METRICS FOR DATASET DEMOGRAPHIC BIAS IN MACHINE LEARNING: A CASE STUDY ON FACIAL EXPRESSION RECOGNITION

Iris Dominguez-Catena*, Mikel Galar, Daniel Paternáin

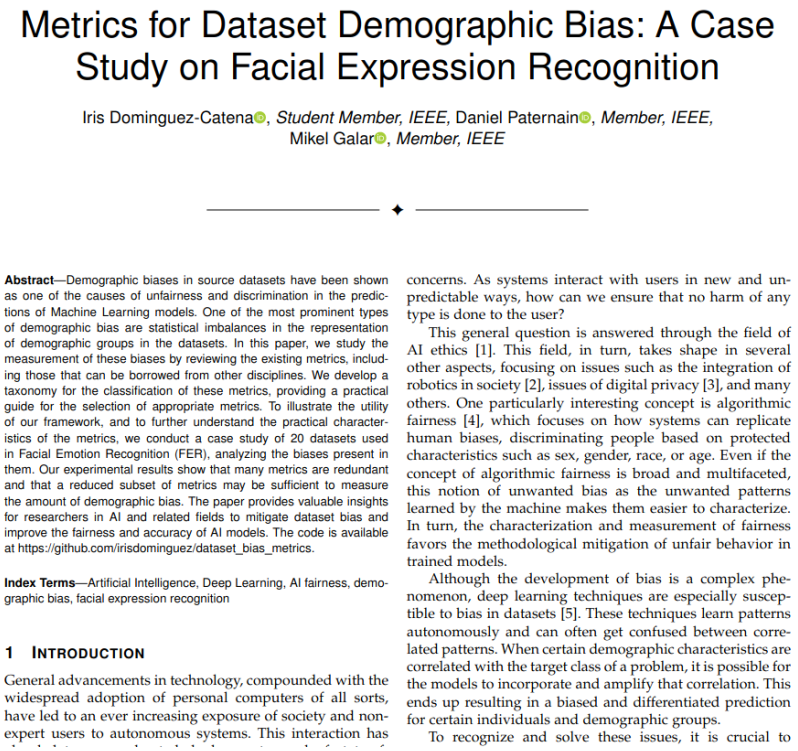
Enero 2024

Instituto de Smart Cities (ISC), Departmento de Estadística, Informática y Matemáticas
Universidad Pública de Navarra (UPNA)

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

*iris.dominguez@unavarra.es, irisai.neocities.org

THE ARTICLE



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

Published in the

IEEE Transactions on Pattern Analysis and Machine Intelligence

(Q1, Impact Factor: 23.6)

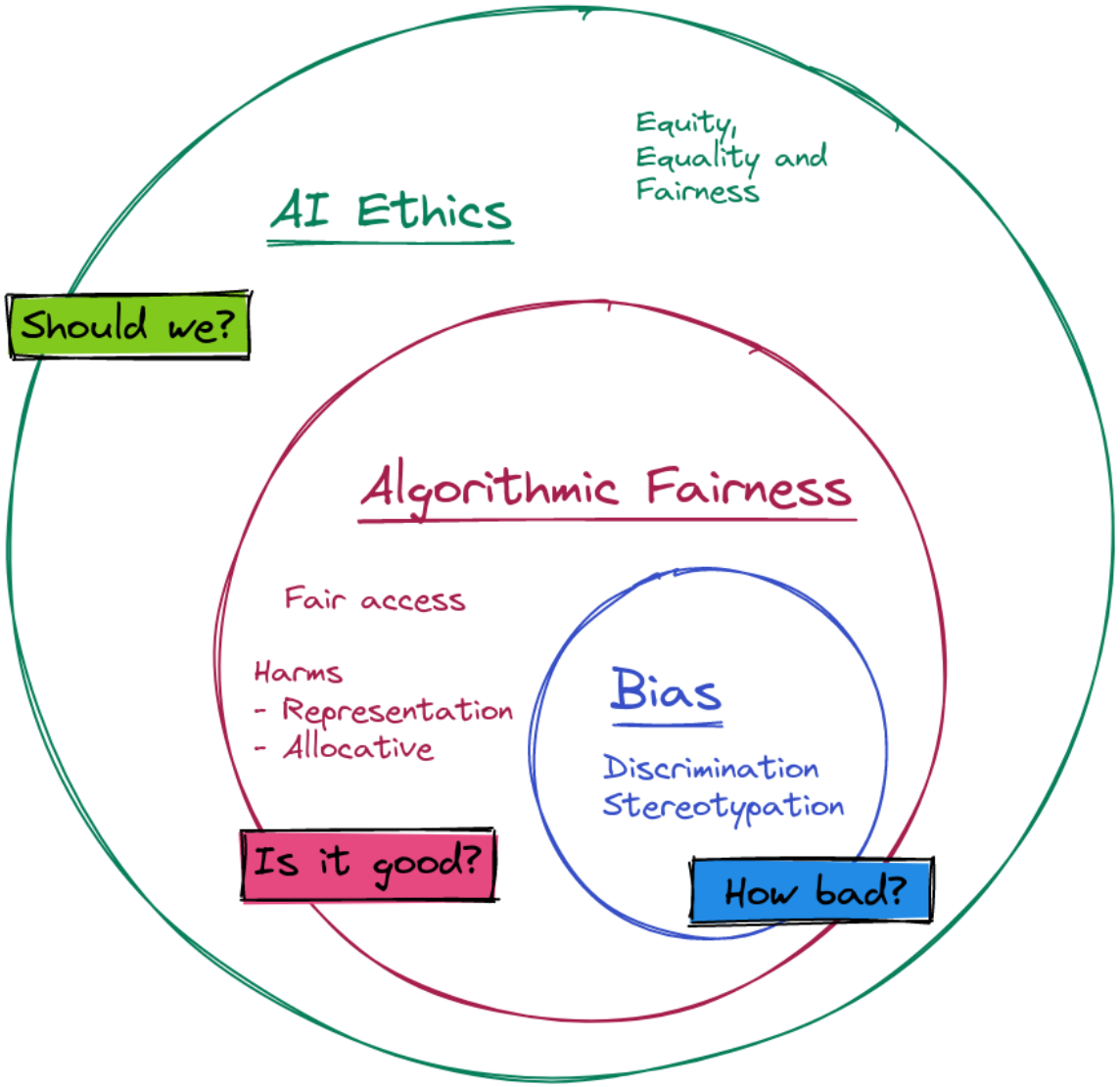
https://doi.org/10.1109/TPAMI.2024.3361979

