Fairness in Big Data

Speaker
Prof. Chiara GALDI
EURECOM

Contact
chiara.galdi@eurecom.fr

Journées nationales 2026 du GDR Sécurité Informatique

9 juin 2026

Outline

- Part I: Introduction to Fairness in Big Data
- Part II: Measuring and Mitigating Bias
- Part III: Use Case
- Part IV: Conclusions

Outline

Let's start from the definition...

- Oxford English Dictionary:
  - "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges." (2013)
  - "Extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions" (2021)
- Cambridge Dictionary:
  - "very large sets of data that are produced by people using the internet, and that can only be stored, understood, and used with the help of special tools and methods" (2025)
- IBM Definition:
  - "Big data refers to massive, complex data sets that traditional data management systems cannot handle. When properly collected, managed and analyzed, big data can help organizations discover new insights and make better business decisions." (2024)

More definitions!

From the ISO/IEC Technical Report 24027: 2021

"Bias in AI systems and AI aided decision making"

Bias in artificial intelligence (AI) systems can manifest in different ways. AI systems that learn patterns from data can potentially reflect existing societal bias against groups. While some bias is necessary to address the AI system objectives (i.e. desired bias), there can be bias that is not intended in the objectives and thus represent unwanted bias in the AI system.

Bias vs. Fairness

Bias is the *"systematic difference in treatment of certain objects, people, or groups in comparison to others."* [1]

Fairness can be described as "a treatment, a behaviour or an outcome that respects established facts, beliefs and norms and is not determined by favouritism or unjust discrimination." [1]

[1] From the ISO/IEC TR 24027: 2021

Overview of Bias in AI

Classification and clustering algorithms cannot function without bias. However, the impact of this bias could be positive, neutral or negative according to the system goals and objectives.

- **Positive effect:** AI developers can introduce bias to ensure a fair result;
- **Neutral effect:** bias does not impact the outcome of the system;
- **Negative effect:** bias can have unintended consequences of limiting the opportunities of those affected.

One challenge with determining the relevance of bias is that what constitutes negative effect can depend on the specific use case or application domain.

Overview of Fairness in AI

Within the context of AI, it is difficult to define fairness in a manner that will apply equally well to all AI systems in all contexts.

- **Unfair allocation:** occurs when an AI system unfairly extends or withholds opportunities or resources for some groups;
- **Unfair quality of service:** AI system performs less well for some groups;
- **Stereotyping:** AI system reinforces existing societal stereotypes;
- **Denigration:** AI system behaves in ways that are derogatory or demeaning;
- **“Over“ or “under“ representation and erasure:** AI system over-represents or under-represents some groups.

Remember that...

