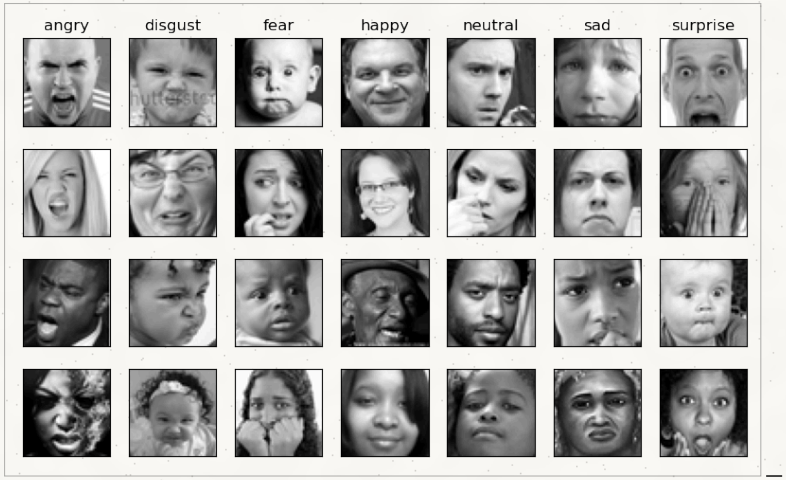
AMCO SEMINAR · UNIVERSITY OF PADOVA · NOVEMBER 2026

Who Is in the Data?

Measuring and mitigating demographic bias, from facial expression recognition to web-scale datasets



Samples of FER+, a popular FER dataset: seven emotions, no demographic labels

Iris Dominguez-Catena

Public University of Navarre (UPNA) · Institute of Smart Cities · ARIN

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

arin
Artificial Intelligence &
Machine Learning Research

About me

2010–15	BSc Computer Engineering, UPNA
2018–20	MSc Computer Engineering: channel-wise OWA operators in CNNs
2020–24	PhD, UPNA: demographic bias in ML (research stay at Ghent University)
2024–	Assistant Professor at UPNA, ISC and ARIN
Now	One month visiting the AMCO team here

> **Interests**: dataset bias · algorithmic fairness · facial analysis · LLM bias



Where I come from

Pamplona

Capital of Navarre, northern Spain. About 210,000 inhabitants, founded as Pompaelo around 75 BC. Home of San Fermín, on the Camino de Santiago

UPNA

Public University of Navarre, created in 1987. About 10,000 students, six faculties and schools, campuses in Pamplona and Tudela

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

ARIN

Artificial Intelligence and Machine Learning Research group, led by Mikel Galar. Constituted in 2024, 13 current members

arin
Artificial Intelligence & Machine Learning Research



Outline

01 **Why look at the data?**
FER as a case study

02 **Chapter I · Taxonomy and metrics**
PhD thesis

03 **Chapter II · DSAP**
PhD thesis

04 **Chapter III · Bias transference**
PhD thesis

05 **Quantification**
Better demographic profiling of datasets · ongoing

06 **LAION-5B**
Demographic profiling at web scale · ongoing

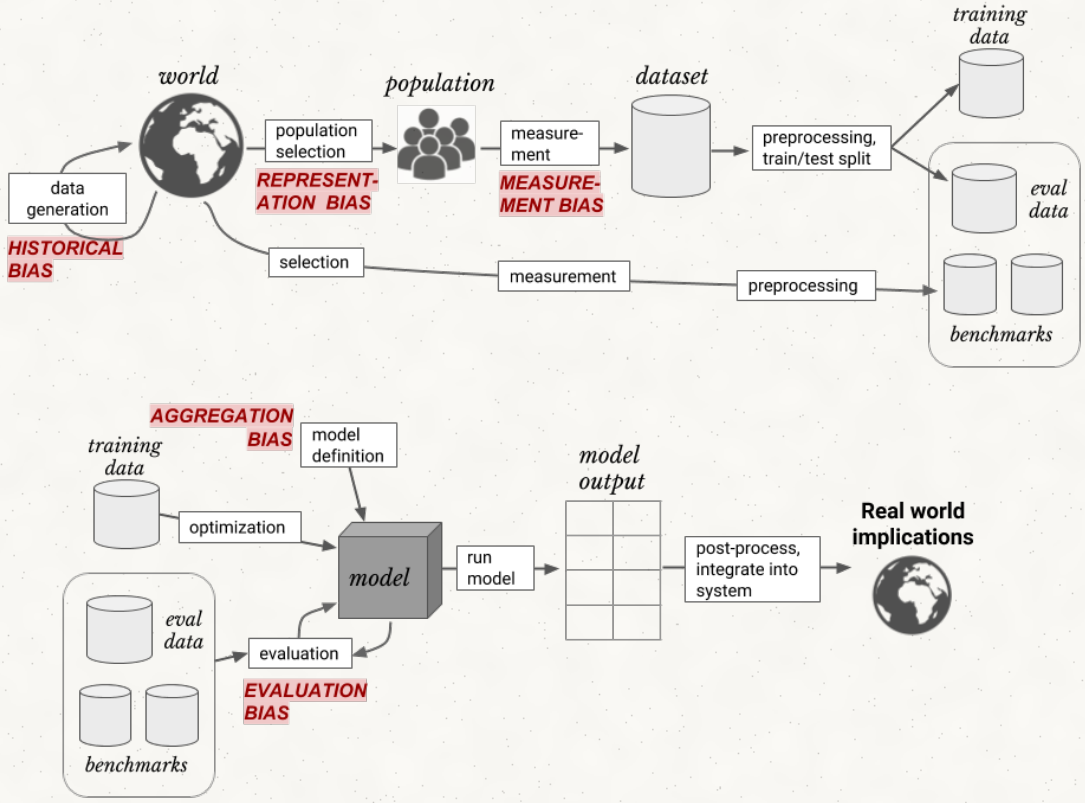
07 **Beyond images**
LLM ideology (Ghent)

01

Why look at the data?

Facial expression recognition as a case study

Models inherit what is in the data

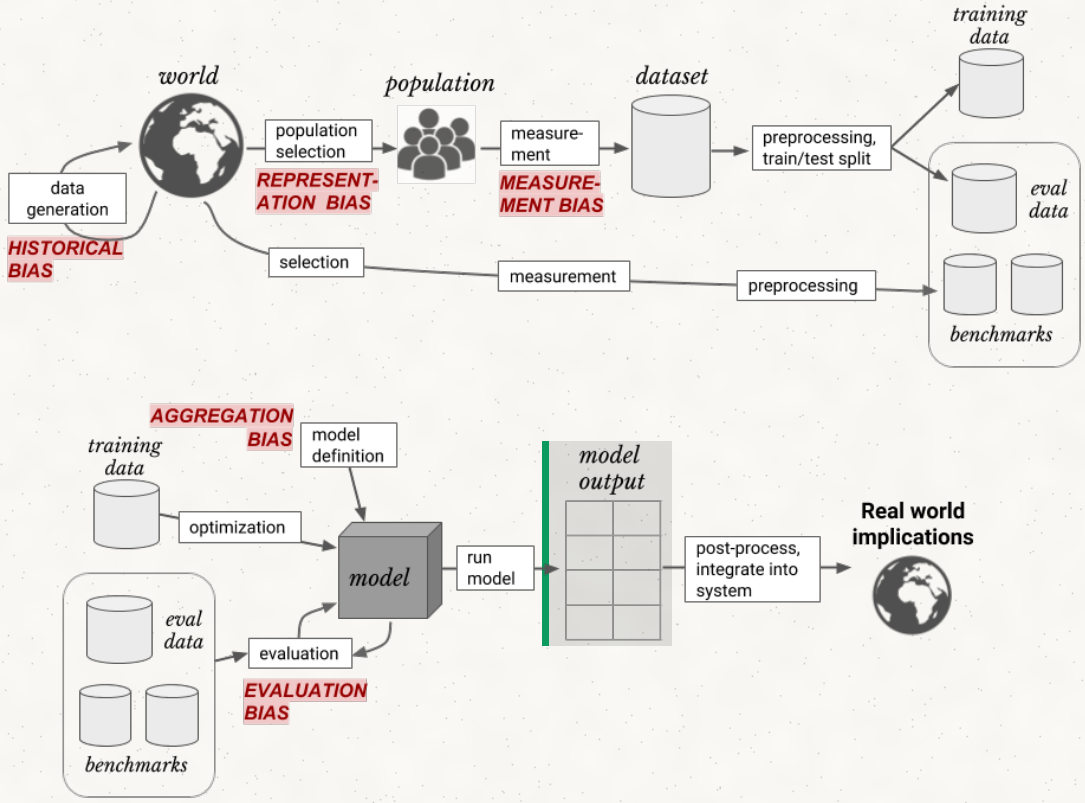


- Bias can enter at every stage of the ML pipeline
- Model bias is well studied; **dataset bias** much less
- Almost no studies on how dataset bias **transfers** to the model

Sources of harm in the ML pipeline: from the world to the dataset (top), from the dataset to the model (bottom). [1]

[1] H. Suresh, J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *EAAMO* 2021

Models inherit what is in the data

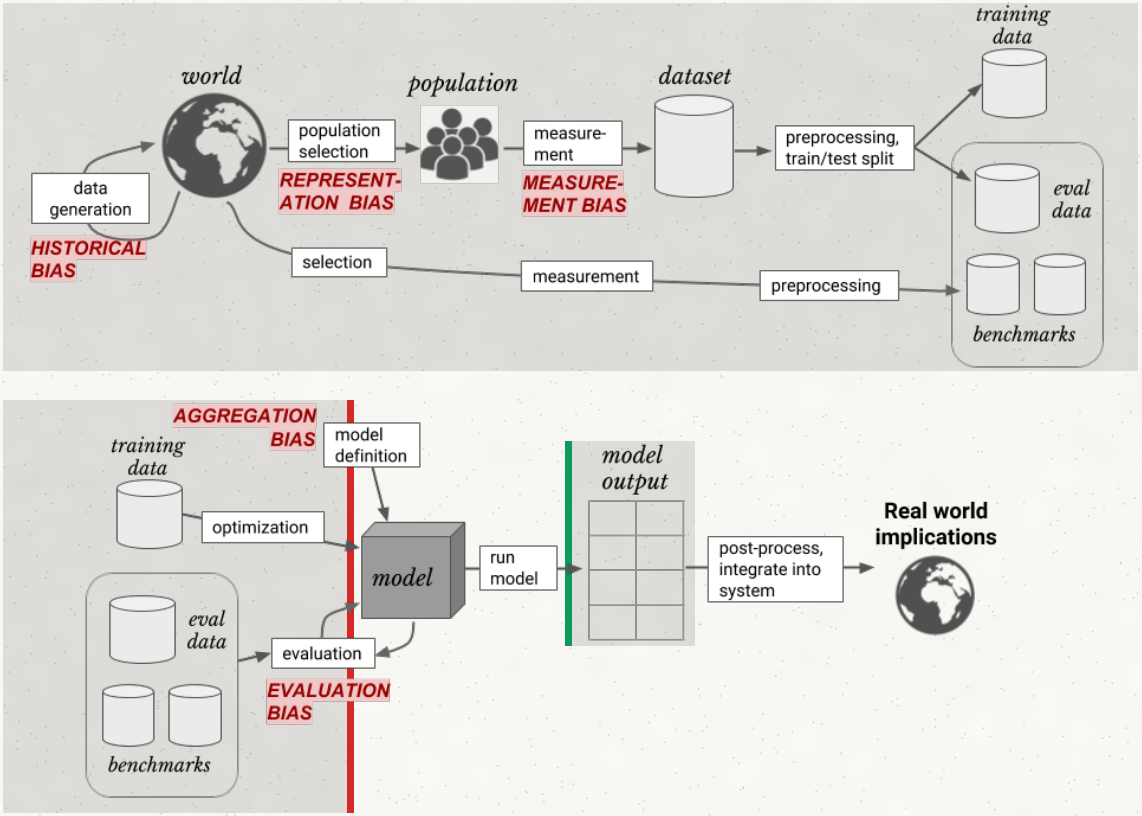


- Bias can enter at every stage of the ML pipeline
- Model bias is well studied; **dataset bias** much less
- Almost no studies on how dataset bias **transfers** to the model

Sources of harm in the ML pipeline: from the world to the dataset (top), from the dataset to the model (bottom). [1]

[1] H. Suresh, J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *EAAMO* 2021

Models inherit what is in the data

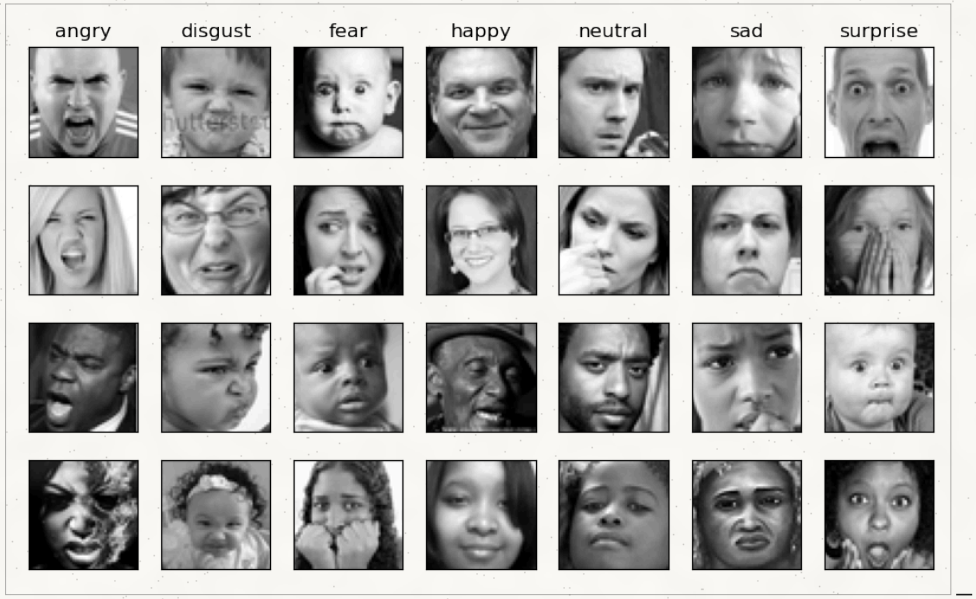


- Bias can enter at every stage of the ML pipeline
- Model bias is well studied; **dataset bias** much less
- Almost no studies on how dataset bias **transfers** to the model