TABLE OF CONTENTS

- Intro to algorithmic bias
- Use case: Facial Expression Recognition
- Types of bias and bias metrics
- Bias transference from dataset to model
- Other real cases

INTRO TO AI BIAS

FIELD MAP FOR AI ETHICS



BIASES

- **Unwanted** patterns
  - Both in *data* and *model predictions*
- Based on **protected attributes**
  - Gender, race, age
  - Inherent and immutable
- **Quantifiable**
  - *Group bias* metrics

BIAS SOURCES

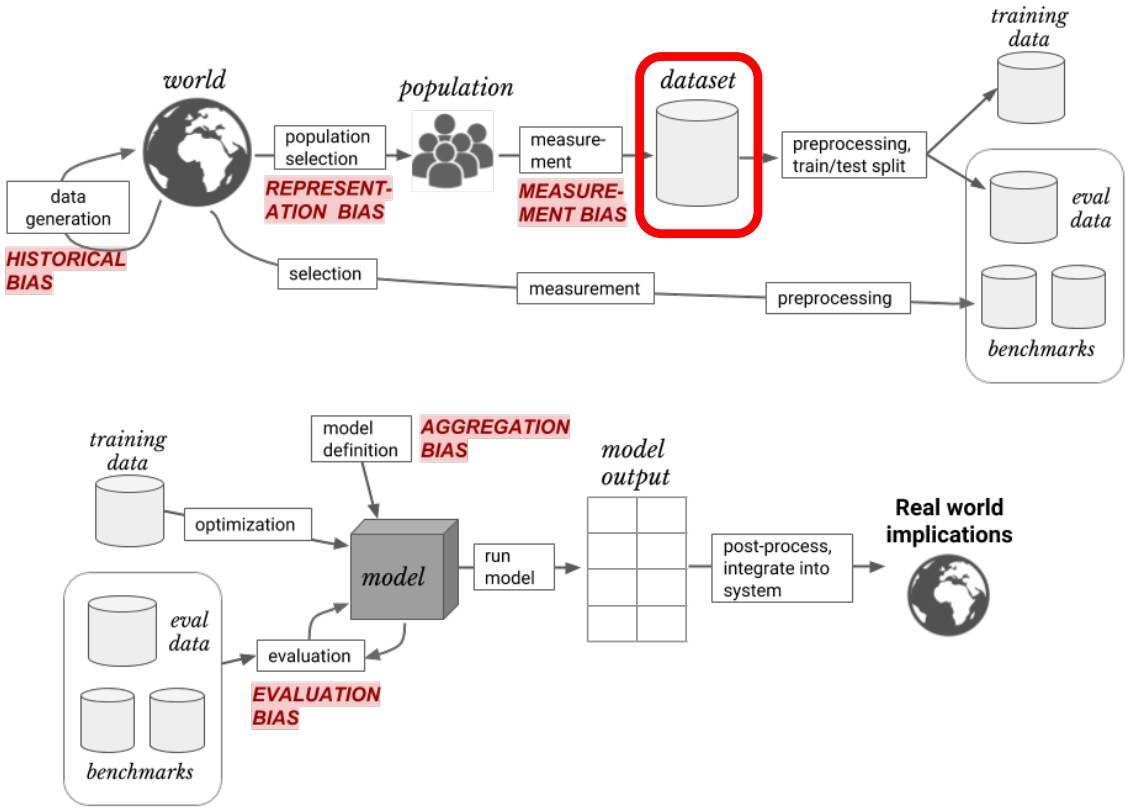


Figure 1: Bias source in the machine learning pipeline[1]

[1] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization. – NY USA: ACM, Oct. 5, 2021, pp. 1–9. ISBN: 978-1-4503-8553-4. DOI: 10.1145/3465416.3483305.

