Sec. III-B.

Each subplot corresponds to one evaluation database. Each radar chart displays nineteen axes, one per fairness metric, and four polygons, one per loss. The distance from the centre reflects the level of bias: the closer a polygon is to the centre, the fairer the model. Metrics whose ideal value is one (SFI, CFI) are inverted ($1 -$ value) so that the centre consistently represents perfect fairness across all axes. Values are normalised per metric by the global maximum observed across all databases, enabling direct visual comparison between subplots.

At the group level, the picture is fragmented. Softmax yields the lowest mean bias on *DemogPairs Ethnicity* and *RFW Ethnicity*, ArcFace on several gender-based groups (*DemogPairs Gender*, *DemogPairs Asian Gender*, *Morph Gender*, *Morph 51+ Gender*), and CosFace on *Morph Age*, *Morph 18-30 Gender* and *CelebaSoft*. On *DemogPairs White Gender*, the four losses are nearly indistinguishable. The loss that ranks first on one group is often among the worst on another: Softmax leads on *DemogPairs Ethnicity* but ranks last on five other groups.

A clearer structural pattern emerges from the aggregated ranking of Figure 3. The nineteen metrics do not rank the four losses arbitrarily: each metric family prefers the loss whose embedding geometry produces the kind of inter-group regularity that family is built to detect. Score-based metrics (SFI and CFI, all variants) rank Softmax first, since the absence of an angular margin yields more uniform score-distribution shapes across demographic groups, which is precisely what SFI and CFI capture. Distribution-divergence metrics (DFI, all variants) and bounded absolute-error metrics (FDR, ISO1) rank

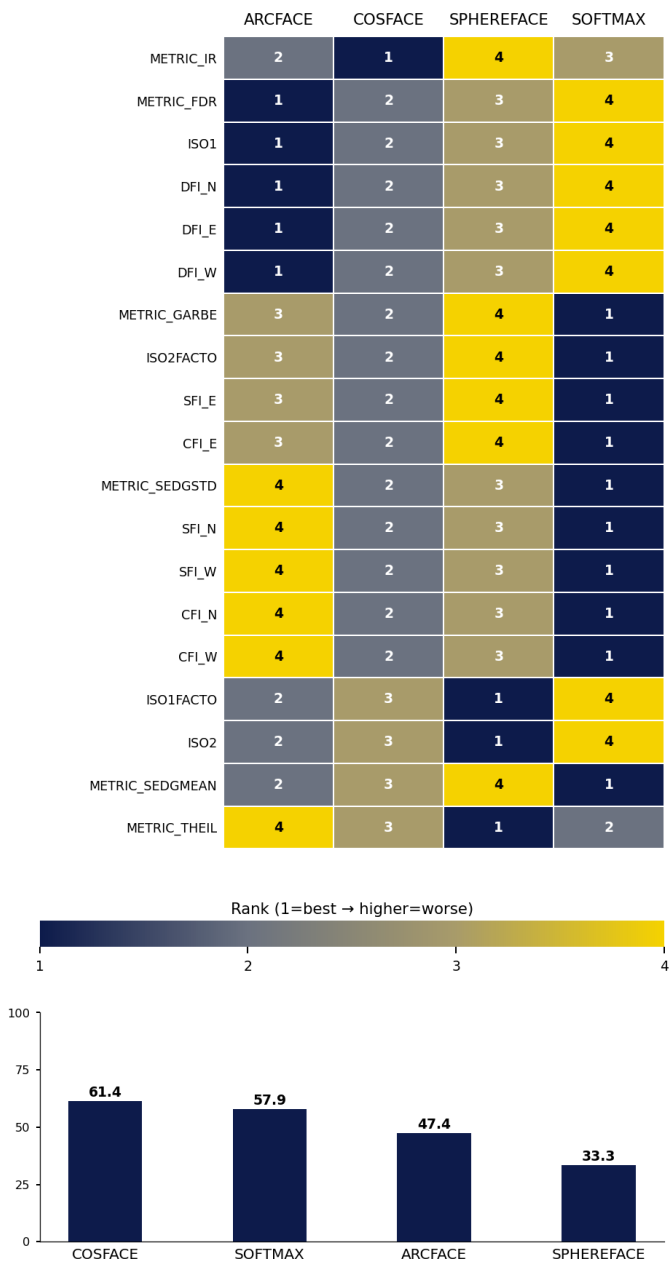

Fig. 3: Aggregated ranking of loss functions across the nineteen fairness metrics (majority vote).

ArcFace first, whose strong angular margin produces compact and globally comparable distributions with small absolute error gaps across groups. Ratio- and geometric-aggregation metrics (ISO1FACTO, ISO2, TM-DP) rank SphereFace first, where the group-wise normalised disparities are most balanced.

Thus, the compatibility between metric structure and loss architecture is itself structurally predictable. The next experiments that follow extend this matching principle.