Sources of harm in the ML pipeline: from the world to the dataset (top), from the dataset to the model (bottom). [1]

[1] H. Suresh, J. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *EAAMO* 2021

Facial expression recognition as a case study



A sample of FER+ [2]

- Applications: healthcare, assistive robotics, driver monitoring, HCI
- Known age, race and gender biases in research and commercial systems
- Shift from lab datasets to large **in-the-wild** web datasets with **no demographic metadata**

> Naturally **multiclass** (7 emotions) and **multigroup** (several races, ages, genders), unlike the usual binary-outcome, two-group setting of fairness work on tabular data

[2] E. Barsoum et al. Training deep networks for facial expression recognition with crowd-sourced label distribution. *ICMI* 2016

20 FER datasets, three sources

Dataset	Year	Source	Images	Videos	Subjects
JAFFE	1998	LAB	213	---	10
KDEF	1998	LAB	4,900	---	70
CK	2000	LAB	8,795	486	97
Oulu-CASIA	2008	LAB	66,000	480	80
CK+	2010	LAB	10,727	593	123
GEMEP	2010	LAB	2,817	1,260	10
MUG	2010	LAB	70,654	---	52
SFEW	2011	ITW-M	1,766	---	330
FER2013	2013	ITW-I	32,298	---	---
WSEFEP	2014	LAB	210	---	30

Dataset	Year	Source	Images	Videos	Subjects
ADFES	2016	LAB	---	648	22
FERPlus	2016	ITW-I	32,298	---	---
AffectNet	2017	ITW-I	291,652	---	---
ExpW	2017	ITW-I	91,793	---	---
RAF-DB	2017	ITW-I	29,672	---	---
CAER-S	2019	ITW-M	70,000	---	---
LIRIS-CSE	2019	LAB	26,000	208	12
MMAFEDB	2020	ITW-I	128,000	---	---
NHFIER	2020	ITW-I	5,558	---	---
iSAFE	2020	LAB	---	395	44

55 candidates → **20 datasets**

- **Lab**
- **ITW-I**: Internet
- **ITW-M**: movies

No demographic labels → we infer them with **FairFace** [3]

[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021

Research questions

1 How do we **describe and measure** demographic bias in a dataset? CHAPTER I

2 How do we **compare** datasets that have no demographic labels? CHAPTER II

3 How does each type of dataset bias **transfer** to the model? CHAPTER III

4 How **reliable** are the demographic estimates we use to audit data, and can we do better? ONGOING

02

Chapter I · Taxonomy and metrics

Dominguez-Catena, Paternain & Galar. IEEE TPAMI 2024

Types of bias in image datasets

Selection bias

Number of samples for each group

EXAMPLE
20,000 images of white people and 20 of Black people

Label bias

Labeling scheme and assignment

EXAMPLE
Not labeling children as “person”

Framing bias

Image properties and context

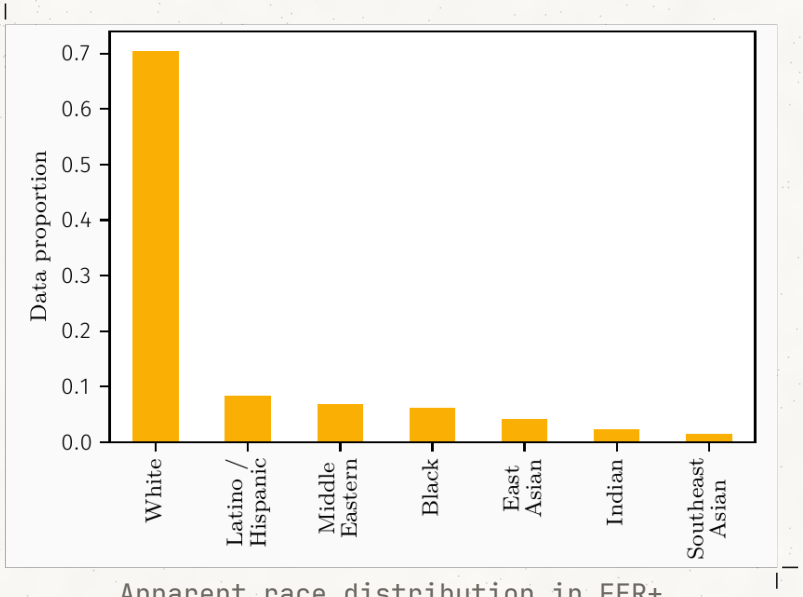
EXAMPLE
Lower image quality in images from certain countries

We focus on **selection bias**: the easiest to measure and quantify, and not limited to image datasets. [4]

[4] S. Fabbrizzi et al. A survey on bias in visual datasets. *Computer Vision and Image Understanding* 2022

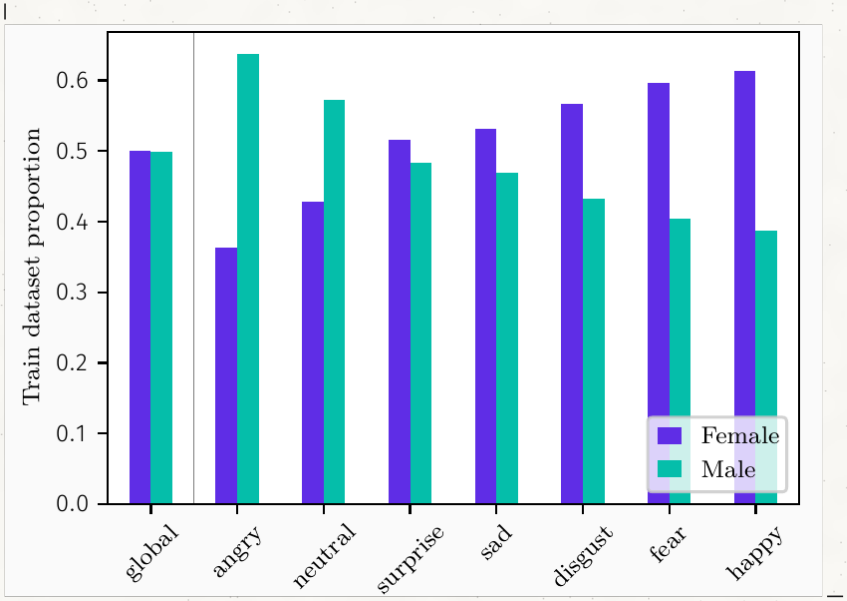
Representational vs. stereotypical bias

REPRESENTATIONAL BIAS



Apparent race distribution in FER+
Global group proportions: mostly white

STEREOTYPICAL BIAS



Apparent per-label gender distribution in FER+
Group-label association: more women “happy” and “fear”, more men “angry”

> The same dataset can look bad on one axis and fine on the other.

A taxonomy of dataset demographic bias

Selection bias

dataset demographic bias

Representational · group proportions

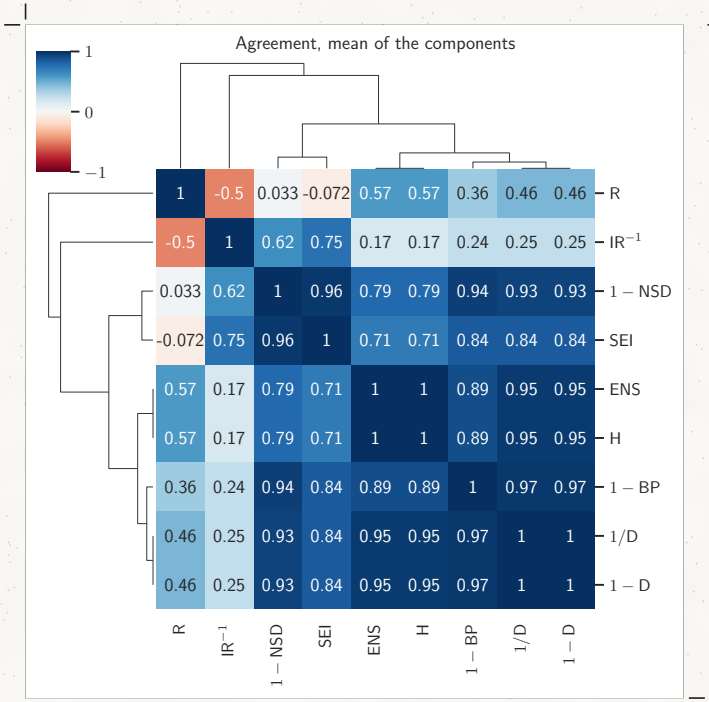
RICHNESS | EVENNESS | DOMINANCE | COMBINED

Stereotypical · group–label association

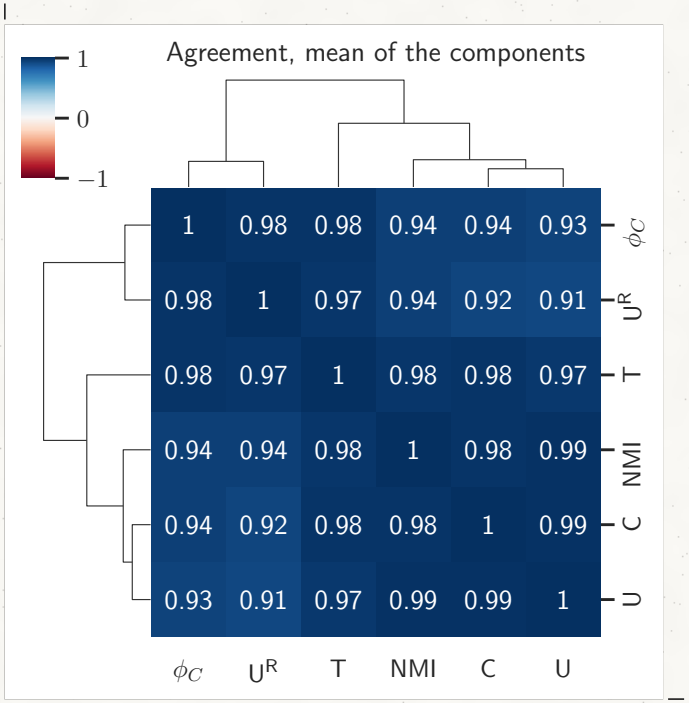
GLOBAL: ONE VALUE PER DATASET | LOCAL: ONE VALUE PER GROUP AND LABEL

- We reviewed **17 metrics** from ecology, information theory and statistics
- Too many for practitioners: next, which ones are redundant?

Which metrics agree?



Representational metrics



Stereotypical metrics

- Measured all metrics on 20 FER datasets (age, race, gender via FairFace)
- Rank correlation → many metrics are **redundant**
- Recommended set: **ENS** (combined), **SEI** (evenness), **BP** (dominance), **Cramér's V** (global stereotypical), **Ducher's Z** (local)

What the metrics say about FER datasets

		MUG	GEMEP	ADFES	Oulu-CASIA	KDEF	CK+	CK	WSEFEP	iSAFE	LIRIS-CSE	JAFFE	Average
Age	9 – ENS	7.164	4.395	7.644	6.376	7.351	7.275	7.298	7.404	7.100	7.732	8.000	7.067 ± 0.932
	1 – SEI	0.124	0.051	0.561	0.304	0.279	0.607	0.516	0.574	0.416	0.657	–	0.409 ± 0.200
	ϕ_C	0.053	$0.165^{\triangle}$	0.000	0.021	0.002	0.125°	0.137°	0.000	0.117°	0.134°	–	0.075 ± 0.063
Race	7 – ENS	4.889	6.000	4.522	4.806	6.000	4.141	4.488	6.000	5.581	5.096	5.320	5.168 ± 0.634
	1 – SEI	0.320	–	0.174	0.285	–	0.460	0.486	–	0.682	0.414	0.252	0.384 ± 0.151
	ϕ_C	0.035	–	0.000	0.038	–	0.073°	0.096°	–	0.197°	$0.256^{\triangle}$	0.039	0.092 ± 0.083
Gender	2 – ENS	0.014	0.001	0.074	0.233	0.098	0.055	0.039	0.000	0.012	0.001	1.000	0.139 ± 0.280
	1 – SEI	0.010	0.000	0.054	0.179	0.073	0.040	0.029	0.000	0.009	0.001	–	0.039 ± 0.052
	ϕ_C	0.054	0.093	0.000	0.026	0.002	0.023	0.053	0.000	0.108°	$0.313^{\triangle}$	–	0.067 ± 0.090

More bias →

Lab datasets

		SFEW	CAER-S	Average
	– ENS	4.561	6.541	5.551 ± 0.990
	1 – SEI	0.283	0.590	0.437 ± 0.154
	ϕ_C	0.115°	0.052°	0.084 ± 0.031
	– ENS	5.151	4.990	5.070 ± 0.080
	1 – SEI	0.684	0.641	0.663 ± 0.021
	ϕ_C	0.076°	0.040	0.058 ± 0.018
	– ENS	0.019	0.129	0.074 ± 0.055
	1 – SEI	0.013	0.096	0.055 ± 0.041
	ϕ_C	0.261°	0.138°	0.199 ± 0.062