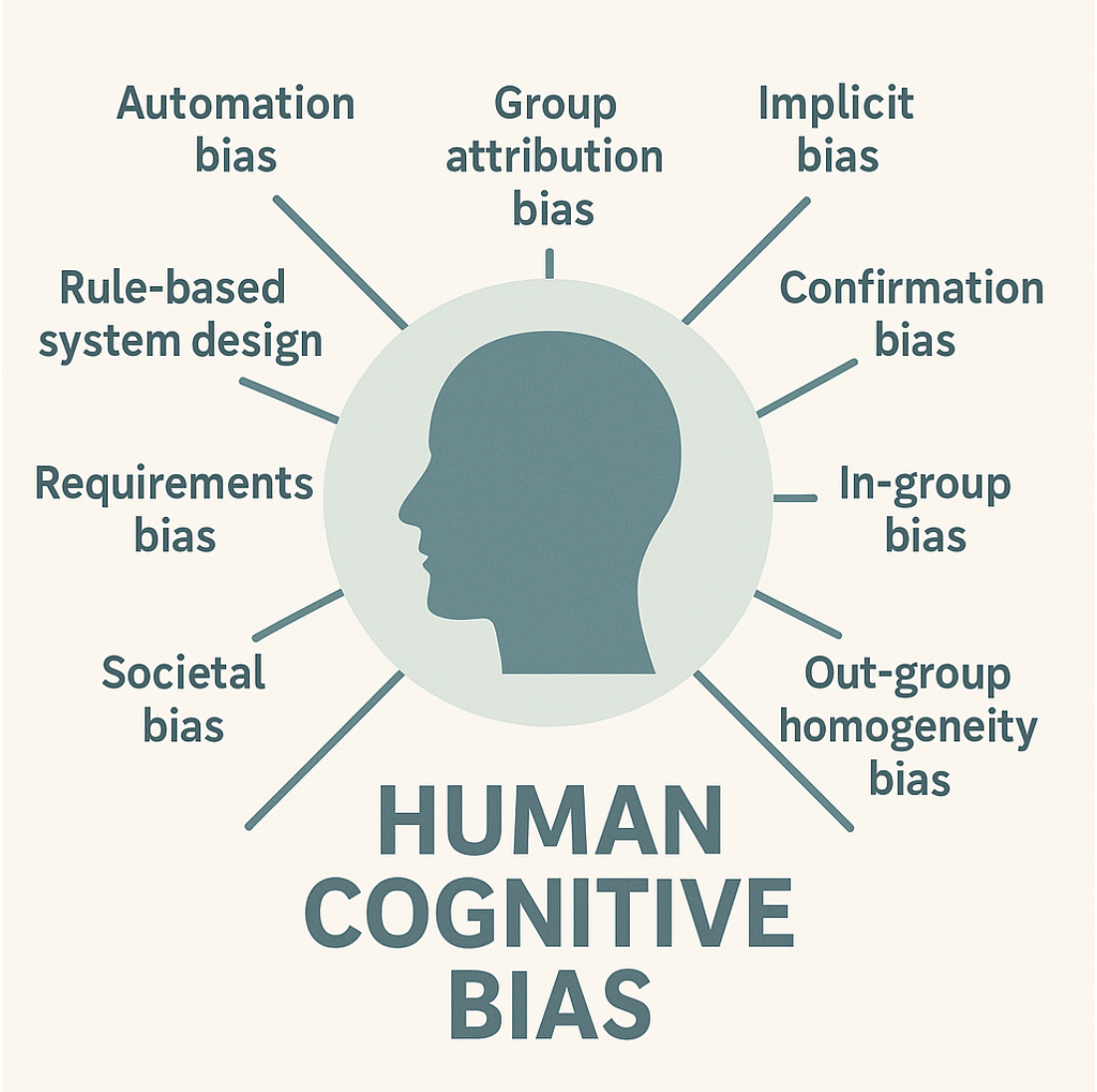
👉 Bias is just one of many elements that can influence fairness.

Achieving fairness in AI systems often means making trade-offs. 👈

Sources of Unwanted Bias

- Human cognitive biases
- Data bias
- Bias introduced by engineering decisions

Human Cognitive Biases



Automation Bias

Automation bias occurs when a human favours recommendations made by an automated system over information made without automation, even when the automation makes errors.

Societal Bias

Societal bias occurs when similar cognitive bias (conscious or unconscious) is being held by many individuals in society. Consequently, this bias can be encoded in datasets and ML models learn or amplify these pre-existing, historical patterns of bias.