FACIAL EXPRESSION RECOGNITION

WHAT IS FACIAL EXPRESSION RECOGNITION?

Figure 2: A sample of FER2013/FER+, a popular FER dataset[2].

[2] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ICMI '16: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION. Tokyo Japan: ACM, Oct. 31, 2016, pp. 279–283. ISBN:

FER AND FER-RELATED KNOWN BIASES

- Gender and skin tone (Fitzpatrick Skin Type) in gender classification[3]
- FER research models[4]: capacitism, age, race and gender
- Commercial FER systems[5]: age, race and gender

[3] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, Feb. 23–24, 2018, pp. 77–91.

[4] Jacqueline J. Greene et al. "The Spectrum of Facial Palsy: The MEEI Facial Palsy Photo and Video Standard Set". In: *The Laryngoscope* 130.1 (2020), pp. 32–37. ISSN: 1531-4995. DOI: 10.1002/lary.27986; Tian Xu et al. "Investigating Bias and Fairness in Facial Expression Recognition". In: *Computer Vision – ECCV 2020 Workshops*. Ed. by Adrien Bartoli and Andrea Fusiello. Cham: Springer International Publishing, 2020, pp. 506–523. ISBN: 978-3-030-65414-6. DOI: 10.1007/978-3-030-65414-6_35.

[5] Eugenia Kim et al. "Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, July 21, 2021, pp. 638–644. ISBN: 978-1-4503-8473-5; Khurshid Ahmad et al. "Comparing the Performance of Facial Emotion Recognition Systems on Real-Life Videos: Gender, Ethnicity and Age". In: *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1*. Ed. by Kohei Arai. Vol. 358. Cham: Springer International Publishing, 2022, pp. 193–210. ISBN: 978-3-030-89905-9 978-3-030-89906-6. DOI: 10.1007/978-3-030-89906-6_14.

FER DATASETS

Short name	Year	Collection	Images	Videos	Subjects
POFA	1976	Lab	110	-	16
JACFEE	1988	Lab	56	-	56
AR-Face	1998	Lab	4000	-	126
JAFFE	1998	Lab	213	-	10
KDEF	1998	Lab	4900	-	70
CK	2000	Lab	8795	486	97
CK+	2010	Lab	10727	593	123
MUG	2010	Lab	70654	-	52
Multi-PIE	2010	Lab	750000	-	337
RaFD	2010	Lab	8040	-	67
SFEW	2011	ITW	1766	-	330
FER2013	**2013**	**ITW**	**32298**	-	-
WSEFEP	2014	Lab	210	-	30
ADFES	2016	Lab	-	648	22
FERPlus	**2016**	**ITW**	**32298**	-	-
Aff-Wild2	2017	ITW	-	558	-
AffectNet	**2017**	**ITW**	**291652**	-	-
ExpW	2017	ITW	91793	-	-
RAF-DB	2017	ITW	29672	-	-
CAER-S	2019	ITW	70000	-	
SEWA	2019	ITW	-	199	398
MMAFEDB	2020	ITW	128000	-	-
NHFIER	2020	ITW	5558	-	-

Representational bias

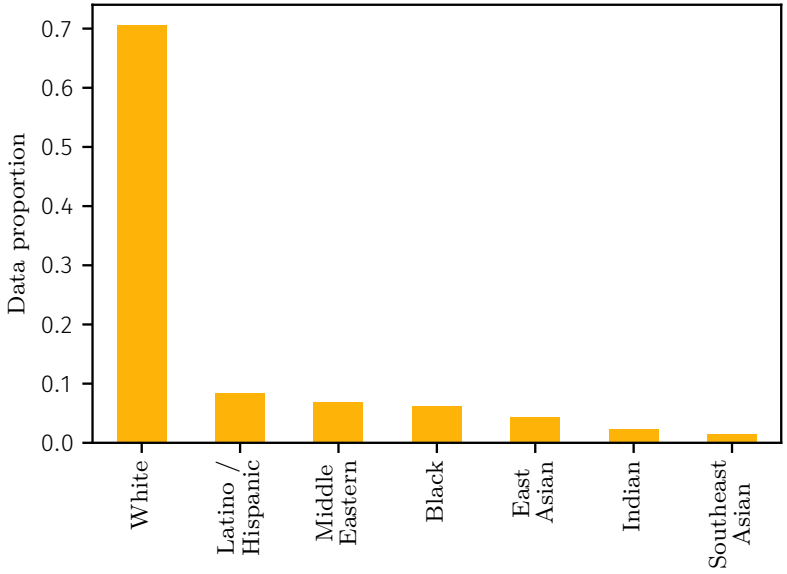


Figure 3: Apparent race distribution in FER+.