More bias →

ITW-M

		EXPW	MMAFEDB	RAF-DB	AFFECTNET	FER+	FER2013	NHFIER	Average
Age	9 – ENS	3.017	3.144	3.258	3.149	3.407	3.414	3.954	3.334 ± 0.286
	1 – SEI	0.186	0.196	0.205	0.196	0.216	0.217	0.263	0.211 ± 0.024
	ϕ_C	0.072°	0.070°	$0.144^{\triangle}$	0.095°	0.105°	0.104°	$0.138^{\triangle}$	0.104 ± 0.027
Race	7 – ENS	3.334	3.445	3.317	3.848	4.011	4.014	4.100	3.724 ± 0.321
	1 – SEI	0.332	0.348	0.330	0.410	0.437	0.438	0.453	0.393 ± 0.050
	ϕ_C	0.041	0.058°	0.058°	0.041	0.068°	0.086°	0.087°	0.063 ± 0.018
Gender	2 – ENS	0.013	0.010	0.008	0.000	0.000	0.000	0.006	0.005 ± 0.005
	1 – SEI	0.009	0.007	0.006	0.000	0.000	0.000	0.004	0.004 ± 0.004
	ϕ_C	0.157°	0.167°	0.131°	0.195°	0.171°	0.178°	0.172°	0.167 ± 0.018

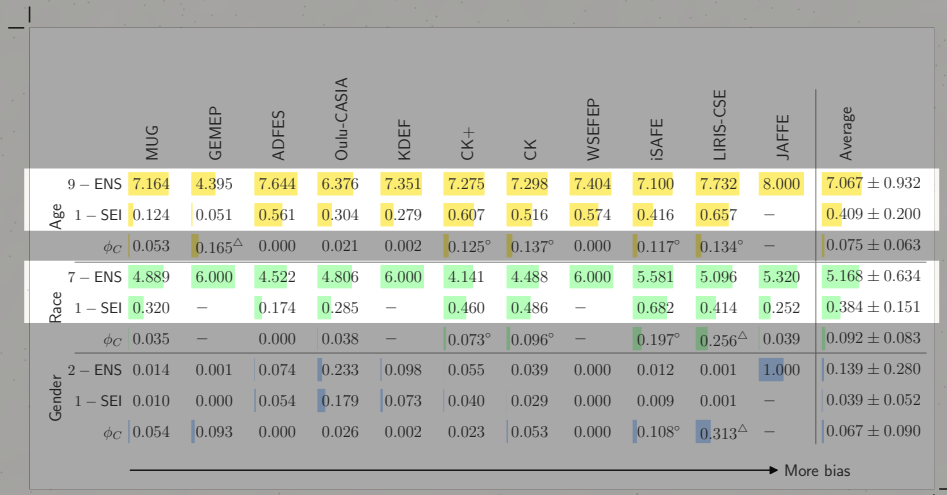
More bias →

ITW-I

Lab: little representational diversity, but low stereotypes

Internet: more diverse, but **more gender stereotypical bias**

What the metrics say about FER datasets



		MUG	GEMEP	ADFES	Oulu-CASIA	KDEF	CK+	CK	WSEFEP	iSAFE	LIRIS-CSE	JAFFE	Average
Age	9 – ENS	7.164	4.395	7.644	6.376	7.351	7.275	7.298	7.404	7.100	7.732	8.000	7.067 ± 0.932
	1 – SEI	0.124	0.051	0.561	0.304	0.279	0.607	0.516	0.574	0.416	0.657	–	0.409 ± 0.200
	ϕ_C	0.053	$0.165^{\triangle}$	0.000	0.021	0.002	0.125°	0.137°	0.000	0.117°	0.134°	–	0.075 ± 0.063
Race	7 – ENS	4.889	6.000	4.522	4.806	6.000	4.141	4.488	6.000	5.581	5.096	5.320	5.168 ± 0.634
	1 – SEI	0.320	–	0.174	0.285	–	0.460	0.486	–	0.682	0.414	0.252	0.384 ± 0.151
	ϕ_C	0.035	–	0.000	0.038	–	0.073°	0.096°	–	0.197°	$0.256^{\triangle}$	0.039	0.092 ± 0.083
Gender	2 – ENS	0.014	0.001	0.074	0.233	0.098	0.055	0.039	0.000	0.012	0.001	1.000	0.139 ± 0.280
	1 – SEI	0.010	0.000	0.054	0.179	0.073	0.040	0.029	0.000	0.009	0.001	–	0.039 ± 0.052
	ϕ_C	0.054	0.093	0.000	0.026	0.002	0.023	0.053	0.000	0.108°	$0.313^{\triangle}$	–	0.067 ± 0.090

More bias →

Lab datasets

		SFEW	CAER-S	Average
	9 – ENS	4.561	6.541	5.551 ± 0.990
	1 – SEI	0.283	0.590	0.437 ± 0.154
	ϕ_C	0.115°	0.052°	0.084 ± 0.031
	7 – ENS	5.151	4.990	5.070 ± 0.080
	1 – SEI	0.684	0.641	0.663 ± 0.021
	ϕ_C	0.076°	0.040	0.058 ± 0.018
	2 – ENS	0.019	0.129	0.074 ± 0.055
	1 – SEI	0.013	0.096	0.055 ± 0.041
	ϕ_C	0.261°	0.138°	0.199 ± 0.062

More bias →

ITW-M

		EXPW	MMAFEDB	RAF-DB	AFFECTNET	FER+	FER2013	NHFIER	Average
Age	9 – ENS	3.017	3.144	3.258	3.149	3.407	3.414	3.954	3.334 ± 0.286
	1 – SEI	0.186	0.196	0.205	0.196	0.216	0.217	0.263	0.211 ± 0.024
	ϕ_C	0.072°	0.070°	$0.144^{\triangle}$	0.095°	0.105°	0.104°	$0.138^{\triangle}$	0.104 ± 0.027
Race	7 – ENS	3.334	3.445	3.317	3.848	4.011	4.014	4.100	3.724 ± 0.321
	1 – SEI	0.332	0.348	0.330	0.410	0.437	0.438	0.453	0.393 ± 0.050
	ϕ_C	0.041	0.058°	0.058°	0.041	0.068°	0.086°	0.087°	0.063 ± 0.018
Gender	2 – ENS	0.013	0.010	0.008	0.000	0.000	0.000	0.006	0.005 ± 0.005
	1 – SEI	0.009	0.007	0.006	0.000	0.000	0.000	0.004	0.004 ± 0.004
	ϕ_C	0.157°	0.167°	0.131°	0.195°	0.171°	0.178°	0.172°	0.167 ± 0.018

More bias →

ITW-I

Lab: little representational diversity, but low stereotypes

Internet: more diverse, but **more gender stereotypical bias**

What the metrics say about FER datasets

		MUG	GEMEP	ADFES	Oulu-CASIA	KDEF	CK+	CK	WSEFEP	iSAFE	LIRIS-CSE	JAFFE	Average
Age	9 – ENS	7.164	4.395	7.644	6.376	7.351	7.275	7.298	7.404	7.100	7.732	8.000	7.067 ± 0.932
	1 – SEI	0.124	0.051	0.561	0.304	0.279	0.607	0.516	0.574	0.416	0.657	–	0.409 ± 0.200
	ϕ_C	0.053	$0.165^{\triangle}$	0.000	0.021	0.002	0.125°	0.137°	0.000	0.117°	0.134°	–	0.075 ± 0.063
Race	7 – ENS	4.889	6.000	4.522	4.806	6.000	4.141	4.488	6.000	5.581	5.096	5.320	5.168 ± 0.634
	1 – SEI	0.320	–	0.174	0.285	–	0.460	0.486	–	0.682	0.414	0.252	0.384 ± 0.151
	ϕ_C	0.035	–	0.000	0.038	–	0.073°	0.096°	–	0.197°	$0.256^{\triangle}$	0.039	0.092 ± 0.083
Gender	2 – ENS	0.014	0.001	0.074	0.233	0.098	0.055	0.039	0.000	0.012	0.001	1.000	0.139 ± 0.280
	1 – SEI	0.010	0.000	0.054	0.179	0.073	0.040	0.029	0.000	0.009	0.001	–	0.039 ± 0.052
	ϕ_C	0.054	0.093	0.000	0.026	0.002	0.023	0.053	0.000	0.108°	$0.313^{\triangle}$	–	0.067 ± 0.090

More bias →

Lab datasets

		SFEW	CAER-S	Average
	– ENS	4.561	6.541	5.551 ± 0.990
	1 – SEI	0.283	0.590	0.437 ± 0.154
	ϕ_C	0.115°	0.052°	0.084 ± 0.031
	– ENS	5.151	4.990	5.070 ± 0.080
	1 – SEI	0.684	0.641	0.663 ± 0.021
	ϕ_C	0.076°	0.040	0.058 ± 0.018
	– ENS	0.019	0.129	0.074 ± 0.055
	1 – SEI	0.013	0.096	0.055 ± 0.041
	ϕ_C	0.261°	0.138°	0.199 ± 0.062

More bias →

ITW-M

		EXPW	MMAFEDB	RAF-DB	AFFECTNET	FER+	FER2013	NHFIER	Average
Age	9 – ENS	3.017	3.144	3.258	3.149	3.407	3.414	3.954	3.334 ± 0.286
	1 – SEI	0.186	0.196	0.205	0.196	0.216	0.217	0.263	0.211 ± 0.024
	ϕ_C	0.072°	0.070°	$0.144^{\triangle}$	0.095°	0.105°	0.104°	$0.138^{\triangle}$	0.104 ± 0.027
Race	7 – ENS	3.334	3.445	3.317	3.848	4.011	4.014	4.100	3.724 ± 0.321
	1 – SEI	0.332	0.348	0.330	0.410	0.437	0.438	0.453	0.393 ± 0.050
	ϕ_C	0.041	0.058°	0.058°	0.041	0.068°	0.086°	0.087°	0.063 ± 0.018
Gender	2 – ENS	0.013	0.010	0.008	0.000	0.000	0.000	0.006	0.005 ± 0.005
	1 – SEI	0.009	0.007	0.006	0.000	0.000	0.000	0.004	0.004 ± 0.004
	ϕ_C	0.157°	0.167°	0.131°	0.195°	0.171°	0.178°	0.172°	0.167 ± 0.018