Confirmation Bias

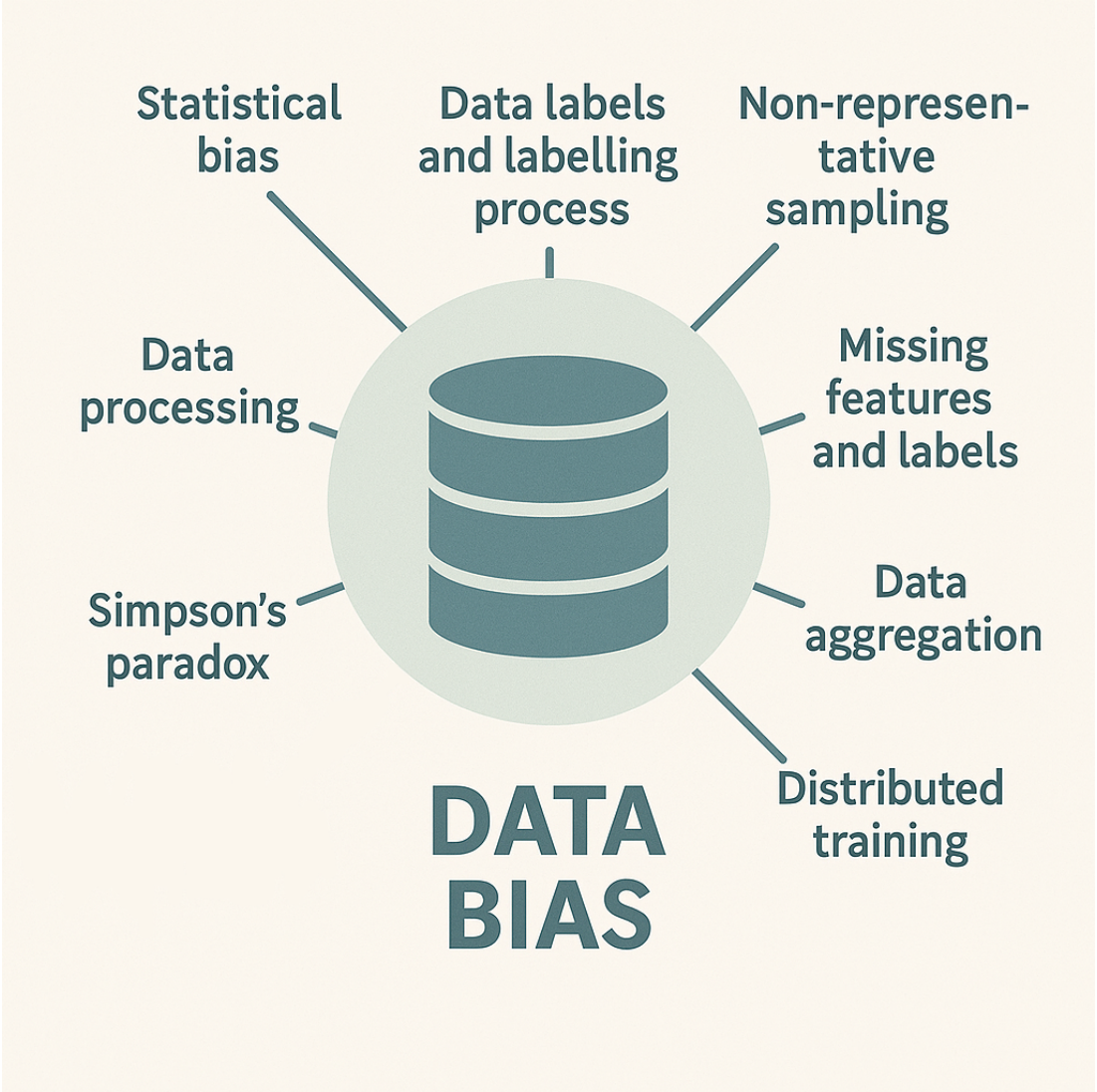
Confirmation bias occurs when hypotheses, regardless of their veracity, are more likely to be confirmed by the intentional or unintentional interpretation of information.

In-Group Bias and Out-Group Homogeneity Bias

In-group bias occurs when showing partiality to one's own group or own characteristics.

Out-group homogeneity bias occurs when seeing out-group members as more alike than in-group members.

Data Bias


Statistical bias
Data labels and labelling process
Non-representative sampling
Data processing
Missing features and labels
Simpson's paradox
Data aggregation
Distributed training
DATA BIAS

Statistical bias

Sampling bias occurs when data records are not collected randomly from the intended population.

Coverage bias occurs when a population represented in a dataset does not match the population that the ML model is making predictions about.

...