[5] (Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. Thirty-First International Joint Conference on Artificial Intelligence and the Twenty-Fifth European Conference on Artificial Intelligence (IJCAI-ECAI-2022). Vienna, Austria, July 24–25, 2022)

Representational bias

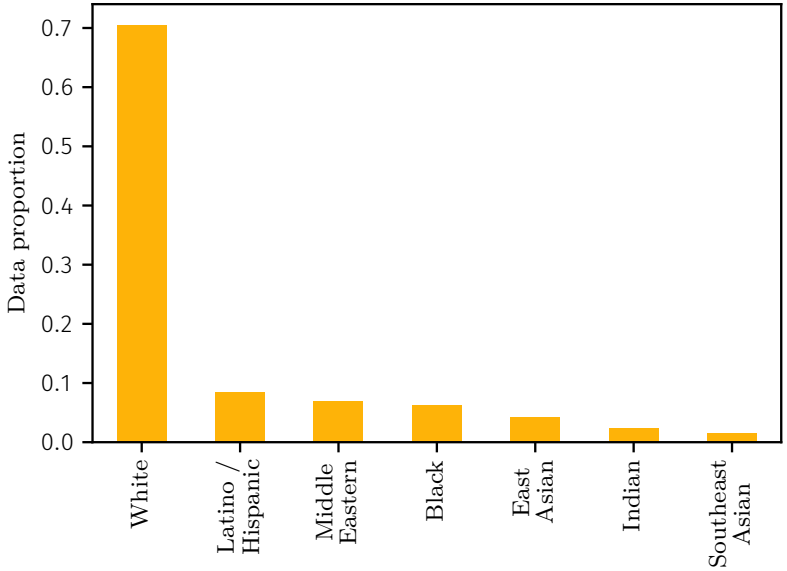


Figure 3: Apparent race distribution in FER+.

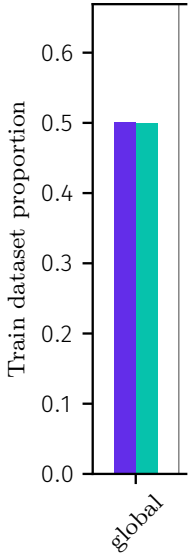


Figure 4: Apparent *per-label* gender distribution in FER+.

[5] (Dominguez-Catena, Paternain, and Galar, "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition")

Representational bias

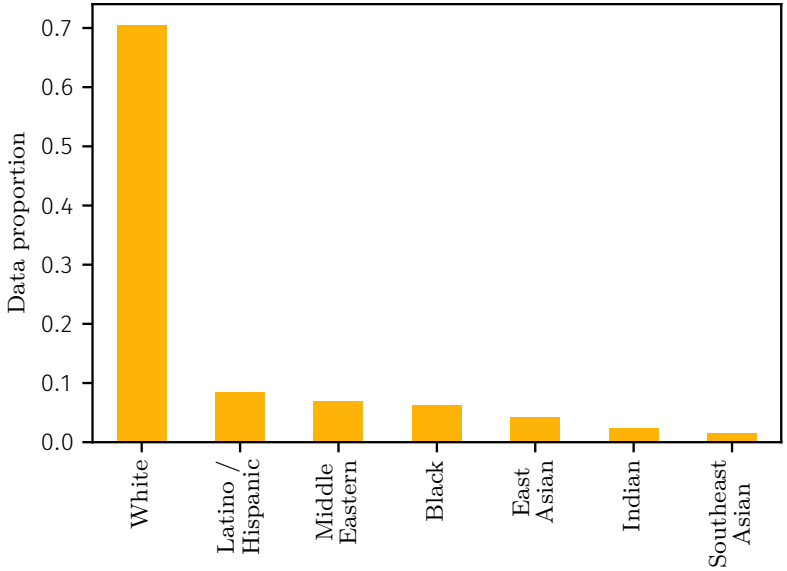


Figure 3: Apparent race distribution in FER+.

Stereotypical bias

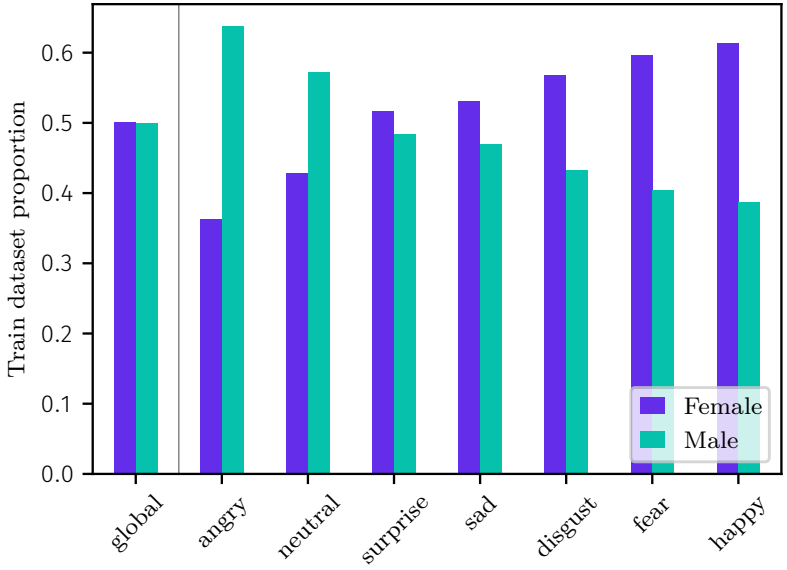


Figure 4: Apparent *per-label* gender distribution in FER+.