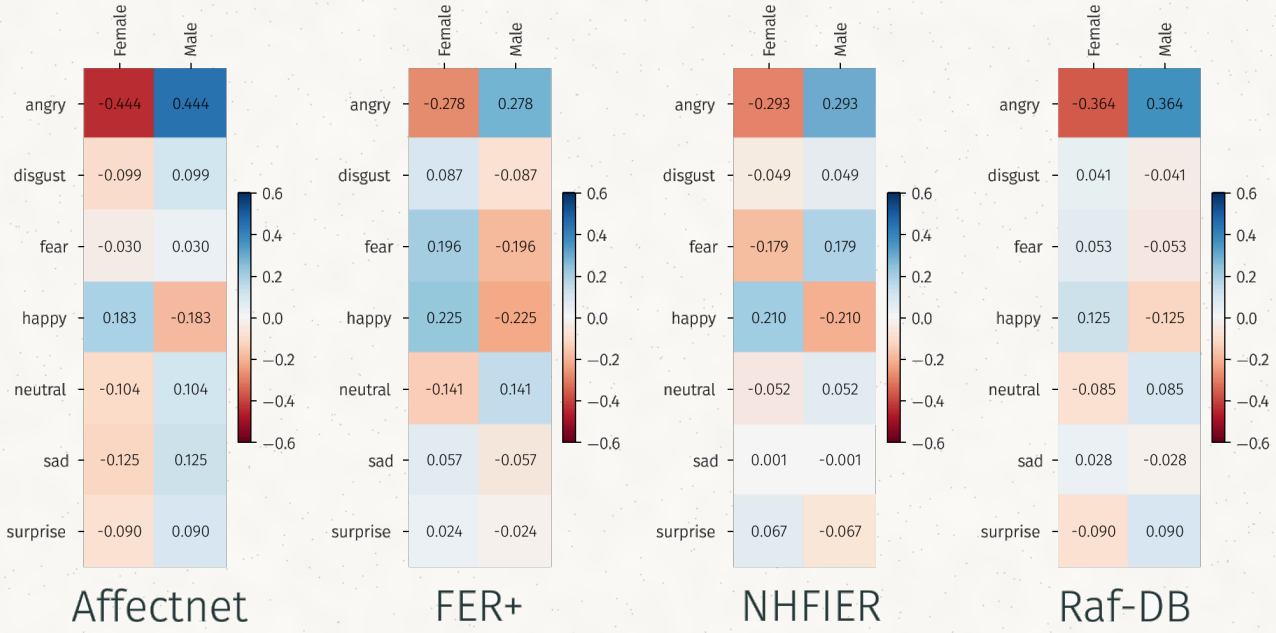
More bias →

ITW-I

Lab: little representational diversity, but low stereotypes

Internet: more diverse, but **more gender stereotypical bias**

Local stereotypical bias

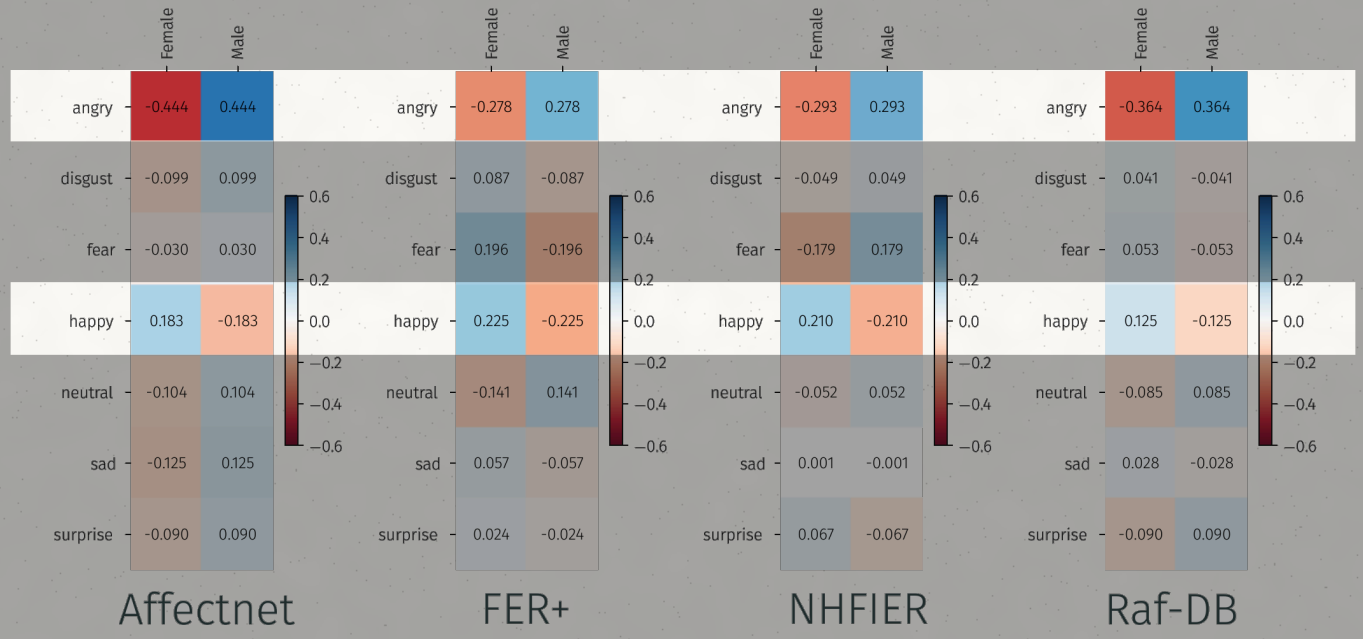


Ducher's Z between emotion and gender in four ITW-I datasets [5]. Blue: overrepresented, red: underrepresented

Across datasets: more **men** labeled "angry", more **women** labeled "happy"

[5] I. Dominguez-Catena, D. Paternain, M. Galar. Metrics for dataset demographic bias: a case study on facial expression recognition. *IEEE TPAMI* 2024

Local stereotypical bias



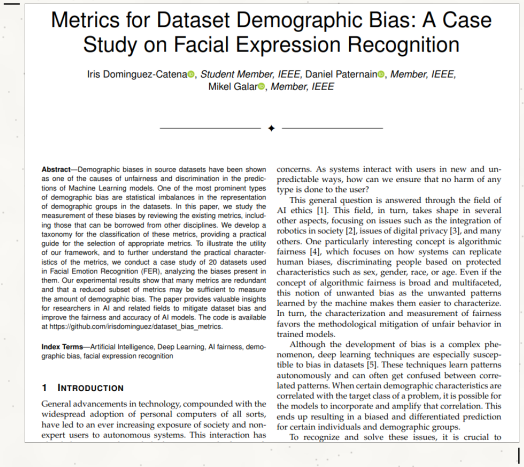
Ducher's Z between emotion and gender in four ITW-I datasets [5]. Blue: overrepresented, red: underrepresented

Across datasets: more **men** labeled "angry", more **women** labeled "happy"

[5] I. Dominguez-Catena, D. Paternain, M. Galar. Metrics for dataset demographic bias: a case study on facial expression recognition. *IEEE TPAMI* 2024

Output: TPAMI paper and library

- **Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition**
  Dominguez-Catena, Paternain, Galar. IEEE TPAMI 2024 [5]
- Open-source library, installable with
  `pip install dataset-bias-metrics`



Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition

Iris Dominguez-Catena, *Student Member, IEEE*, Daniel Paternain, *Member, IEEE*, Mikel Galar, *Member, IEEE*

Abstract—Demographic biases in source datasets have been shown as one of the causes of unfairness and discrimination in the predictions of Machine Learning models. One of the most prominent types of demographic bias are statistical imbalances in the representation of demographic groups in the datasets. In this paper, we study the measurement of these biases by reviewing the existing metrics, including those that can be borrowed from other disciplines. We develop a taxonomy for the classification of these metrics, providing a practical guide for the selection of appropriate metrics. To illustrate the utility of our framework, and to further understand the practical characteristics of the metrics, we conduct a case study of 20 datasets used in Facial Emotion Recognition (FER), analyzing the biases present in them. Our experimental results show that many metrics are redundant and that a reduced subset of metrics may be sufficient to measure the amount of demographic bias. The paper provides valuable insights for researchers in AI and related fields to mitigate dataset bias and improve the fairness and accuracy of AI models. The code is available at https://github.com/irisdominguez/dataset_bias_metrics.

Index Terms—Artificial Intelligence, Deep Learning, AI fairness, demographic bias, facial expression recognition

1 Introduction

General advancements in technology, compounded with the widespread adoption of personal computers of all sorts, have led to an ever increasing exposure of society and non-expert users to autonomous systems. This interaction has concerns. As systems interact with users in new and unpredictable ways, how can we ensure that no harm of any type is done to the user?

This general question is answered through the field of AI ethics [1]. This field, in turn, takes shape in several other aspects, focusing on issues such as the integration of robotics in society [2], issues of digital privacy [3], and many others. One particularly interesting concept is algorithmic fairness [4], which focuses on how systems can replicate human biases, discriminating people based on protected characteristics such as sex, gender, race, or age. Even if the concept of algorithmic fairness is broad and multifaceted, this notion of unwanted bias as the unwanted patterns learned by the machine makes them easier to characterize. In turn, the characterization and measurement of fairness favors the methodological mitigation of unfair behavior in trained models.

Although the development of bias is a complex phenomenon, deep learning techniques are especially susceptible to bias in datasets [5]. These techniques learn patterns autonomously and can often get confused between correlated patterns. When certain demographic characteristics are correlated with the target class of a problem, it is possible for the models to incorporate and amplify that correlation. This ends up resulting in a biased and differentiated prediction for certain individuals and demographic groups.

To recognize and solve these issues, it is crucial to

Paper

GitHub

PyPI

[5] I. Dominguez-Catena, D. Paternain, M. Galar. Metrics for dataset demographic bias: a case study on facial expression recognition. *IEEE TPAMI* 2024

03

Chapter II · DSAP

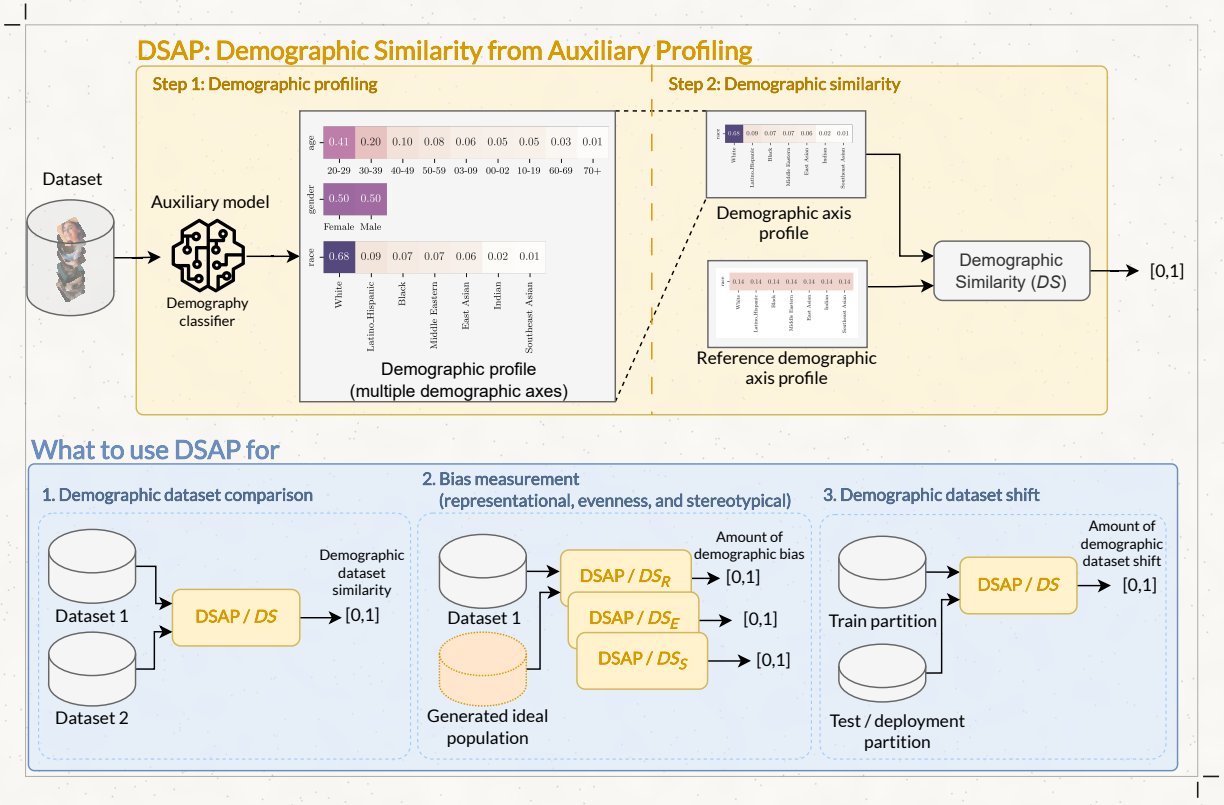
Demographic Similarity from Auxiliary Profiling · Information Fusion 2024

Metrics are hard to interpret

- What does ENS = 3.4 mean? Ranges differ across attributes
- Most ITW datasets have no demographic labels at all
- Often the real question is **relative**: “is my test set like my training set?”, “which datasets look alike?”

> Idea: compare demographic **profiles** instead of single numbers

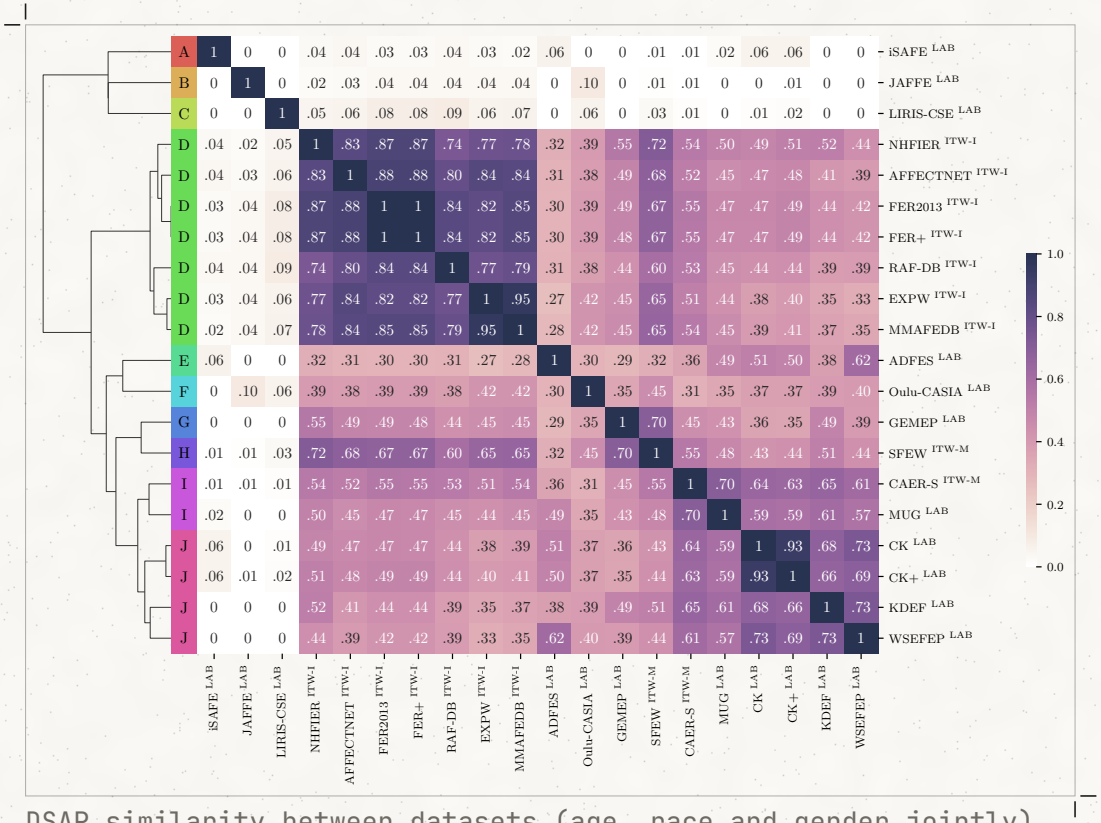
DSAP: Demographic Similarity from Auxiliary Profiling



- **1. Profile**: run FairFace on every image → joint age·race·gender distribution
- **2. Compare**: similarity between profiles
- Three uses: dataset comparison, bias against an ideal profile, dataset shift [6]