B. Data-related bias: dataset resampling

Following the global comparison of extractors presented in the previous section, CosFace is fixed for the subsequent analysis, as it achieves both the highest classification accuracy

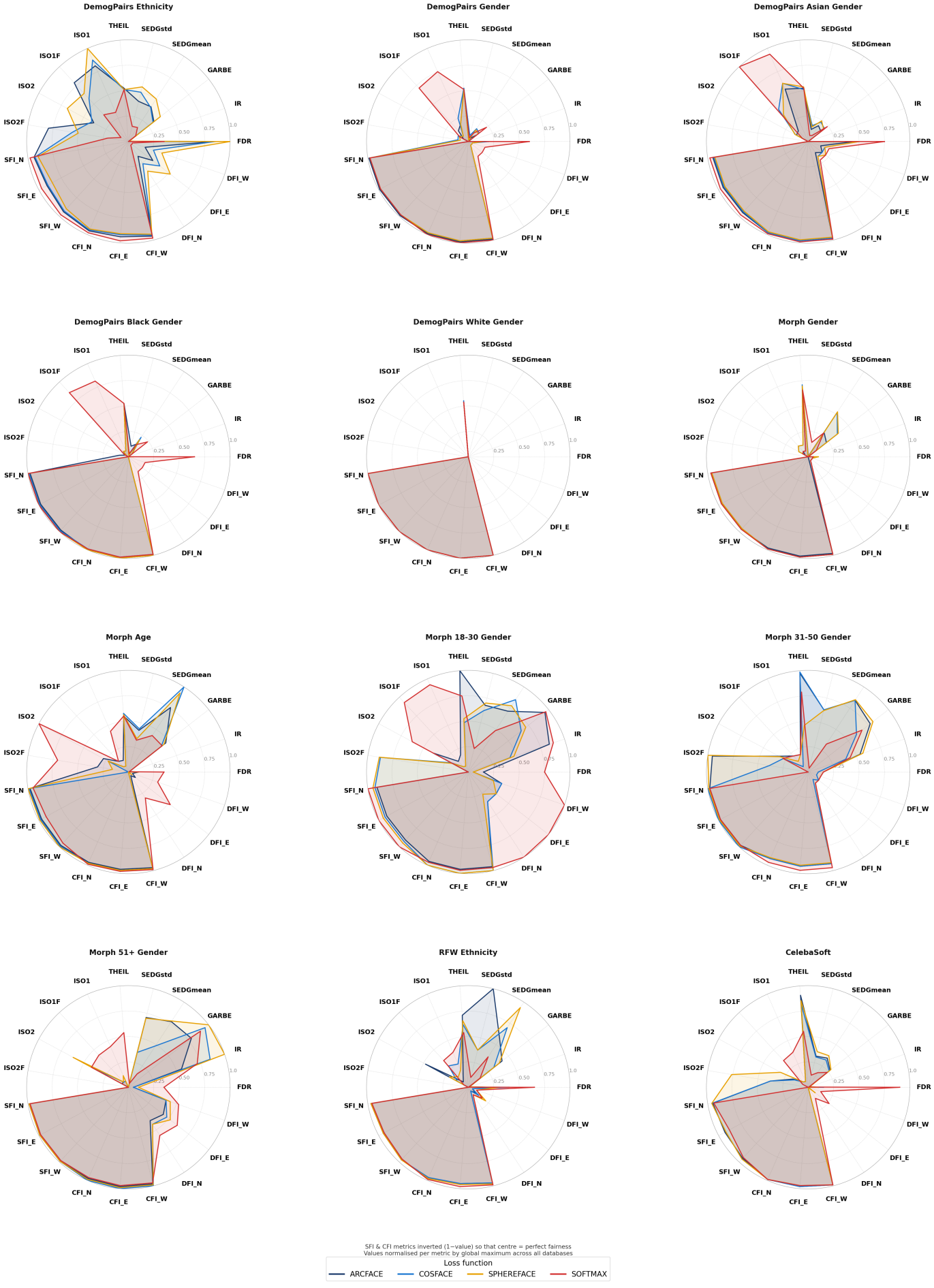

Fig. 4: Comparison of fairness metrics across the four loss functions for each demographic group. Smaller polygon area indicates greater fairness.

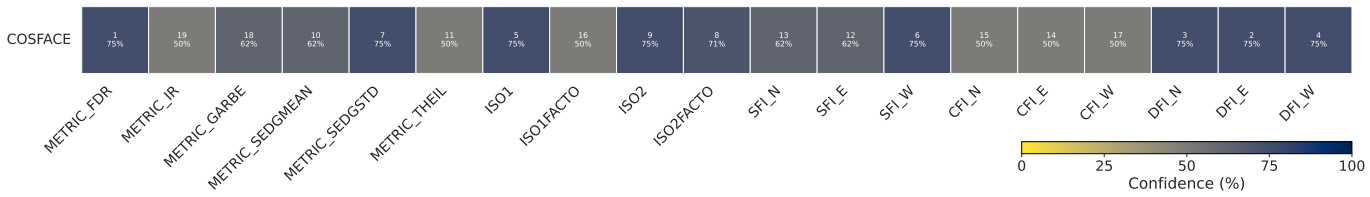

Fig. 5: Ranking of metrics under controlled resampling. Each cell reports the numerical rank (1 = best) and the associated win rate; cell color encodes the win rate, with darker tones indicating higher confidence.

and the most favorable fairness ranking across demographic groups. This choice allows us to isolate data-related bias while minimizing the influence of algorithmic variability.