Data Labels and Labelling Process

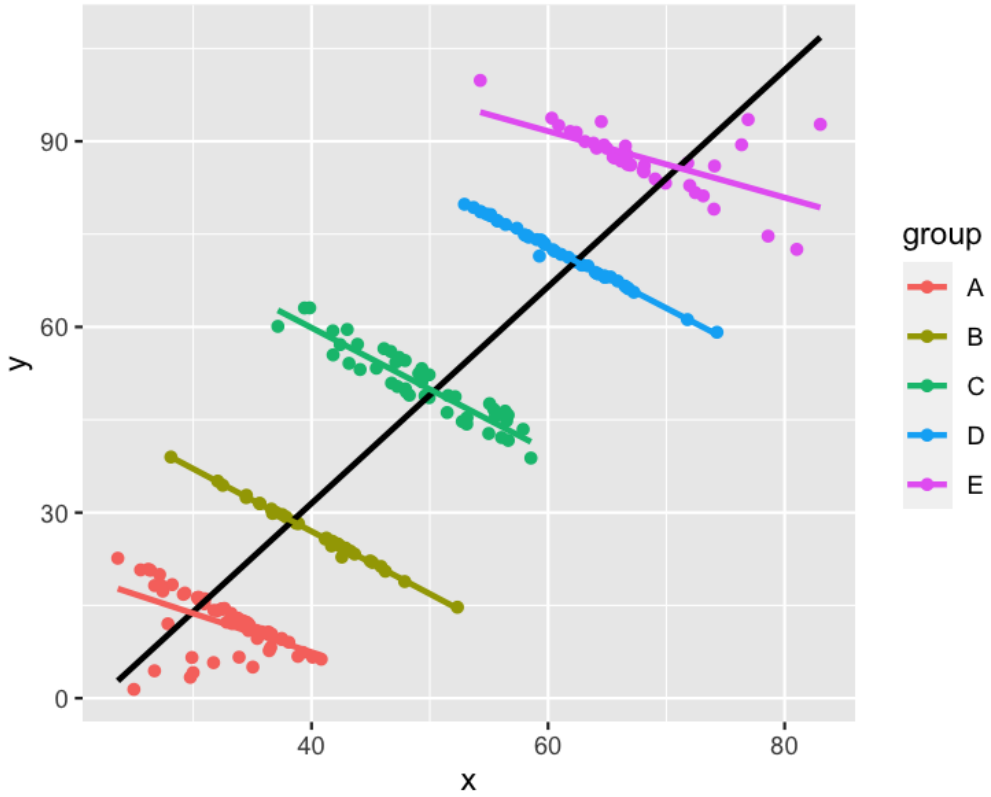
The **labelling process** itself potentially introduces the cognitive or societal biases to the data. For example, by deciding to classify people into male or female, or old and young, people are cast into discrete categories that do not necessarily represent the full reality being modelled. Another example is when the true labels are inaccessible and proxies for ground truth are used and are accepted as sufficiently close for most purposes.

Missing Features and Labels

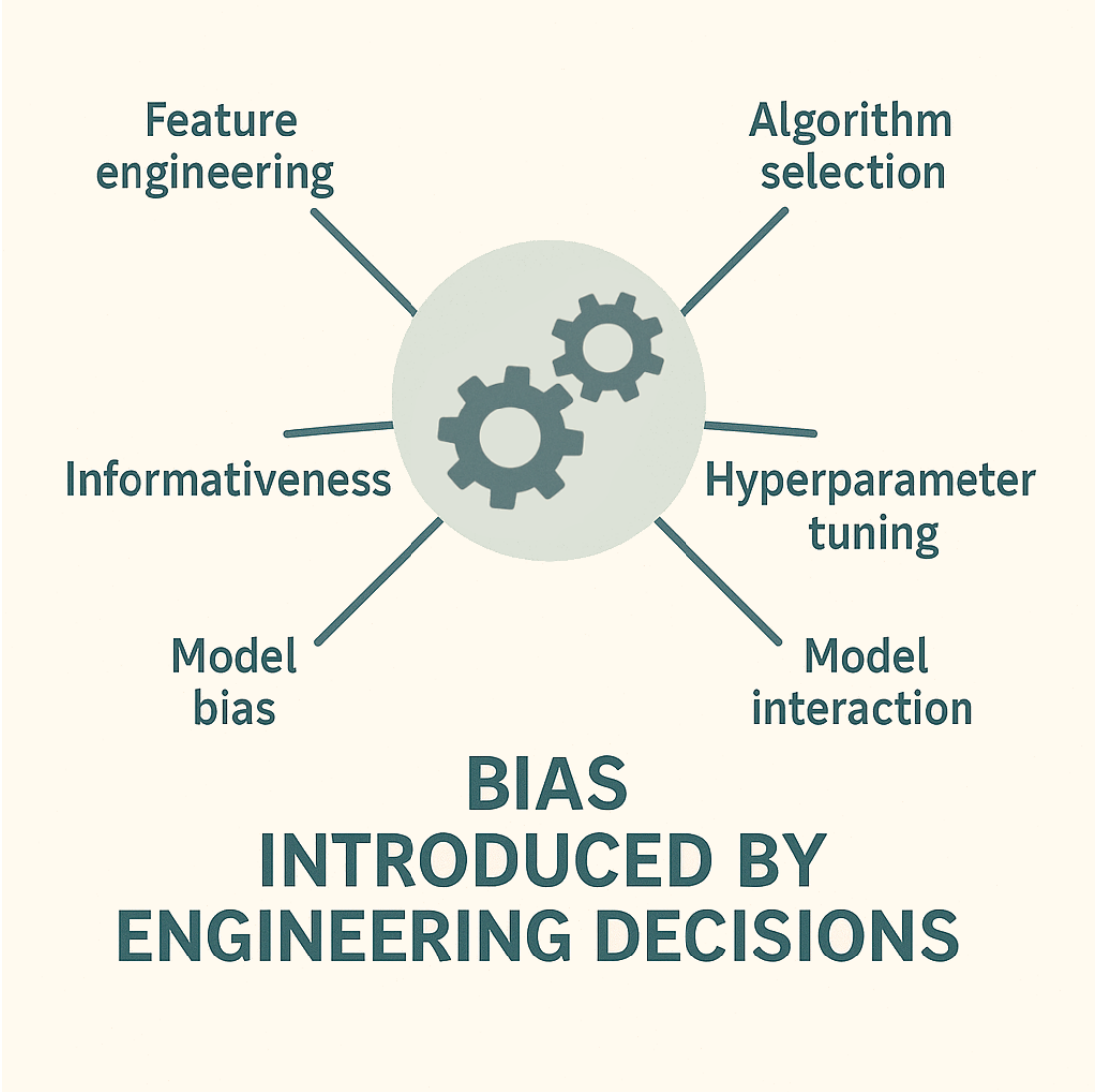
Real world data are rarely complete. In particular, **features are often missing** from individual training samples. If the frequency of missing features is higher for one group than another then this presents another vector for bias.