[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021 [6] I. Dominguez-Catena, D. Paternain, M. Galar. DSAP: analyzing bias through demographic comparison of datasets. *Information Fusion* 2024

Clustering 20 FER datasets



DSAP similarity between datasets (age, race and gender jointly)

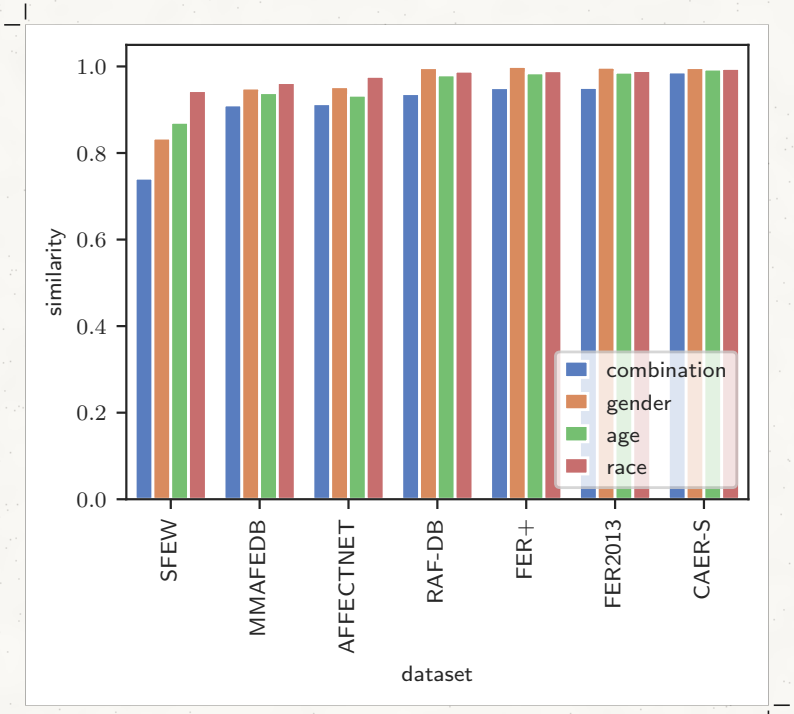
- Datasets cluster by **source**, not by size or year
- Many ITW datasets are demographically near-identical
- DSAP bias scores correlate with classical metrics, but are interpretable

Dataset shift between train and test

- Some official train/test partitions differ demographically
- DSAP detects it without demographic labels [6]
- Paper: Information Fusion 2024
- Code: github.com/irisdominguez/DSAP

Paper

Code



Similarity between train and test partitions

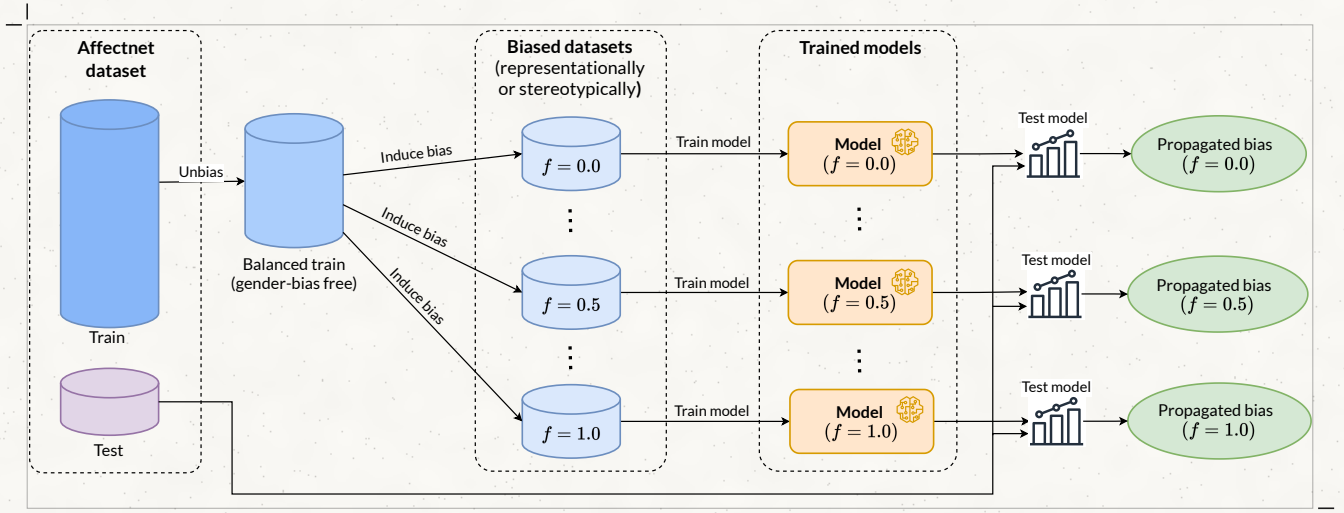
[6] I. Dominguez-Catena, D. Paternain, M. Galar. DSAP: analyzing bias through demographic comparison of datasets. *Information Fusion* 2024

04

Chapter III · Bias transference

Progress in AI 2024 · IEEE Trans. Affective Computing 2026 (with Ghent)

Does each bias type transfer differently?



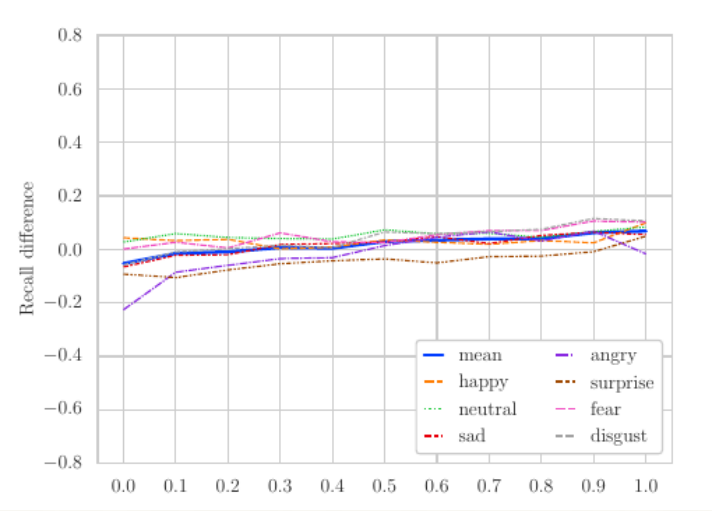
Induce controlled gender bias on AffectNet, train ResNet50 / ViT, measure recall gaps

Representational: fewer women overall

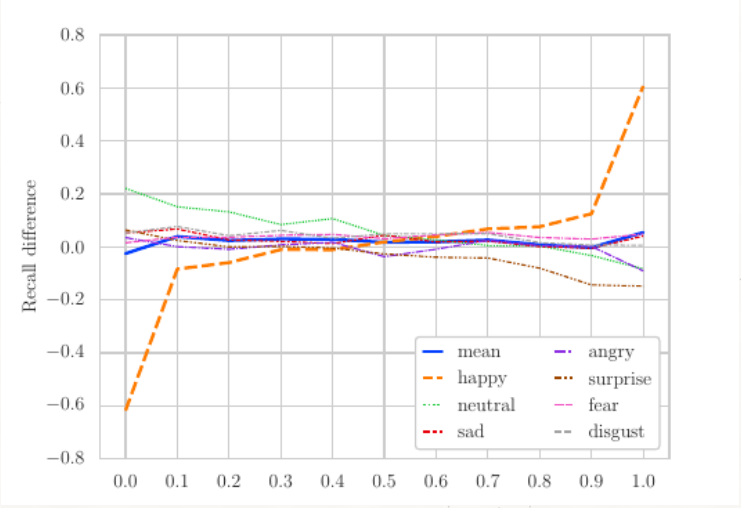
Stereotypical: fewer women in one emotion class

Less can be more

Representational



Stereotypical (happy)

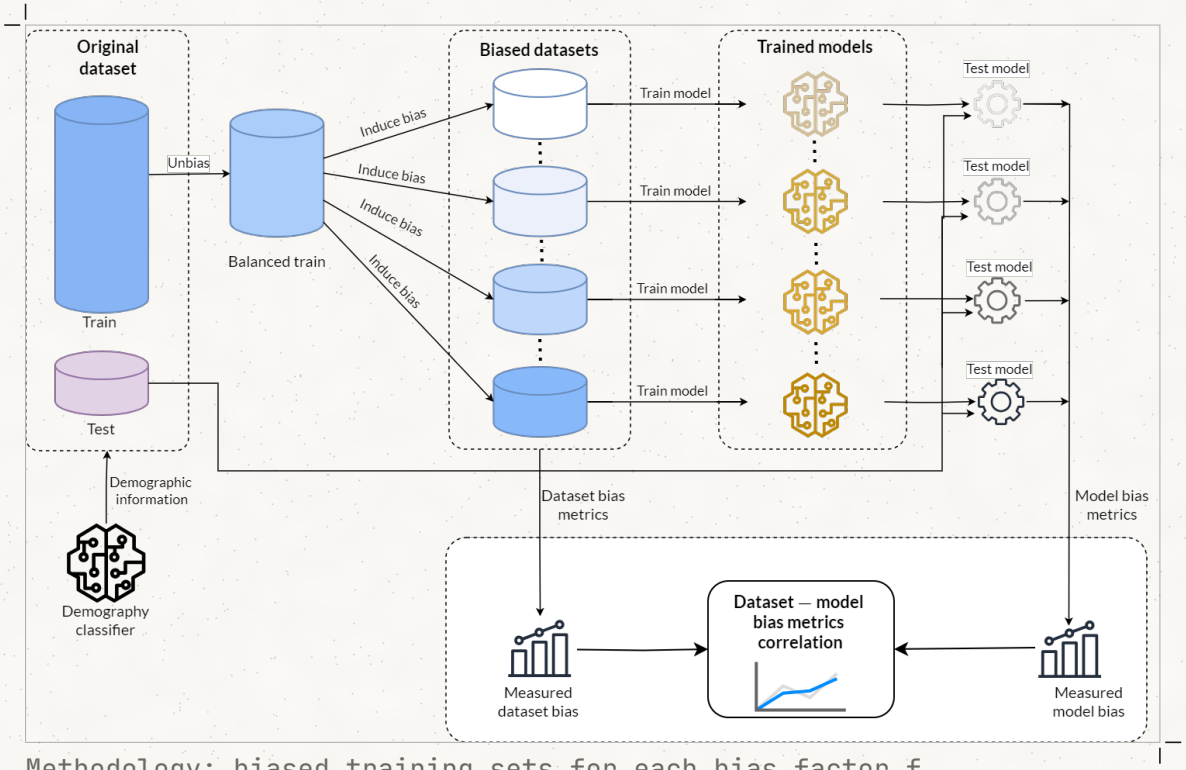


Recall difference (female minus male) as bias is induced

- Representational bias: **weak** effect on the model
- Stereotypical bias: **strong**, and localized to the biased class
- Removing data to break stereotypes can help more than adding data [7]

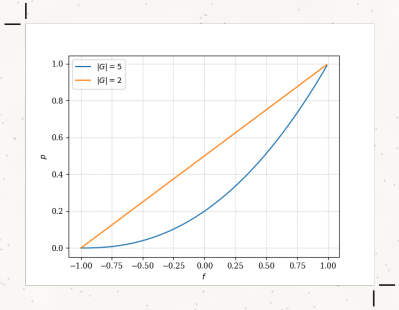
[7] I. Dominguez-Catena et al. Less can be more: representational vs. stereotypical gender bias in facial expression recognition. *Progress in AI* 2024

Measuring transference at scale



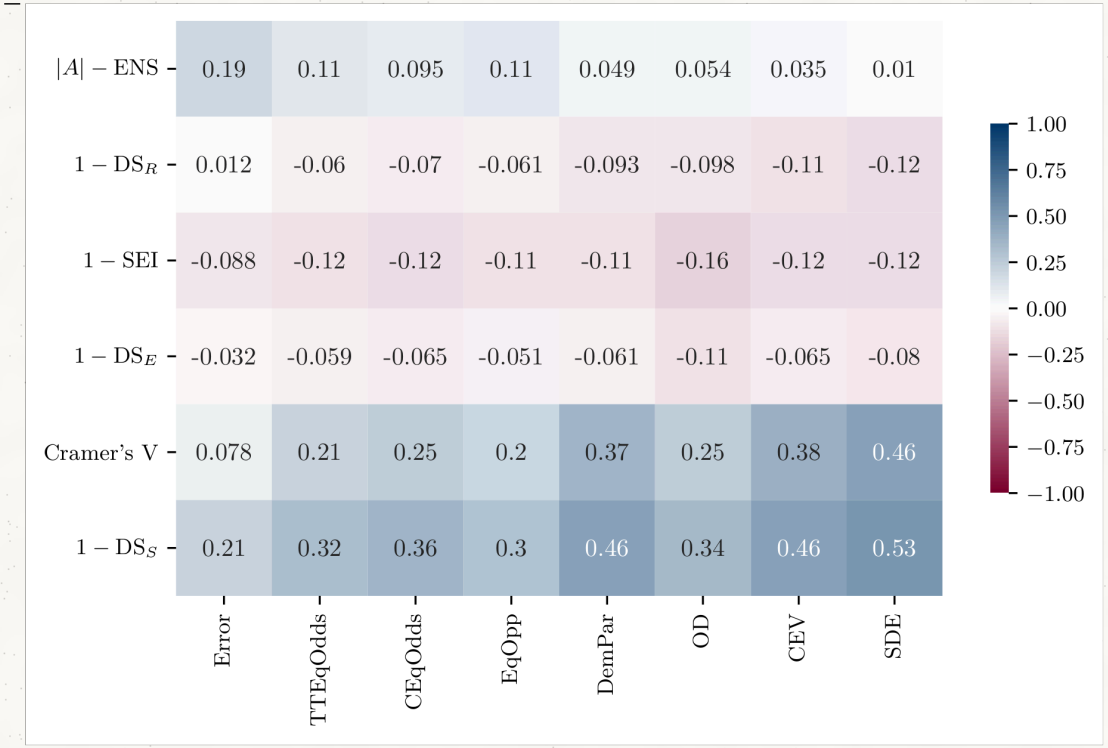
Methodology: biased training sets for each bias factor f

- Systematic bias scenarios: multiclass (7 emotions) and multigroup (race, gender) [8]
- Continuous bias factor $f \in [-1, 1]$
- Adapted multiclass, multigroup model bias metrics (TTEqOdds, CEqOdds, ...)