To evaluate the sensitivity of fairness metrics to changes in the composition of the evaluation data, we progressively reduce the proportion of samples retained within each subgroup. The variation relative to the complete dataset is defined as

$$\Delta_m(p) = V_m(p) - V_m(1.0), \tag{2}$$

where p denotes the retained proportion and $V_m(p)$ is the value of metric m at that proportion. Metric orientations are harmonized so that larger values correspond to greater measured bias. The metrics are then ranked according to the magnitude, monotonicity, and stability of their responses across sub-datasets and resampling levels. The final aggregated ranking is shown in Figure 5, where each cell reports the metric rank (1 = best), while color intensity represents the associated win rate.

The ranking reveals a direct relationship between the mathematical structure of a metric and its suitability for sampling-induced bias. FDR obtains the highest rank, followed by DFI_E, DFI_N, and DFI_W, while ISO1 ranks fifth. Thus, the strongest metrics belong to two complementary structural families: bounded absolute error-disparity metrics and full-distribution divergence metrics.

FDR and ISO1 are well suited to subgroup resampling because they measure worst-case inter-group differences using bounded absolute deviations. Their formulations do not divide by a minimum group-specific error rate and are therefore less affected by fluctuations caused by small subgroup sizes. They capture changes in FMR and FNMR while remaining relatively robust to denominator instability.

The high ranks of all three DFI variants show that resampling also acts as a distributional perturbation. Reducing the number of samples available for one subgroup modifies the empirical estimation of its score distribution, particularly its density and tails. Because DFI compares complete score distributions through KL divergence, it is structurally adapted to detecting these changes. The strong performance of its normal, extremal, and weighted variants indicates that this sensitivity is not limited to a single aggregation strategy.

By contrast, IR and ISO1FACTO rank near the bottom because their maximum-to-minimum ratio formulations depend on potentially small and noisy denominators. Minor variations in the minimum group error rate can therefore produce disproportionately large metric changes. The CFI variants also rank lower because subgroup resampling does not necessarily produce a systematic change in score compactness. Similarly, SFI reacts mainly when resampling alters genuine–impostor separation, which explains its more variable performance across sub-datasets.

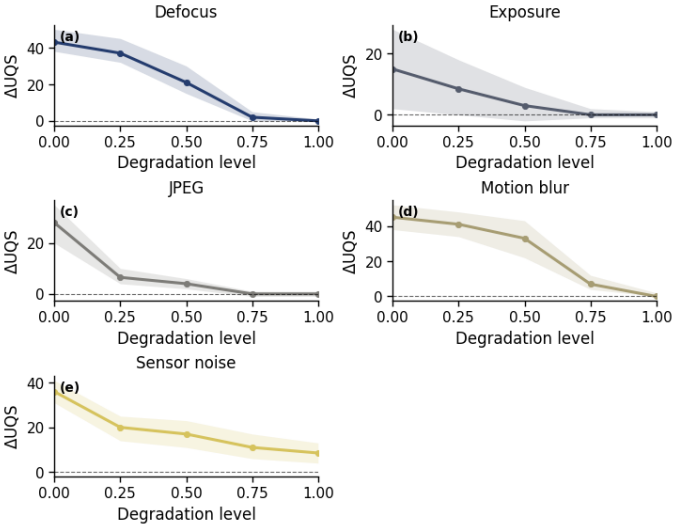

Fig. 6: OFIQ quality score evolution with degradation severity, confirming the monotonic decrease of perceived quality.

Overall, the results identify an alignment between metric architecture and the manifestation of bias. Bounded absolute error metrics such as FDR and ISO1 are adapted to stable group-wise error shifts, whereas DFI variants are adapted to resampling-induced changes in empirical score distributions. Ratio-based metrics such as IR and ISO1FACTO are less appropriate when subgroup sizes become small. For dataset resampling, FDR and DFI therefore provide complementary views of decision-level and score-distribution-level bias.

C. Data-related bias: image quality degradation

To assess metric responses to image quality degradation, we consider five realistic perturbations: defocus blur, motion blur, sensor noise, low exposure, and JPEG compression. Each degradation is applied at four increasing severity levels.

Prior to metric comparison, all degradations were validated using OFIQ quality scores, as shown in Figure 6. The monotonic decrease in these scores confirms that the perturbations introduce controlled and progressively severe quality deterioration. The corresponding metric rankings are reported in Figure 8. The results show that the most suitable