[5] (Dominguez-Catena, Paternain, and Galar, "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition")

BIAS METRICS

- Representational
  - Richness
    - Richness
  - Evenness
    - Shannon Evenness Index
    - Normalized Standard Deviation
  - Dominance
    - Imbalance Ratio
    - Berger-Parker Index
  - Combined
    - Effective Number of Species
    - Simpson Index
    - Simpson's Reciprocal
    - Simpson's Index of Diversity
    - Shannon Entropy
- Stereotypical
  - Global
    - Cramer's V
    - Tschuprow's T
    - Pearson's Contingency Coefficient
    - Theil's U
    - Normalized Mutual Information
  - Local
    - Normalized Pointwise Mutual Information
    - Ducher's Z

[5] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition*. Mar. 28, 2023. DOI: 10.48550/arXiv.2303.15889. arXiv: 2303.15889 [cs]. URL: http://arxiv.org/abs/2303.15889 (visited on 05/26/2023). preprint

1. Dataset preprocessing, homogenize images and labels
2. Demographic analysis of the datasets
   - FairFace[6]
3. Measure bias with all metrics
4. Discard *redundant* metrics, prioritizing *interpretable* metrics

[6] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1547–1557. ISBN: 978-1-66540-477-8. DOI: 10.1109/WACV48630.2021.00159.

REPRESENTATIONAL BIAS

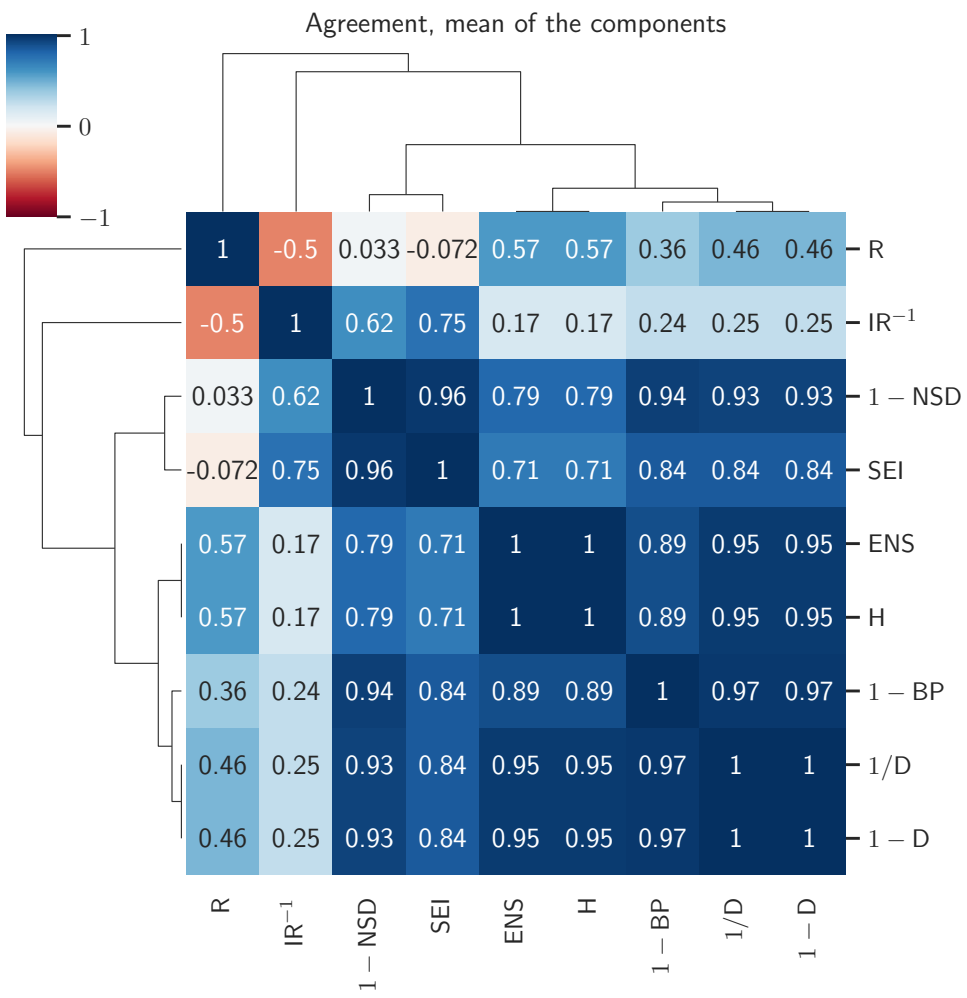


Figure 5: Spearman's ρ correlation between representational bias metrics

[6] Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

BEST CHOICES

- General representational
  - **Effective Number of Species** (ENS)
- Evenness between represented groups
  - **Shannon Evenness Index** (SEI)
- Good approximation: Dominance
  - **Berger-Parker Index** (BP)