[8] I. Dominguez-Catena et al. Biased heritage: how datasets shape models in facial expression recognition. *IEEE Trans. Affective Computing* 2026

Which model metric sees which dataset bias?

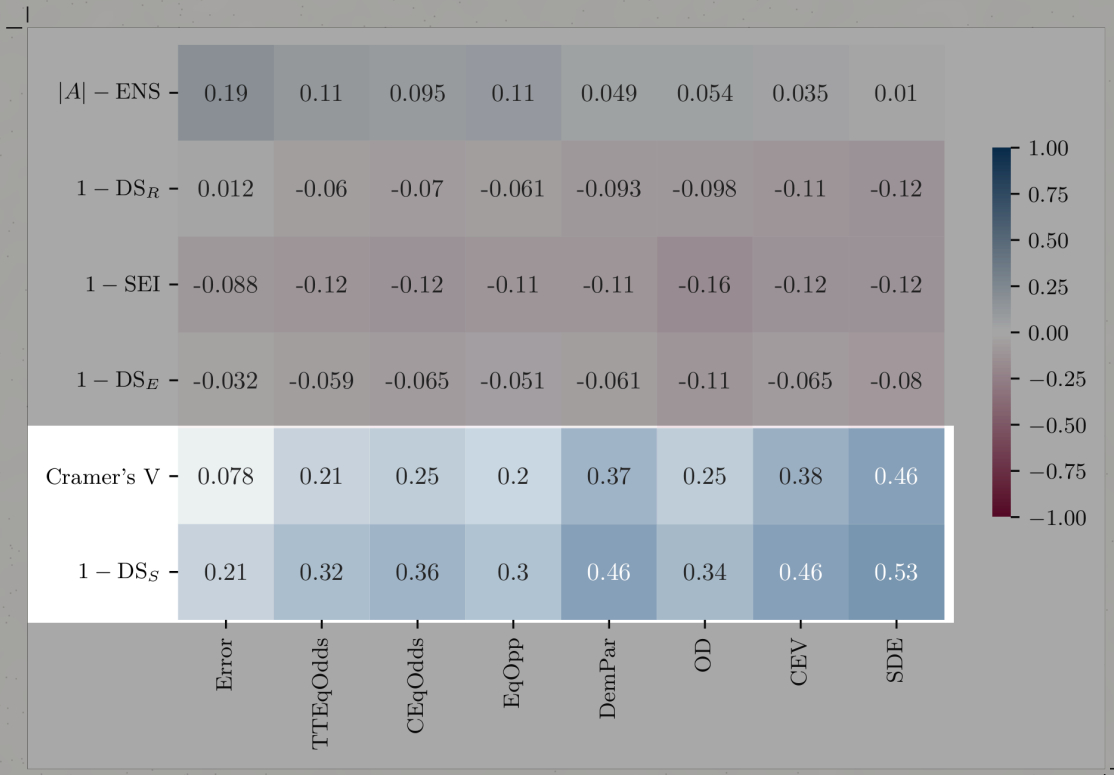


Spearman rank correlation between dataset bias (rows) and model bias metrics

- Stereotypical dataset bias ↔ strongly visible in equalized-odds metrics
- Representational bias probably has a smaller impact on the model
- But if there is any, current model bias metrics do not pick it up [8]

[8] I. Dominguez-Catena et al. Biased heritage: how datasets shape models in facial expression recognition. *IEEE Trans. Affective Computing* 2026

Which model metric sees which dataset bias?

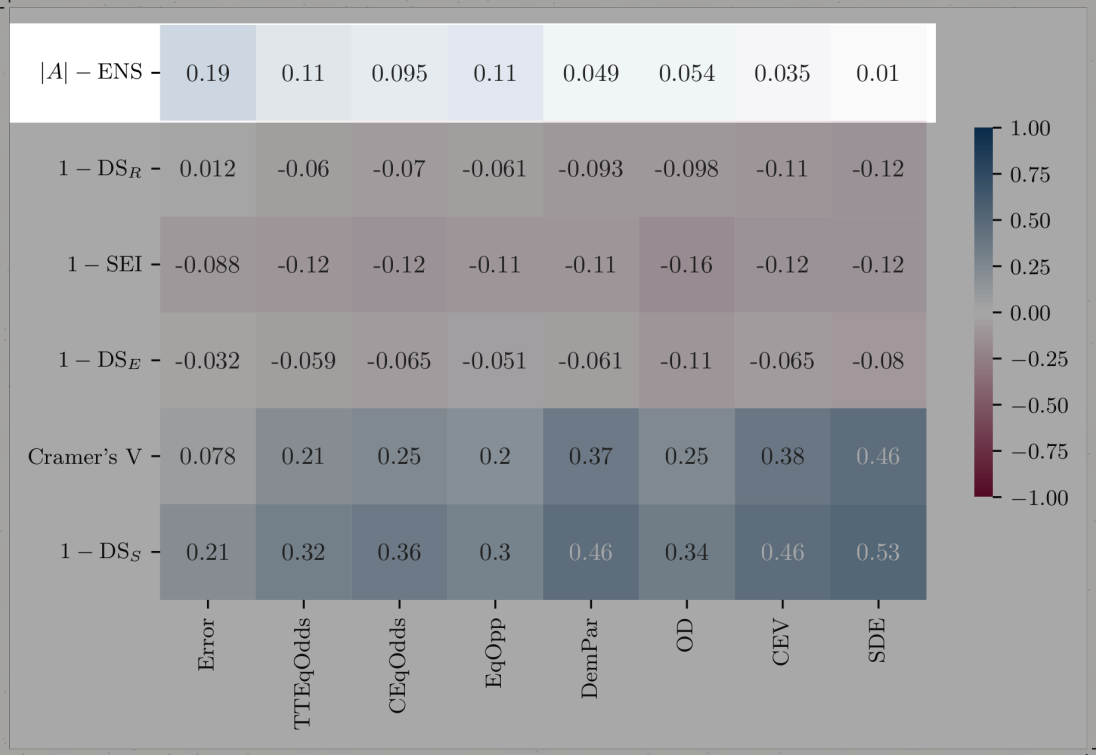


Spearman rank correlation between dataset bias (rows) and model bias metrics

- Stereotypical dataset bias ↔ strongly visible in equalized-odds metrics
- Representational bias probably has a smaller impact on the model
- But if there is any, current model bias metrics do not pick it up [8]

[8] I. Dominguez-Catena et al. Biased heritage: how datasets shape models in facial expression recognition. *IEEE Trans. Affective Computing* 2026

Which model metric sees which dataset bias?



Spearman rank correlation between dataset bias (rows) and model bias metrics

- Stereotypical dataset bias ↔ strongly visible in equalized-odds metrics
- Representational bias probably has a smaller impact on the model
- But if there is any, current model bias metrics do not pick it up [8]

[8] I. Dominguez-Catena et al. Biased heritage: how datasets shape models in facial expression recognition. *IEEE Trans. Affective Computing* 2026

Output: two papers



Less can be more: representational vs. stereotypical gender bias in facial expression recognition

Iris Dominguez-Catena*, Daniel Paternain, Aranzazu Jurio, Mikel Galar

Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Arrosadia Campus, Pamplona, 31006, Navarre, Spain.

*Corresponding author(s). E-mail(s): iris.dominguez@unavarra.es;
Contributing authors: daniel.paternain@unavarra.es; aranzazu.jurio@unavarra.es; mikel.galar@unavarra.es;

Abstract

Machine learning models can inherit biases from their training data, leading to discriminatory or inaccurate predictions. This is particularly concerning with the increasing use of large, unsupervised datasets for training foundational models. Traditionally, demographic biases within these datasets have not been well-understood, limiting our ability to understand how they propagate to the models themselves. To address this issue, this paper investigates the propagation of demographic biases from datasets into machine learning models. We focus on the gender demographic component, analyzing two types of bias: representational and stereotypical. For our analysis, we consider the domain of facial expression recognition (FER), a field known to exhibit biases in most popular datasets. We use AffectNet, one of the largest FER datasets, as our baseline for carefully designing and generating subsets that incorporate varying strengths of both representational and stereotypical bias. Subsequently, we train several models on these biased subsets, evaluating their performance on a common test set to assess the propagation of bias into the models' predictions. Our results show that representational bias has a weaker impact than expected. Models exhibit a good generalization ability even in the absence of one gender in the training dataset. Conversely, stereotypical bias has a significantly stronger impact, primarily concentrated on the biased class, although it can also influence predictions for unbiased classes. These results highlight the need for a bias analysis that differentiates between types of bias, which is crucial for the development of effective bias mitigation strategies.

406.17405v1 [cs.CV] 25 Jun 2024

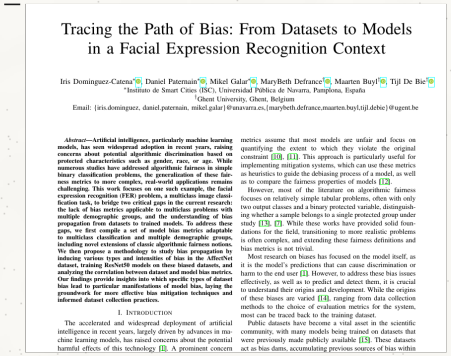
Less can be more: representational vs. stereotypical gender bias in facial expression recognition

I. Dominguez-Catena, D. Paternain, A. Jurio, M. Galar

Progress in Artificial Intelligence, 2024

doi.org/10.1007/s13748-024-00345-w

Paper



Tracing the Path of Bias: From Datasets to Models in a Facial Expression Recognition Context

Iris Dominguez-Catena*, Daniel Paternain*, Mikel Galar*, MaryBeth Defrance, Maarten Buyl, Tijl De Bie

*Instituto de Smart Cities (ISC), Universidad Pública de Navarra, Pamplona, España

Ghent University, Ghent, Belgium

Email: {iris.dominguez, daniel.paternain, mikel.galar}@unavarra.es,{marybeth.defrance,maarten.buyl,tijl.debie}@ugent.be

Abstract—**Artificial intelligence, particularly machine learning models, has seen widespread adoption in recent years, raising concerns about potential algorithmic discrimination based on protected characteristics such as gender, race, or age. While numerous studies have addressed algorithmic fairness in simple binary classification problems, the generalization of these fairness metrics to more complex, real-world applications remains challenging. This work focuses on one such example, the facial expression recognition (FER) problem, a multiclass image classification task, to bridge two critical gaps in the current research: the lack of bias metrics applicable to multiclass problems with multiple demographic groups, and the understanding of bias propagation from datasets to trained models. To address these gaps, we first compile a set of model bias metrics adaptable to multiclass classification and multiple demographic groups, including novel extensions of classic algorithmic fairness notions. We then propose a methodology to study bias propagation by inducing various types and intensities of bias in the AffectNet dataset, training ResNet50 models on these biased datasets, and analyzing the correlation between dataset and model bias metrics. Our findings provide insights into which specific types of dataset bias lead to particular manifestations of model bias, laying the groundwork for more effective bias mitigation techniques and informed dataset collection practices.**

I. Introduction

The accelerated and widespread deployment of artificial intelligence in recent years, largely driven by advances in machine learning models, has raised concerns about the potential harmful effects of this technology [1]. A prominent concern metrics assume that most models are unfair and focus on quantifying the extent to which they violate the original constraint [10], [11]. This approach is particularly useful for implementing mitigation systems, which can use these metrics as heuristics to guide the debiasing process of a model, as well as to compare the fairness properties of models [12].

However, most of the literature on algorithmic fairness focuses on relatively simple tabular problems, often with only two output classes and a binary protected variable, distinguishing whether a sample belongs to a single protected group under study [13], [7]. While these works have provided solid foundations for the field, transitioning to more realistic problems is often complex, and extending these fairness definitions and bias metrics is not trivial.

Most research on biases has focused on the model itself, as it is the model's predictions that can cause discrimination or harm to the end user [1]. However, to address these bias issues effectively, as well as to predict and detect them, it is crucial to understand their origins and development. While the origins of these biases are varied [14], ranging from data collection methods to the choice of evaluation metrics for the system, most can be traced back to the training dataset.

Public datasets have become a vital asset in the scientific community, with many models being trained on datasets that were previously made publicly available [15]. These datasets act as bias dams, accumulating previous sources of bias within

Biased Heritage: how datasets shape models in facial expression recognition

I. Dominguez-Catena, D. Paternain, M. Galar, M. B. Defrance, M. Buyl, T. De Bie

IEEE Transactions on Affective Computing, 2026

doi.org/10.1109/TAFFC.2026.3693403

Paper

Thesis takeaways

Dataset bias has (at least) **two faces**, and they behave differently

We can **measure both**, even without demographic labels

Stereotypical bias is the one that hurts models most (at least in FER)

But all of it relies on an auxiliary demographic model. How good is it?