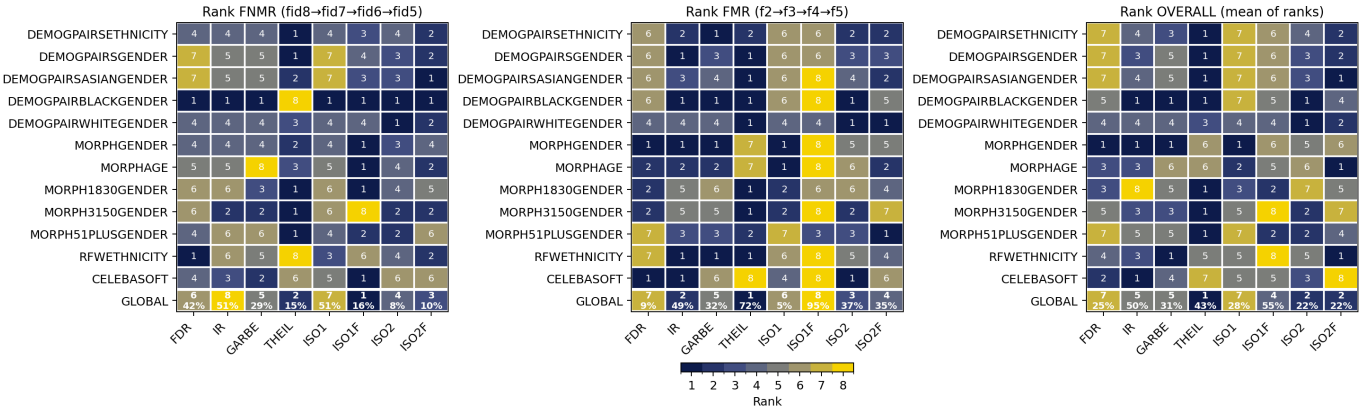

Fig. 7: Ranking of metrics under progressive threshold constraints. Left: user-oriented FNMR regime; middle: security-oriented FMR regime; right: combined ranking.

metric depends on how each degradation affects the score distributions and the resulting error rates.

a) Defocus blur: ISO1 ranks first, followed by $\text{SEDG}_{\text{mean}}$ and SFI_E. Defocus reduces image sharpness and progressively modifies genuine–impostor separation, leading to group-dependent error shifts. ISO1 captures these shifts through absolute FMR and FNMR differences, while $\text{SEDG}_{\text{mean}}$ measures deviations from the global error behavior. The high rank of SFI_E further indicates that defocus also affects the worst-case separation between genuine and impostor scores.

b) Motion blur: A clearer score-level pattern appears for motion blur, for which SFI_E, SFI_W, and SFI_N occupy the first three ranks. Motion blur increases genuine-score dispersion and overlap with impostor scores, directly altering the group-specific separation $|\mu_G - \mu_I|$. Because this quantity is central to the SFI formulation, the SFI family is naturally better aligned with this type of degradation than threshold-dependent metrics.

c) Sensor noise: Under sensor noise, ISO1, FDR, and DFI_N obtain the highest ranks. Noise introduces heterogeneous variance and reshapes the score distributions, which explains the sensitivity of DFI_N. These distributional changes also translate into group-specific FMR and FNMR differences, captured by the absolute-disparity formulations of ISO1 and FDR. The leading metrics therefore reflect complementary score-level and decision-level effects.

d) Low exposure: Low exposure favors $\text{SEDG}_{\text{mean}}$, DFI_E, and DFI_N. Reduced contrast and dynamic range produce gradual deviations from global error behavior while also reshaping the complete score distributions. $\text{SEDG}_{\text{mean}}$ is therefore responsive to the resulting error shifts, whereas the DFI variants capture their broader distributional effects.

e) JPEG compression: For JPEG compression, FDR ranks first, followed by CFI_N and CFI_W. Compression artifacts mainly increase score dispersion and modify distribution compactness, which corresponds directly to the CFI formulation. FDR provides a complementary decision-level view by capturing the largest group-wise differences in FMR and FNMR.

f) Cross-degradation trends: The mean ranking places ISO1 first, followed by DFI_E, FDR, SFI_E, DFI_N, and DFI_W. Thus, the results reveal a structural correspondence: absolute error metrics are better suited to group-wise error shifts, DFI to full-distribution changes, SFI to separation changes, and CFI to compactness variations. Metric suitability therefore depends on whether its mathematical architecture matches the way quality degradation manifests as bias.

D. Operational bias: decision threshold variation

The objective of this analysis is to determine whether fairness metrics respond consistently when progressively stricter decision constraints are applied. Figure 7 presents the ranking of threshold-dependent metrics under two operational regimes.

The left panel corresponds to the FNMR-driven user-oriented regime, the middle panel to the FMR-driven security-oriented regime, and the right panel to the combined ranking. Lower ranks indicate a more consistent response to threshold tightening, while the percentages report confidence in the observed monotonic behavior.

a) FNMR-driven regime:
ISO1FACTO ranks first, followed by THEIL and ISO2FACTO. Their strong response is consistent with their relative-disparity formulations: ISO1FACTO and ISO2FACTO rely on maximum-to-minimum error ratios, while THEIL measures errors relative to their group mean. These structures are sensitive to changes in relative group disparities when the FNMR constraint is tightened. However, the poor performance of ISO1FACTO in the FMR regime indicates that its sensitivity is specific to the operating condition rather than uniformly robust.

b) FMR-driven regime:
THEIL ranks first, followed by IR and ISO2. The strong performance of IR is structurally expected because it is defined directly from the ratio between the maximum and minimum group-specific FMR values. However, its weaker FNMR ranking limits its use across both regimes. THEIL remains highly ranked because its normalized inequality formulation captures changes in the relative distribution of group errors, whereas