STEREOTYPICAL BIAS

STEREOTYPICAL BIAS, AN EXAMPLE I

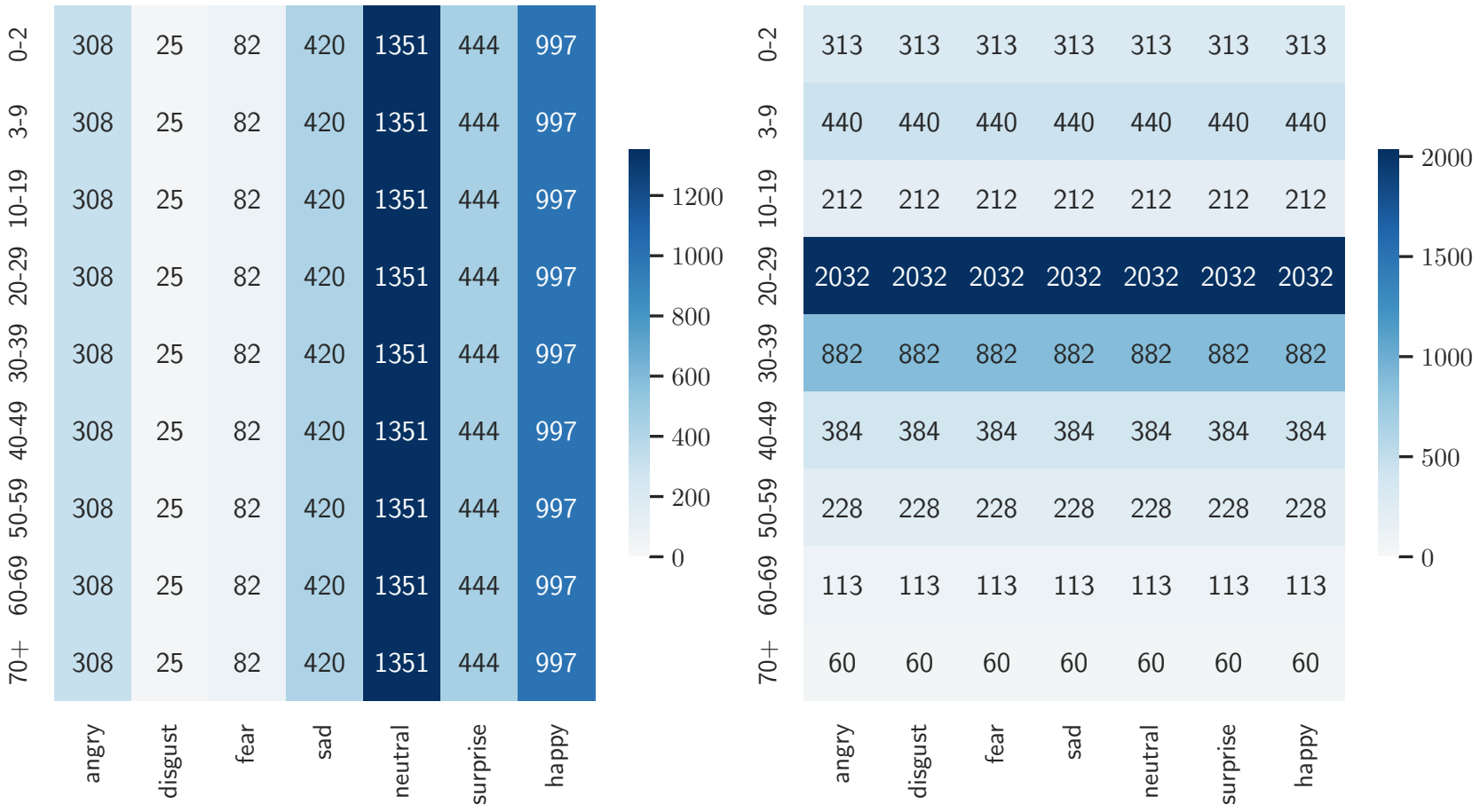


Figure 6: Contingency tables of two datasets **without** stereotypical bias

STEREOTYPICAL BIAS, AN EXAMPLE II

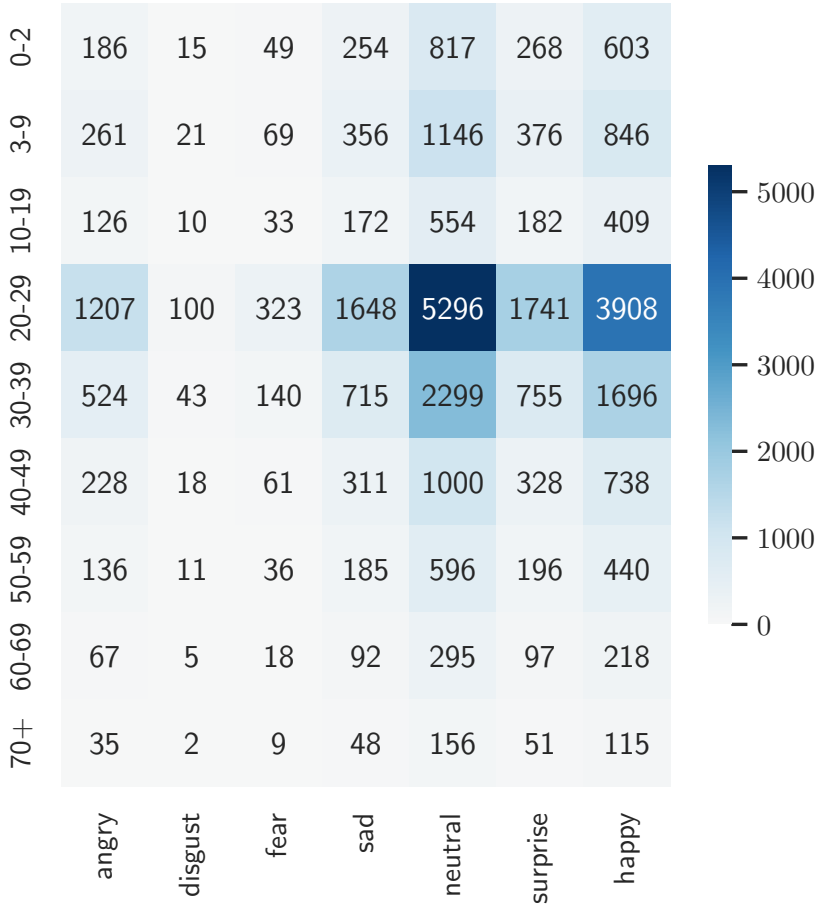


Figure 7: Contingency tables of a datasets **without** stereotypical bias

STEREOTYPICAL BIAS, AN EXAMPLE III

	angry	disgust	fear	sad	neutral	surprise	happy
0-2	134	12	24	529	547	381	569
3-9	248	16	30	588	842	445	911
10-19	77	3	15	174	728	171	322
20-29	869	72	307	1209	5590	2186	3993
30-39	863	79	239	655	2027	518	1795
40-49	355	25	74	351	1027	161	695
50-59	163	16	45	149	742	62	425
60-69	59	7	7	87	417	38	179