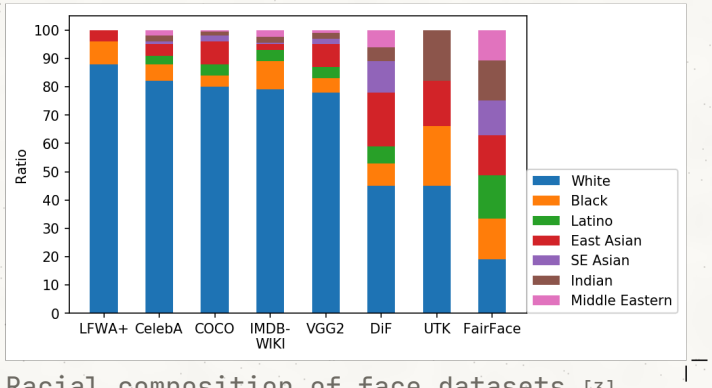
05

Better profiling through quantification

With P. González, J. J. del Coz, D. Paternain and M. Galar

FairFace: the dataset and the model

The dataset



Racial composition of face datasets [3]

- 108,501 face images from YFCC-100M (Flickr)
- Built to be **balanced across race groups**
- **Perceived** race, gender and age from external annotators

The trained model

- CNN trained on that dataset, used as our **frozen auxiliary model**

Predicts per face

`RACE · 7` White, Black, Latino/Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern

`GENDER · 2` male, female

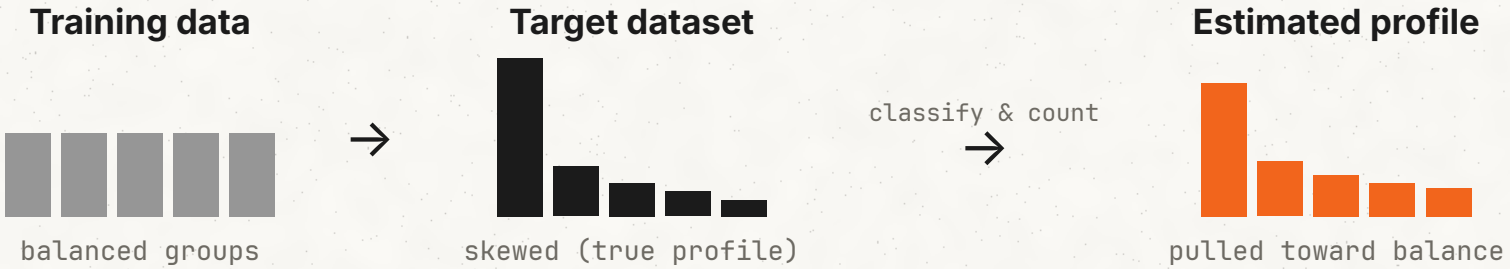
`AGE · 9` 0–2, 3–9, 10–19, …, 60–69, 70+

- Outputs **probabilities**, which is what quantification works on

[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021

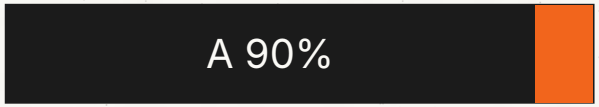
The weak link: auxiliary demographic models

- Audits of large datasets rely on classifiers like FairFace
- Ethical caveat: binary gender, 7 census-style race groups, **perceived** attributes
- Fine per image, but we care about the **aggregate** profile
- Under distribution shift, classify-and-count is biased toward the training prior



The intuition: correcting a 90/10 count

What the classifier says

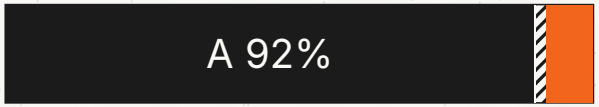


Its confusion matrix (balanced)

	pred. A	pred. B
true A	97.5%	2.5%
true B	2.5%	97.5%

same error rate for both groups

Corrected estimate



There are 9× more A faces to get wrong, so about 90% of the mistakes are **A→B** and only 10% are B→A.

- The B count is **inflated** by A faces mislabeled as B → the real share of B is smaller, **≈ 92/8**
- Quantification does this systematically, from the classifier's confusion or posteriors

Toy numbers: true 92/8, error rate 2.5% per group → observed ≈ 90/10.

Quantification in one slide

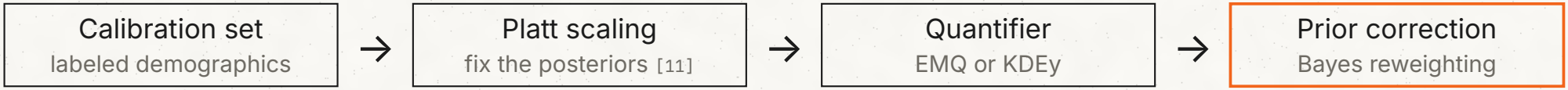
- Goal: estimate class **prevalences** π_y in a bag, not individual labels
- **EMQ** [9]: EM on the posteriors, re-estimating priors
- **KDEy** [10]: match KDE of posteriors, robust without calibration
- Then push the corrected priors back to each sample:

$$\hat{p}_{\text{test}}(y|x) \propto \frac{\hat{\pi}_y}{P_{\text{tr}}(y)} p_{\text{tr}}(y|x)$$

[9] M. Saerens, P. Latinne, C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities. *Neural Computation* 2002

[10] A. Moreo, P. González, J. J. del Coz. Kernel density estimation for multiclass quantification. *Machine Learning* 2025

Setup



The classifier (FairFace) stays frozen. The correction is fitted once and applied to each target dataset.

Data

Six face datasets: FairFace, RAF-DB, ADFES, KDEF, FACES, CAFE

Gender (binary), race (nominal), age (ordinal)

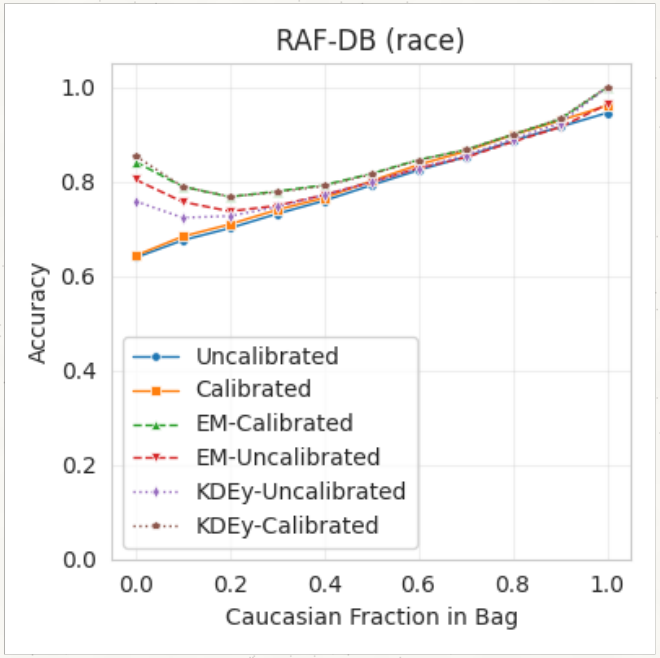
Three calibration scenarios

On FairFace · on RAF-DB · on a labeled 60% of the target

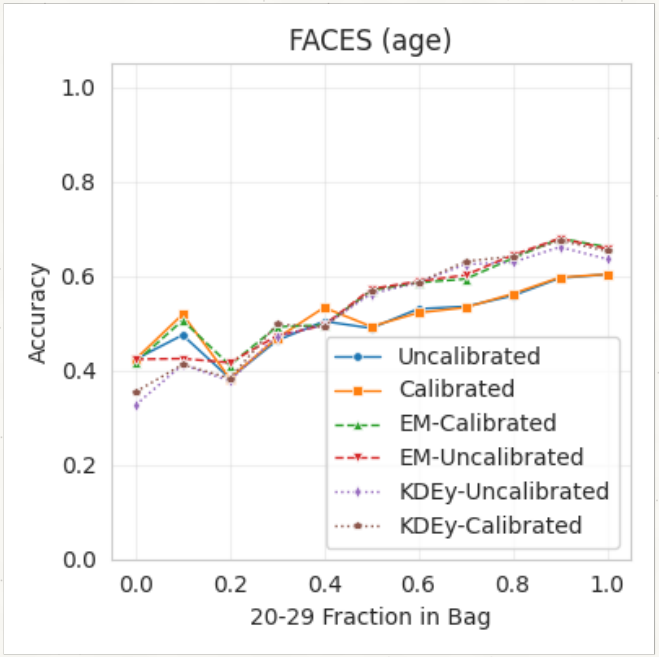
Artificial bags sweep one group's prevalence from 0 to 1

[11] J. Platt. Probabilities for SV machines. *Advances in Large-Margin Classifiers* 2000

Per-sample accuracy vs. bag composition



RAF-DB, race (calibrated on RAF-DB)



FACES, age (calibrated on FairFace)

- Largest gains where the bag is **far** from the calibration prior
- RAF-DB race: 0.65 → 0.69 avg. accuracy, up to +0.2 on skewed bags
- Gains on nearly all datasets and attributes

Which method, when?

Labeled sample of the target
Platt + EMQ best by far

External calibration set
depends on how close it is to the target

KDEy without calibration
never the best, always second: a safe default (but slower)

By attribute
binary gender: small consistent gains; age and race: larger gains

Better Demographic Profiling of Datasets through Quantification, ECAF 2026 poster

06

Current focus: LAION-5B

Unmasking LAION-5B · ICLR 2026 DATA-FM workshop · with D. Paternain and M. Galar

Why web-scale datasets?

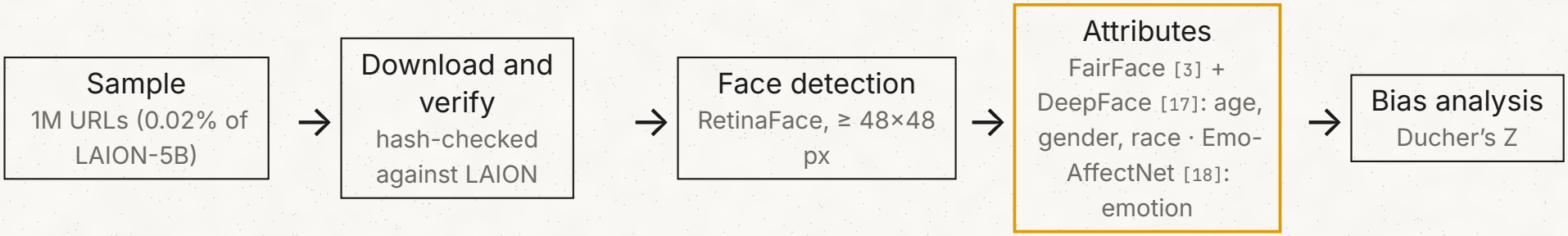


Faces detected in our LAION sample

- Foundation and generative models (Stable Diffusion, DALL-E, FLUX) are trained on web-scraped image–text data such as **LAION-5B** [15]
- Billions of pairs, scraped and barely curated
- Instead of auditing the outputs of trained models, we audit the **underlying demographic composition** of the data [16]

[15] C. Schuhmann et al. LAION-5B: an open large-scale dataset for training next generation image-text models. *NeurIPS Datasets and Benchmarks* 2022 [16] I. Dominguez-Catena, D. Paternain, M. Galar. Facial demography analysis of the LAION dataset. *EWAF* 2025

Profiling pipeline at scale



1,004,277
URLs sampled

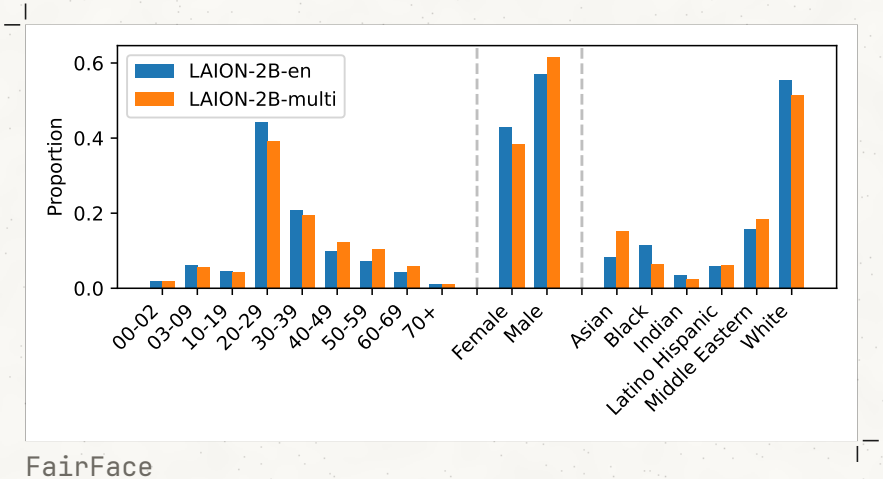
464,161
images downloaded and verified

79,902
faces (37,331 English · 42,571 multilingual)

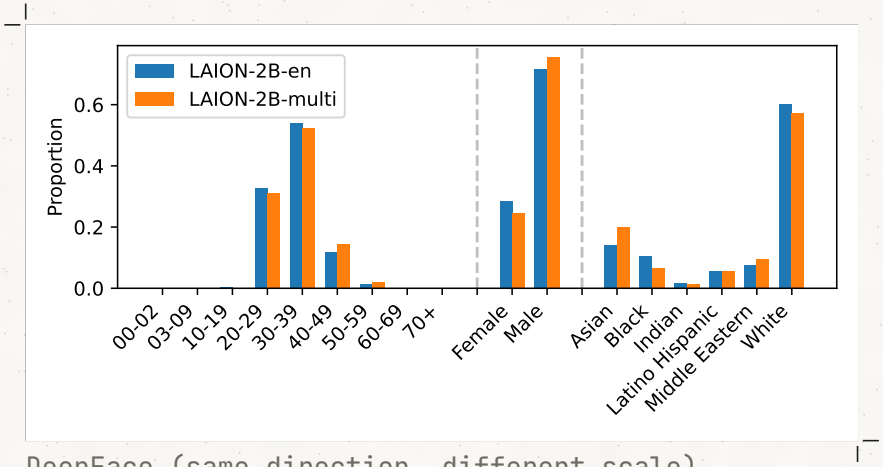
- 2024 re-release, both main components analysed separately: **LAION-2B-en** and **LAION-2B-multi**
- Worst-case margin of error: ±0.51% at 95% confidence

[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021 [16] I. Dominguez-Catena, D. Paternain, M. Galar. Facial demography analysis of the LAION dataset. *EWAF* 2025 [17] S. Serengil, A. Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *J. Information Technologies* 2024 [18] E. Ryumina et al. In search of a robust facial expressions recognition model: a large-scale visual cross-corpus study. *Neurocomputing* 2022

Who is in LAION?