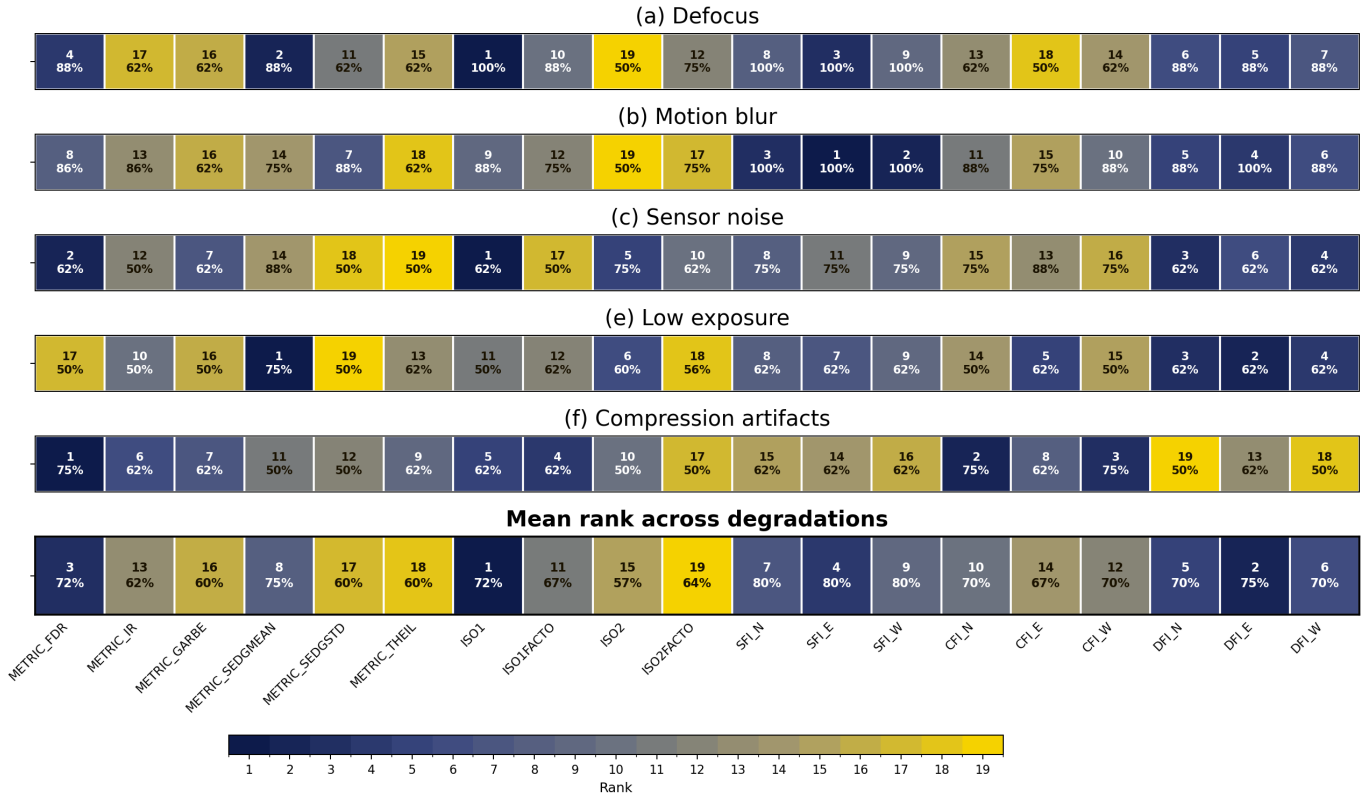

Fig. 8: Ranking heatmaps for the five degradation types. Darker cells indicate better satisfaction of the validity criteria.

ISO2 captures the joint evolution of absolute FMR and FNMR disparities through geometric aggregation.

c) Overall ranking:

When both regimes are combined, THEIL ranks first, while ISO2 and ISO2FACTO share the second position. This indicates that threshold-induced bias is best captured by metrics that measure relative inequality or jointly aggregate changes in FMR and FNMR. By contrast, FDR and ISO1 rank lower overall: their bounded absolute differences are stable, but may respond less monotonically when threshold tightening changes group errors proportionally. The group-specific results nevertheless remain heterogeneous, showing that no metric is uniformly optimal across all demographic configurations.

E. Cross-experiment summary

Across the four experiments, no single metric satisfies both validity criteria in every situation, and the metrics that pass them shift systematically with the source of bias.

This variation follows the mathematical structure of the metrics. Score-based metrics operate before thresholding: SFI detects separation changes, CFI captures variations in distribution compactness, and DFI responds to complete distributional reshaping. They are therefore most suitable when bias affects score-space geometry, as in algorithmic variations and most image-quality degradations.

Error-rate-based metrics operate after thresholding and differ mainly in their aggregation form. Bounded absolute metrics such as FDR and ISO1 are robust to sampling variability; Gini or geometric aggregations such as GARBE and ISO2 are more stable in multi-group settings; and inequality-based metrics such as THEIL respond strongly to operating-point variation. In contrast, ratio-based metrics such as IR and ISO1FACTO are more sensitive to small or noisy denominators.