70+	6	0	1	44	245	37	88

5000

0

Figure 8: Real contingency table of a FER, **with** stereotypical bias

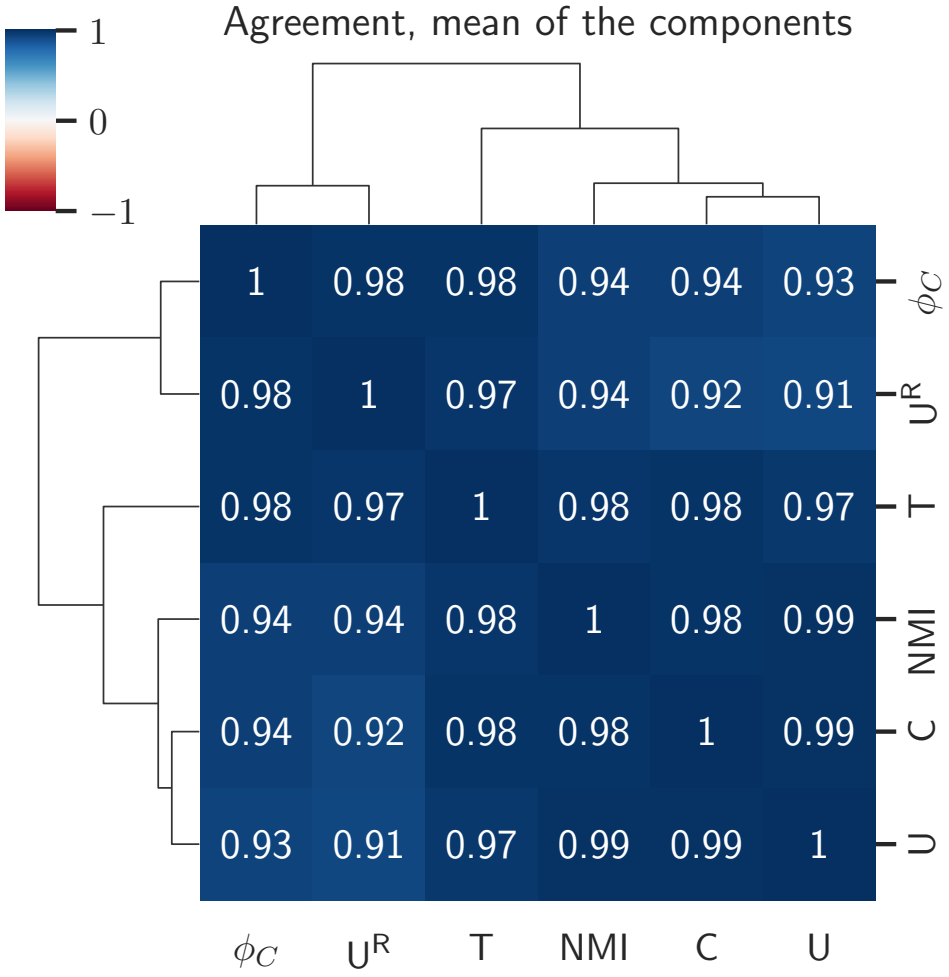


Figure 9: Spearman's ρ correlation between stereotypical bias metrics

[6] Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

BEST CHOICE

- Stereotypical bias (global)
  - **Cramer's V** (ϕ_C)
- Stereotypical bias (local)
  - **Ducher's Z** (Z)

METRICS IN ACTION

Laboratory 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

ITW-I datasets

		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

Figure 10: Representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of lab (top) and ITW-I (bottom) datasets.