FairFace



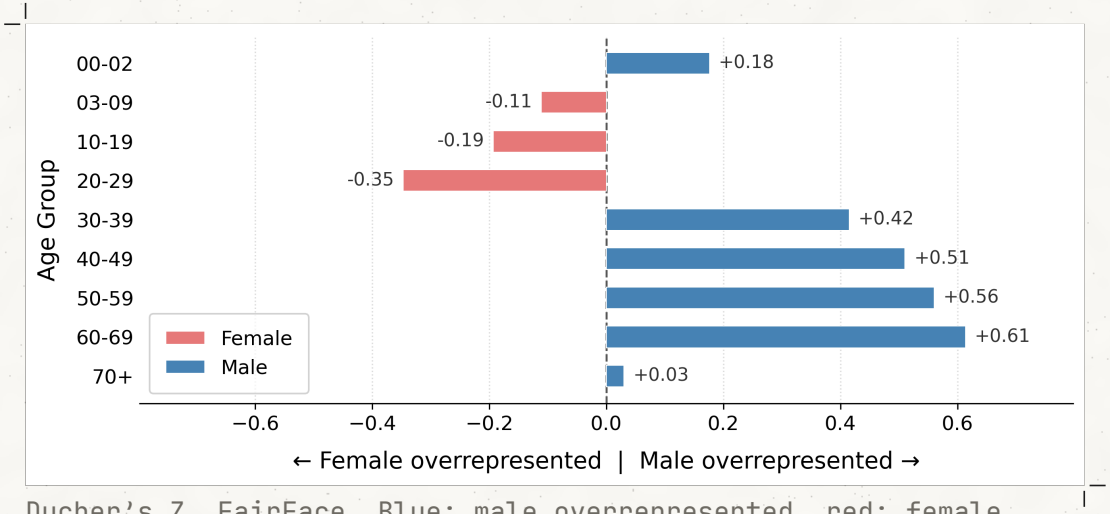
DeepFace (same direction, different scale)

Age
strong overrepresentation of **20–39**

Gender
57–61% male (FairFace), more than 70% (DeepFace)

Race
White 50–60%, against a balanced 16% baseline

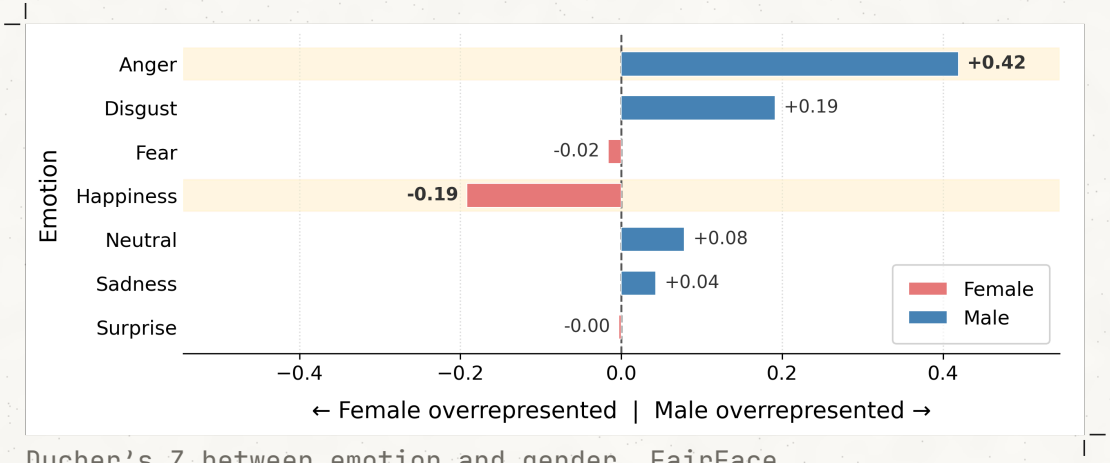
Intersectional bias: gender × age



Ducher's Z, FairFace. Blue: male overrepresented, red: female overrepresented

- **Younger women, older men**: women overrepresented below 30 (Z = 0.35 at 20–29), men above 30 (Z = 0.61 at 60–69)
- DeepFace shows the same pattern, with the threshold at 40
- Age × race: oldest groups underrepresented; White infants overrepresented

Stereotypes in LAION



Ducher's Z between emotion and gender, FairFace

- The **"angry man, happy woman"** stereotype is in the data
- Anger: Z = 0.42 for men (95% CI 0.40–0.43). Happiness: Z = 0.19 for women (0.18–0.20)
- Weaker age and race effects: older people linked to negative emotions
- The multilingual component is **not** more diverse: a bit more racial and age diversity, but **more** gender disparity

What comes next

- **Trace** how these biases propagate through generative pipelines (Stable Diffusion, FLUX)
- **Validate** with human annotations: the auxiliary models have their own biases
- Extend the audit to other web-scale datasets (COYO-700M, DataComp)
- Link to the earlier parts: correct the profiles with **quantification**, and compare datasets with **DSAP**

Conclusions

01 Dataset bias is measurable, even without labels (taxonomy, metrics, DSAP)

02 Different bias types transfer differently; stereotypes matter most

03 Demographic estimates for auditing can be improved with **quantification**

04 Applied to LAION-5B: strong imbalances and an “angry man, happy woman” stereotype

07

Beyond images

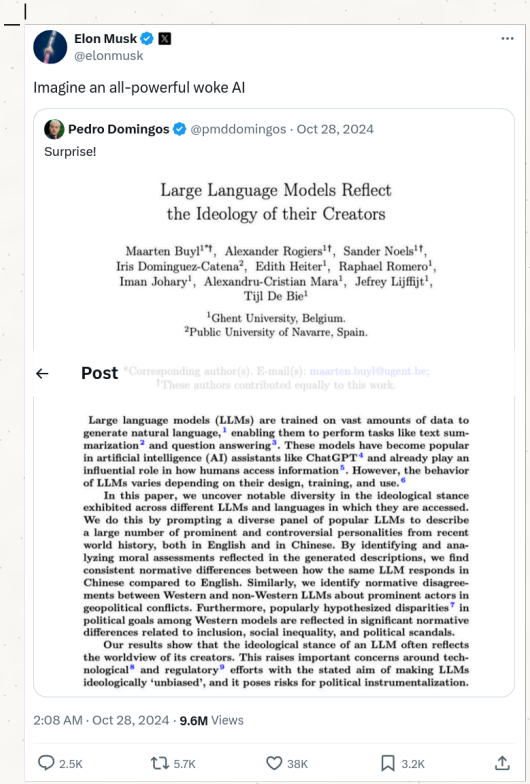
Same question, different modality

Large language models reflect the ideology of their creators

- With Ghent University (Buyl, De Bie et al.) [19]
- Asked LLMs to describe thousands of political figures, in English and Chinese
- Ideological stance differs by language and by the model's creator
- Same question as before: **who and what is in the data?**

UNIVERSITEIT GENT

Paper



[19] M. Buyl et al. Large language models reflect the ideology of their creators. *npj Artificial Intelligence* 2026

Thank you! Questions?

Thank you!

iris.dominguez@unavarra.es
irisdominguez.es

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

arin
Artificial Intelligence &
Machine Learning Research



Website and publications

08

Backup · Cross-modal bias reduction

Dominguez-Catena, Paternain, Jurio & Galar. IEEE CAI 2026

Complementary failure modes

FairFace
misgenders women showing anger or fear [3]

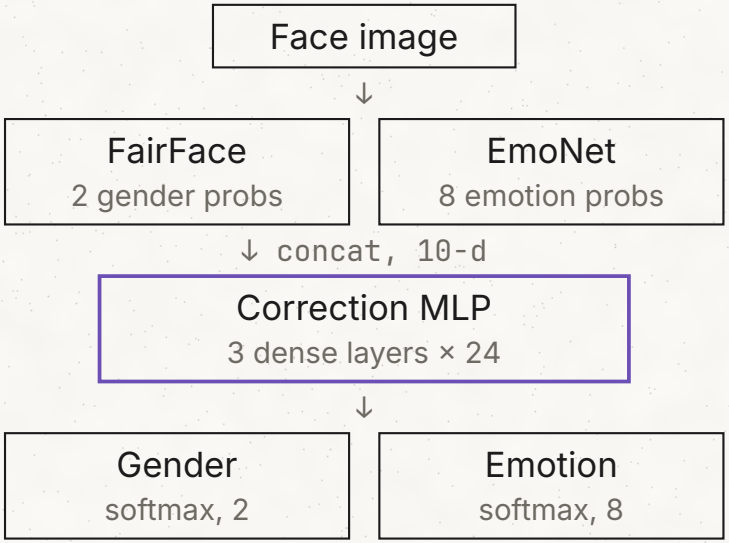
EmoNet
emotion accuracy depends on gender [13]

- Idea: let each model **see the other's prediction**
- "Fairness through awareness" [12]: no explicit fairness constraint

[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021 [12] C. Dwork et al. Fairness through awareness. *ITCS* 2012 [13] A. Toisoul et al. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 2021

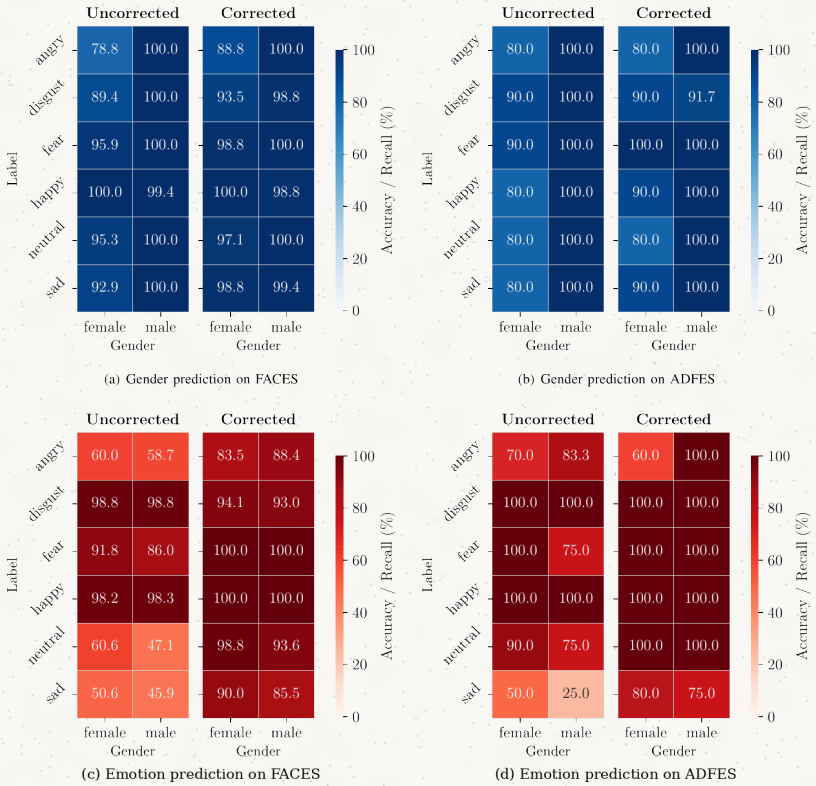
A tiny correction network

- Plain cross-entropy on both tasks [14]
- Base models (FairFace, EmoNet) are **not** retrained
- Train on FACES, test on the unseen ADFES



[3] K. Kärkkäinen, J. Joo. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *WACV* 2021 [13] A. Toisoul et al. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* 2021 [14] I. Dominguez-Catena et al. Leveraging cross-modal information to reduce gender bias in facial analysis systems. *IEEE CAI* 2026

Results



- FACES gender: 95.98 → 97.85 % macro acc., avg gap 8.0 → 3.7 [14]
- FACES emotion: 74.6 → 93.9 %
- ADFES (unseen) emotion: 80.7 → 92.9 %, avg gap 13.1 → 7.5

> Accuracy and fairness improve **together**

[14] I. Dominguez-Catena et al. Leveraging cross-modal information to reduce gender bias in facial analysis systems. *IEEE CAI* 2026

Limitations and next steps

- Small lab datasets, binary gender
- Largest gap on ADFES can still grow (few subjects)
- Next: more attributes (age, race), in-the-wild data