These findings lead to a practical selection rule: use score-based metrics for distributional distortions, bounded absolute error metrics for sampling variability, Gini or geometric aggregation for multi-group settings, and inequality-based metrics for operational threshold changes.

The complete mapping is consolidated in Table III and discussed in Section V.

V. Discussion

A. The bias source dominates the metric

Across the controlled resampling, image-quality degradation, and decision-threshold experiments, the results show that metric validity depends strongly on the source and manifestation of bias. The exploratory training-loss comparison reveals a similar dependence on loss-induced embedding geometry, although it does not provide an ordered severity of algorithmic bias. A metric may therefore satisfy the validity criteria under one controlled condition and fail under another because its mathematical structure is aligned with a specific form of disparity. Bounded absolute-disparity metrics such as FDR and ISO1 are robust to sampling variability, whereas ratio-based metrics such as IR and ISO1FACTO are sensitive to small or unstable denominators. DFI captures full score-distribution reshaping, SFI responds to changes in genuine–impostor separation, and inequality-based or geometrically

TABLE III: Structure-aware guideline for selecting fairness metrics according to the expected manifestation of bias. For data-related and operational conditions, recommendations reflect the metrics that most consistently satisfied monotonicity and stability. Algorithmic entries summarize structural indications from the training-loss architecture.

Bias source	**Bias manifestation**	**Recommended metrics**	**Structural rationale**
Algorithmic bias	Change in genuine–impostor separation	SFI_N, SFI_E, SFI_W	SFI directly measures inter-group variation in $\lvert\mu_G - \mu_I\rvert$ and is therefore suited to training objectives that modify embedding margins.
	Change in score compactness	CFI_N, CFI_E, CFI_W	CFI measures differences in genuine and impostor score dispersion, making it appropriate when the training objective modifies cluster compactness.
	Full score-distribution reshaping	DFI_N, DFI_E, DFI_W	DFI compares complete group-wise score distributions and captures changes in their shape, overlap, and tails.
	Decision-level error redistribution	FDR, ISO1	Their bounded absolute FMR and FNMR differences capture the decision-level consequences of model changes without unstable ratio normalization.
Data-related bias	Subgroup resampling or sampling variability	FDR, DFI_E/DFI_N/DFI_W, ISO1	FDR and ISO1 capture stable absolute error gaps, while DFI detects resampling-induced changes in empirical score distributions.
	Defocus blur	ISO1, $\text{SEDG}_{\text{mean}}$, SFI_E	Defocus produces progressive group-wise error shifts and reduces genuine–impostor separation.
	Motion blur	SFI_E, SFI_W, SFI_N	Motion blur directly alters genuine–impostor separation, which is the quantity measured by the SFI family.
	Sensor noise	ISO1, FDR, DFI_N	Noise reshapes score distributions and produces heterogeneous group-wise error differences, requiring complementary distribution- and error-based measures.
	Low exposure	$\text{SEDG}_{\text{mean}}$, DFI_E, DFI_N	Low exposure causes progressive deviations from global error behavior and broader changes in score-distribution shape.
	JPEG compression	FDR, CFI_N, CFI_W	Compression mainly modifies score dispersion and compactness, while FDR captures the resulting worst-case error differences.
Operational bias	FNMR-driven threshold tightening	ISO1FACTO, THEIL, ISO2FACTO	Ratio-based and normalized inequality formulations are sensitive to changes in relative group disparities under stricter FNMR constraints.
	FMR-driven threshold tightening	THEIL, IR, ISO2	IR directly measures relative FMR disparities, while THEIL and ISO2 capture the redistribution and joint evolution of group error rates.
	Combined FNMR/FMR assessment	THEIL, ISO2, ISO2FACTO	These metrics provide the most consistent response across both operating regimes. For large multi-group settings, GARBE or FDR may be used as complementary measures.

aggregated metrics such as THEIL and ISO2 respond consistently to threshold tightening.

Fairness should therefore not be regarded as an intrinsic system property measurable by a single universally optimal metric. Instead, it is a context-dependent property whose assessment depends on how bias manifests in score distributions and error rates. The same biometric system may consequently receive different fairness assessments under different metrics without these assessments being contradictory, provided that each metric captures a distinct aspect of the observed disparity. Metric selection should therefore be guided by the expected bias mechanism and the operational conditions of the deployment.

B. From experimental evidence to operational guidelines

Table III translates these findings into practical recommendations. For the controlled data-related and operational conditions, it reports the metrics that most consistently satisfied monotonicity and stability, together with the structural rationale for their selection. The algorithmic entries summarize the tendencies observed in the exploratory training-loss comparison and should therefore be interpreted as structural indications rather than formally validated recommendations.

C. Practical implications

The proposed framework provides three practical benefits for fairness auditing. First, it reduces the selection problem from nineteen candidate metrics to a smaller set of context-specific recommendations. Auditors can select suitable metrics according to the expected manifestation of bias, such as subgroup sampling variability, acquisition-quality degradation, or restrictive operating thresholds.