Simpson's Paradox

Simpson's paradox manifests when a trend that is indicated in individual groups of data reverses when the groups of data are combined.



Bias Introduced by Engineering Decisions



Model Bias

Given that ML often uses functions like a maximum likelihood estimator (MLE) to determine parameters, if there is under-representation present in the data, the MLE tends to amplify any underlying bias in the distribution. That is, the ML model learns patterns proportional to what it sees most often in the training data. As a result, minority patterns get underrepresented because they contribute less to the overall likelihood function.

Feature Engineering

The AI developers can directly use any of the input features or can **create complex features** for the ML model. Steps such as encoding, data type conversion, dimensionality reduction and feature selection are subject to choices made by the AI developer and can introduce bias in the ML model.

Informativeness

For some groups the mapping between inputs present in the data and outputs are more difficult to learn. This can happen when some features are **highly informative** about one group, while a different set of features is highly informative about another group. If this is the case, then a model that only has one feature set available, can be biased against the group whose relationships are difficult to learn from available data.

Outline

How to Assess Fairness?

Metrics of fairness seek to evaluate differences between average observed values and true values. $\rightarrow$ **Protected groups should be treated similarly.**

However, there are different fairness notions used in machine learning:

- Demographic Parity;
- Equality of Opportunity;
- Equalized Odds;
- Predictive equality.

Demographic Parity

The **decision** (prediction $\hat{Y}$) should be independent of sensitive attributes A (like gender or ethnicity).

Equal prediction rates between categories:

$$P(\hat{Y} = \hat{y}|A = m) = P(\hat{Y} = \hat{y}|A = n)$$

for all values m, n that A can take.

Drawbacks of Demographic Parity

A ML model could achieve demographic parity (i.e., its predictions could be independent of sensitive group membership), but still generate more false positive predictions for one group versus others. **Stricter metrics** are designed that requires that the model's predictions are not only independent of sensitive group membership, but that groups experience the same TP or/and FP rates.

Equality of Opportunity

Equality of opportunity requires equal true positive rates across groups. An algorithm's decisions that $\hat{Y} = 1$ are independent of a category A given the input $Y = 1$.

$$P(\hat{Y} = 1 | Y = 1, A = m) = P(\hat{Y} = 1 | Y = 1, A = n)$$

for all values m, n that A can take.

Equalized Odds

The model's true positive rate and false positive rate should be equal across groups.

Equalized odds means that an algorithm's decisions are independent of a category A given the input Y.

$$P(\hat{Y} = \hat{y} | Y = y, A = m) = P(\hat{Y} = \hat{y} | Y = y, A = n)$$

for all values m, n that A can take.

Predictive Equality

Predictive equality implies equal false positive rates across demographic categories.

$$P(\hat{Y} = 1 | Y = 0, A = m) = P(\hat{Y} = 1 | Y = 0, A = n)$$

for all values m, n that A can take.

How can we Mitigate Bias?

To ensure fairness in ML models, a holistic approach must be adopted in order to mitigate bias at all levels of the system where it may occur:

- Data representation and labelling;
- Training and tuning;
- Adversarial methods.

Bias mitigation: Data Representation and Labelling

Decide how to best represent the training data in **features** that are interpretable by the model.

For example: Which data records must be included in training? What features must be selected? How the data relates to the purpose of the system? Which individuals are choosing features and what is their rationale?

It is important to evaluate the chosen features for any data and human cognitive biases. Where crowd workers are used, it can be useful to understand the diversity and goals of the people who annotate the data and how they are incentivised.

Bias mitigation: Training and tuning

Training data: Ensure that training data is balanced (sampling) and analyse features to find out the feature contribution and the relative significance of each feature in a model's prediction (feature selection).

Tuning:

- Data-based: Up-sampling of under-represented populations or the use of synthetic data;
- Model-based: addition of regularization terms that enforce fairness during optimization;
- Post-hoc: identifying group-specific decision thresholds based on predicted outcomes.

Adversarial Methods to Mitigate Bias

An **adversary** is a secondary model that tries to predict a sensitive attribute (like gender or ethnicity) from the output of the main model.

During training, the main model is updated not only to perform its main task well but also to make it hard for the adversary to correctly guess the sensitive attribute.

As a result, the main model learns to make predictions that are independent (or orthogonal) to those sensitive characteristics, effectively reducing bias related to them.

Outline

Use case: Bias in Face Recognition

In 2019, NIST has publish a report "Face Recognition Vendor Test Part 3: Demographic Effects"

Included 18.27 million images of 8.49 million individuals processed by 189 algorithms from 99 developers.

The **majority** of algorithms showed variation in error rates depending on ethnicity, gender, or age.

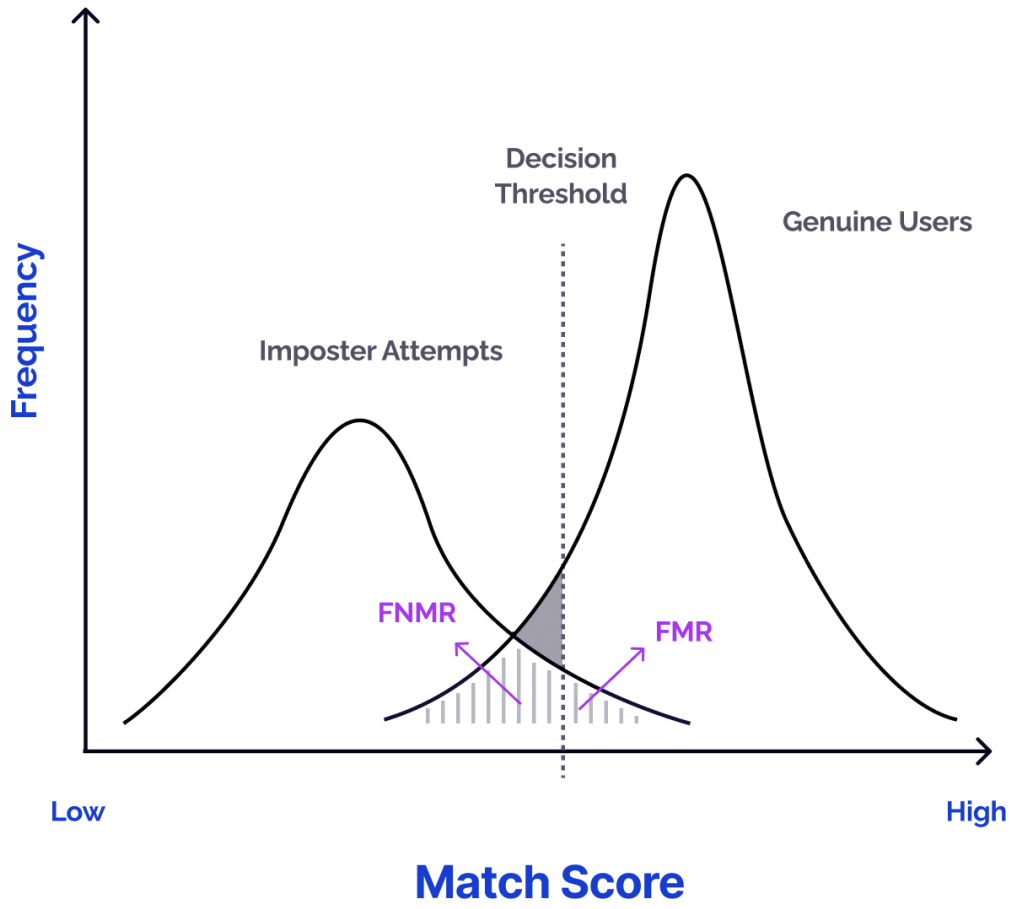
Caucasian		Asian		Black	
Male	Female	Male	Female	Male	Female
					