Second, the framework can also support black-box evaluations in which only group-wise FMR and FNMR values are available, as in public evaluations of commercial systems. In such settings, practitioners can use the recommended error-rate-based metrics reported in Table III, including FDR, ISO1, GARBE, THEIL, and ISO2, without requiring access to score distributions or internal model parameters. When score distributions are available, score-based metrics can provide complementary information about separation, compactness, and distributional changes.

Third, the structural analysis supports the use of complementary metric combinations. Pairing a bounded absolute-disparity metric such as FDR or ISO1 with DFI captures both decision-level disparities and score-distribution distortions. Similarly, combining an inequality-based metric such as THEIL with an extremal difference metric provides complementary information about overall inequality and worst-

case group behavior. Such combinations should be selected according to the expected bias mechanism rather than applied systematically in every evaluation.

D. Implications for biometric fairness standards

The findings also have implications for biometric fairness standards. ISO/IEC 19795-10 defines four Aggregated Error Measure variants (ISO1, ISO1FACTO, ISO2, and ISO2FACTO) based on linear or geometric aggregation and difference- or ratio-based disparities. Although these variants are closely related, they respond differently to the same controlled conditions. ISO1 is comparatively robust to sampling variability, whereas ISO1FACTO is more sensitive to denominator instability. By contrast, ISO2 and ISO2FACTO respond more consistently to threshold tightening across the two operational regimes. Without explicit criteria for selecting among these variants, laboratories applying the same standard may therefore reach different fairness conclusions.

Future standards could reduce this ambiguity by linking metric variants to specific bias manifestations and operational contexts [35]. The recommendations provided in Table III offer an empirical and structural basis for developing such context-dependent guidance.

VI. Conclusion

This work introduced a unified framework for comparing nineteen fairness metrics on real biometric data under three complementary sources of variation: algorithmic, data-related, and operational. Controlled and ordered perturbations in the data-related and operational experiments provide an explicit reference for assessing metric validity through monotonicity and stability. The training-loss comparison provides a complementary exploratory analysis of how algorithmic choices and loss-induced embedding geometry influence fairness conclusions.

The main finding is that metric suitability depends on how bias manifests in score distributions and error rates. Bounded absolute-disparity metrics are robust to sampling variability, distribution-based metrics detect complete score-distribution reshaping, separation-based metrics respond to changes in genuine-impostor separation, and inequality-based or geometrically aggregated metrics capture threshold-dependent disparities. The exploratory algorithmic results further show that different metric families favor different loss-induced embedding properties. Fairness should therefore be treated as a context-dependent property rather than as an intrinsic system property measurable by a single metric. These findings are operationalized through Table III, which maps bias manifestations to suitable metric structures and provides guidance for fairness audits in both white-box and black-box settings. The framework does not require access to internal model parameters: depending on the evaluation setting, it can be applied using score distributions or group-wise error rates. It is therefore applicable to biometric systems evaluated through public benchmarks or independent auditing protocols.

Several directions extend this work. First, current biometric fairness standards, particularly ISO/IEC 19795-10, define multiple variants of aggregated error measures without normative criteria for selecting among them. Our results show that these variants respond differently to the same controlled conditions, suggesting that future standardization work should incorporate context-dependent selection criteria. Second, extending the framework to other biometric modalities, such as fingerprint recognition [36] and speaker verification [37], would make it possible to determine whether the observed structure–bias relationships generalize beyond face recognition. Finally, future work could investigate whether monotonicity and stability under controlled bias induction can be incorporated as objectives in fairness-aware training, supporting the development of biometric systems whose fairness behavior is both measurable and robust across operational conditions.

Acknowledgment

The present work was performed using computing resources of CRIANN (Normandy, France) under the allocation 2026023. In addition, the authors would like to thank the FIME EMEA company and the National Association for Research and Technology (ANRT) for the financial support.

References

[1] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," in *IEEE Winter Conf. on Applications of Computer Vision Workshops (WACVW)*, 2020.

[2] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[3] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Bias and fairness in face recognition: A survey," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 381–396, 2020.

[4] K. Kotwal and S. Marcel, "Review of demographic bias in face recognition," *arXiv preprint arXiv:2502.02309*, 2025.

[5] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?" *IEEE Trans. on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 101–111, 2021.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[8] T. De Freitas Pereira and S. Marcel, "Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, Jan. 2020.

[9] J. J. Howard, E. J. Laird, R. E. Rubin, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Evaluating Proposed Fairness Models for Face Recognition Algorithms," in *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, ser. Lecture Notes in Computer Science, J.-J. Rousseau and B. Kapralos, Eds. Cham: Springer Nature Switzerland, 2023, pp. 431–447.

[10] K. Kotwal and S. Marcel, "Fairness Index Measures to Evaluate Bias in Biometric Recognition," Jun. 2023.

[11] A. Elobaid, N. Ramoly, L. Younes, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris, "Sum of group error differences: A critical examination of bias evaluation in biometric verification and a dual-metric measure," 2024. [Online]. Available: https://arxiv.org/abs/2404.15385

[12] K. Sanon, J. Manno, C. Charrier, and C. Rosenberger, "A new fairness evaluation metric of biometric systems based on the theil inequality," in *Proceedings of the International Carnahan Conference on Security Technology (ICCST)*, Oct. 2025, pp. 1–6.