98%	98%	93%	93%	93%	95%

Image source[2]

[2] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome and J. Fierrez, "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition", in IAPR Iberoamerican Congress on Pattern Recognition (CIARP), LNCS, Springer, vol. 11401, Madrid, Spain, November 2018

Metrics of Fairness in Face Recognition

Most of the fairness metrics in Face Recognition are based on the observation of differentials in the error rates of a fare recognition system.



FMR: False Match Rate; **FNMR:** False Non-Match Rate

Proposed Fairness Metrics in Face Recognition

There is still **no agreement** on the best metric for measuring the fairness of a facial recognition system.

In 2022 *Howard et al.* proposed Gini Aggregation Rate for Biometric Equitability (GARBE): a fairness metrics that uses a Gini-coefficient style aggregation of error-rates to summarise fairness across multiple demographic groups.

GARBE

Functional Fairness Measure Criteria (FFMC):

1. The net contributions of FMR and FNMR differentials to the overall fairness measure should be intuitive.
2. The fairness measure must be bounded, with a minimum and maximum possible value.
3. The fairness measure should be calculable when no errors are observed for a demographic group.

Outline

Conclusions

- Fairness is not a formula, it's a process.
- In Big Data and AI, fairness depends on choices made at every step: what data we collect, how we model it, which objectives we optimize, and how the results are used.
- Achieving fair outcomes requires more than technical fixes, it requires awareness, accountability, and collaboration between disciplines and people.
- Building fair AI means understanding the entire system, from data to decisions.

News

- ANR JCJC GOOD-BIAS! project
  - Holistic and Generalizable Solutions for Bias Mitigation in Face Recognition
- Course on Responsible AI at EURECOM

The end...

Thank you.

Questions?

References

1. ISO/IEC Technical Report 24027: 2021. Information technology – Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
2. A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome and J. Fierrez, "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition", in IAPR Iberoamerican Congress on Pattern Recognition (CIARP), LNCS, Springer, vol. 11401, Madrid, Spain, November 2018
3. John J. Howard, Eli J. Laird, Rebecca E. Rubin, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 2022. "Evaluating Proposed Fairness Models for Face Recognition Algorithms". In Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges: Montreal, QC, Canada, August 21–25, 2022, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 431–447.

Our Work on Fairness and Biometric System Vulnerabilities

- Fair-Gate: Fairness-Aware Interpretable Risk Gating for Sex-Fair Voice Biometrics. Y Qu, M Todisco, C Galdi, N Evans. IWBF 2026.
- A comparison of differential performance metrics for the evaluation of automatic speaker verification fairness. O Chouchane, C Busch, C Galdi, N Evans, M Todisco. ODYSSEY 2024, The Speaker and Language Recognition Workshop (pp. 209-216).
- Fairness and privacy in voice biometrics: A study of gender influences using wav2vec 2.0. O Chouchane, M Panariello, C Galdi, M Todisco, N Evans. 2023 International Conference of the Biometrics Special Interest Group
- Lost in light field compression: understanding the unseen pitfalls in computer vision. A Zizien, C Galdi, K Fliegel, JL Dugelay. Signal Processing: Image Communication 136, 117304
- Facial biometrics in the social media era: An in-depth analysis of the challenge posed by beautification filters. N Mirabet-Herranz, C Galdi, JL Dugelay. IEEE TBIOM 7 (1), 108-117