[13] "Information technology — biometric performance testing and reporting — part 10: Quantifying biometric system performance variation across demographic groups," Tech. Rep. ISO/IEC 19795-10:2024, 2024.

[14] S. Gong, X. Liu, and A. K. Jain, "Mitigating Face Recognition Bias via Group Adaptive Classifier," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 3413–3423.
[15] L. Huang, M. Wang, J. Liang, W. Deng, H. Shi, D. Wen, Y. Zhang, and J. Zhao, "Gradient Attention Balance Network: Mitigating Face Recognition Racial Bias via Gradient Attention," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 38–47.
[16] I. Serna, A. Morales, J. Fierrez, and N. Obradovich, "Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," *Artificial Intelligence*, vol. 305, p. 103682, Apr. 2022.
[17] P. J. Phillips, T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "An introduction to the Face Recognition Vendor Test (FRVT)," *IEEE Computer*, vol. 44, no. 1, pp. 59–66, 2011.
[18] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*, pp. 77–91, 2018.
[19] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition accuracy and demographic effects," National Institute of Standards and Technology (NIST), Tech. Rep., 2019.
[20] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.
[21] P. Grother, "Demographic differentials in face recognition algorithms," EAB Virtual Event Series - Demographic Fairness in Biometric Systems, 2021.
[22] A. F. Shorrocks, "The class of additively decomposable inequality measures," *Econometrica*, vol. 48, no. 3, pp. 613–625, 1980.
[23] I. Hupont and C. Fernández, "Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–7.
[24] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
[25] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *IEEE 7th International Conference on Automatic Face and Gesture Recognition (FGR)*, Southampton, UK, 2006, pp. 341–345.
[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.
[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," May 2019, arXiv:1709.01507 [cs]. [Online]. Available: http://arxiv.org/abs/1709.01507
[28] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive Margin Softmax for Face Verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, Jul. 2018, arXiv:1801.05599 [cs]. [Online]. Available: http://arxiv.org/abs/1801.05599
[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," Jan. 2018, arXiv:1704.08063 [cs]. [Online]. Available: http://arxiv.org/abs/1704.08063
[30] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," Apr. 2018, arXiv:1801.09414 [cs]. [Online]. Available: http://arxiv.org/abs/1801.09414
[31] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, arXiv:1801.07698 [cs]. [Online]. Available: http://arxiv.org/abs/1801.07698
[32] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
[33] B.-O. P. Contributors, "Ofiq: Open source face image quality (v1.0.3)," 2025, reference implementation for ISO/IEC 29794-5. [Online]. Available: https://github.com/BSI-OFIQ/OFIQ-Project
[34] H. Theil, *Economics and Information Theory*. North-Holland, 1967.
[35] European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (ai act)," Official Journal of the European Union, 2024.
[36] Y. Godbole, S. Grosz, and A. K. Jain, "On demographic bias in fingerprint recognition," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2022.
[37] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *International Conference on Computational Science and its Applications (ICCSA)*, 2020.

Kaïra Neily SANON is a Ph.D. candidate within the SAFE research team at the GREYC laboratory in Caen, Normandy University, in partnership with Fime, where she also works as an engineer. With a background in applied mathematics from the French engineering school Polytech Lyon, her research focuses on the fairness, trustworthiness, and responsible governance of artificial intelligence and biometric systems. She studies demographic and intersectional disparities, with particular attention to the robustness, reliability, and interpretability of fairness assessments, as well as to the ethical and governance challenges associated with the development and deployment of these technologies.

Joël DI MANNO is a consultant specializing in biometrics, artificial intelligence, digital identity, and authentication technologies. He received his engineering degree from ENSICAEN, the French National Graduate School of Engineering of Caen, in 2005. He currently works as a Biometrics and AI Consultant at Fime, where he contributes to the evaluation, security, and deployment of biometric and identity solutions.

Alongside his consulting activities, he delivers lectures and professional training modules at universities and engineering schools. His teaching covers biometrics, authentication, e-banking, and environmental, social, and governance topics.

Christophe CHARRIER obtained his Ph.D. in Computer Science from Jean Monnet University, Saint-Étienne, in 1998. From 1998 to 2001, he was a Research Assistant with the Laboratory of Radio Communications and Signal Processing at Laval University, Quebec, Canada. In 2001, he joined the Cherbourg Institute of Technology at Saint-Lô as an Associate Professor and was promoted to Full Professor in 2024.

In 2008, he joined the GREYC laboratory and its SAFE research group. His current research interests include digital image and video forensics, deepfake detection, image and video coding, quality assessment, computational vision, and biometrics. He has authored or co-authored more than 150 international publications.

Christophe Rosenberger obtained his Ph.D. in Information Technology from the University of Rennes 1 in 1999. He has been a Full Professor at ENSICAEN, Caen, since 2007. His research focuses on cybersecurity, particularly biometrics, including keystroke dynamics, soft biometrics, biometric system evaluation, and fingerprint quality assessment, as well as digital forensics. He has authored or co-authored more than 200 international publications.