[6] Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

		Laboratory	ITW-I
		Average	Average
Age	$9-$ ENS	7.067 ± 0.932	3.334 ± 0.286
	$1-$ SEI	0.409 ± 0.200	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.104 ± 0.027
Race	$7-$ ENS	5.168 ± 0.634	3.724 ± 0.321
	$1-$ SEI	0.384 ± 0.151	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.063 ± 0.018
Gender	$2-$ ENS	0.139 ± 0.280	0.005 ± 0.005
	$1-$ SEI	0.039 ± 0.052	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.167 ± 0.018

Figure 11: Representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of lab (left) and ITW-I (right) datasets.

[6] Dominguez-Catena, Paternain, and Galar, *Metrics for Dataset Demographic Bias*

LOCAL STEREOTYPICAL BIAS

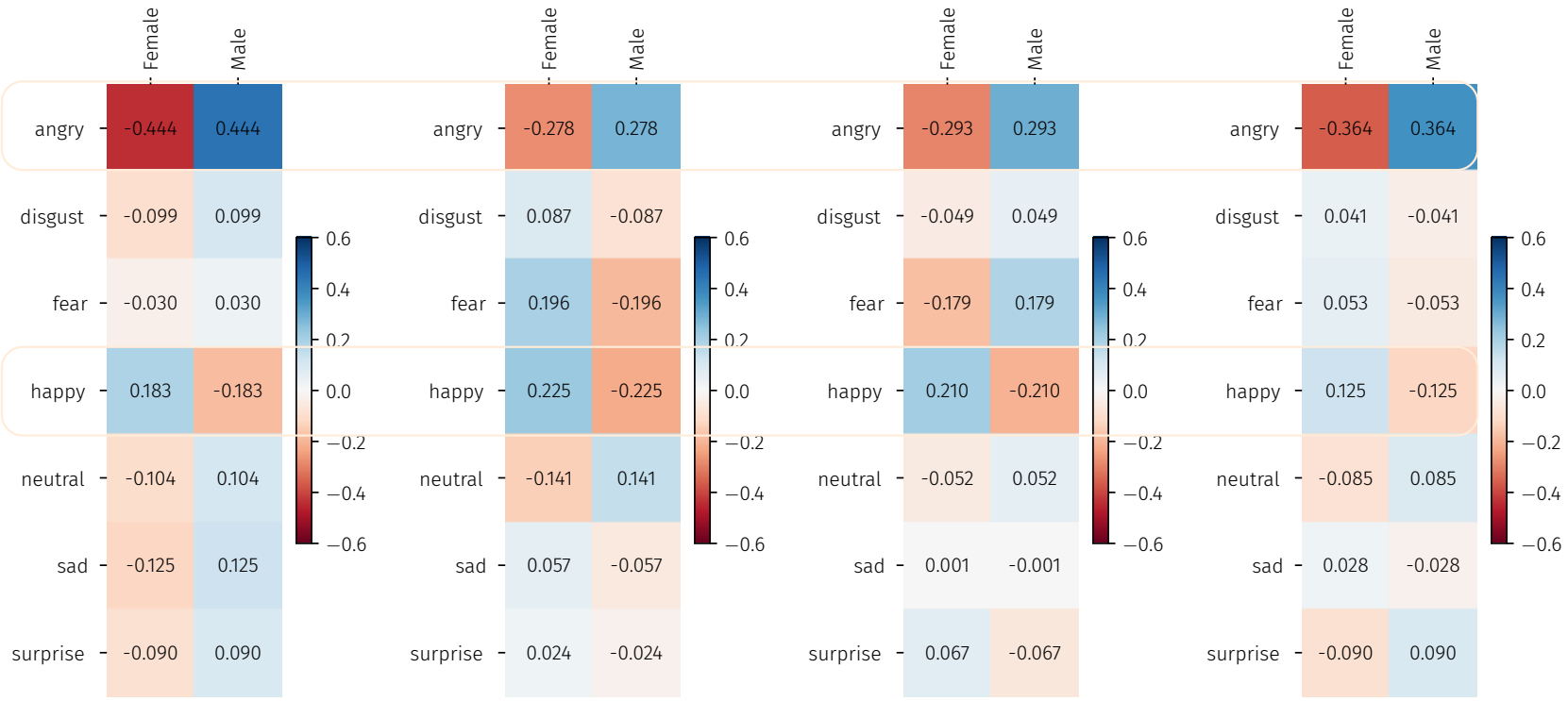


Figure 12: Local stereotypical bias for gender in Affectnet, Fer+, NHFIER y Raf-DB (Ducher's Z). (F: Female, M: Male)

LOCAL STEREOTYPICAL BIAS

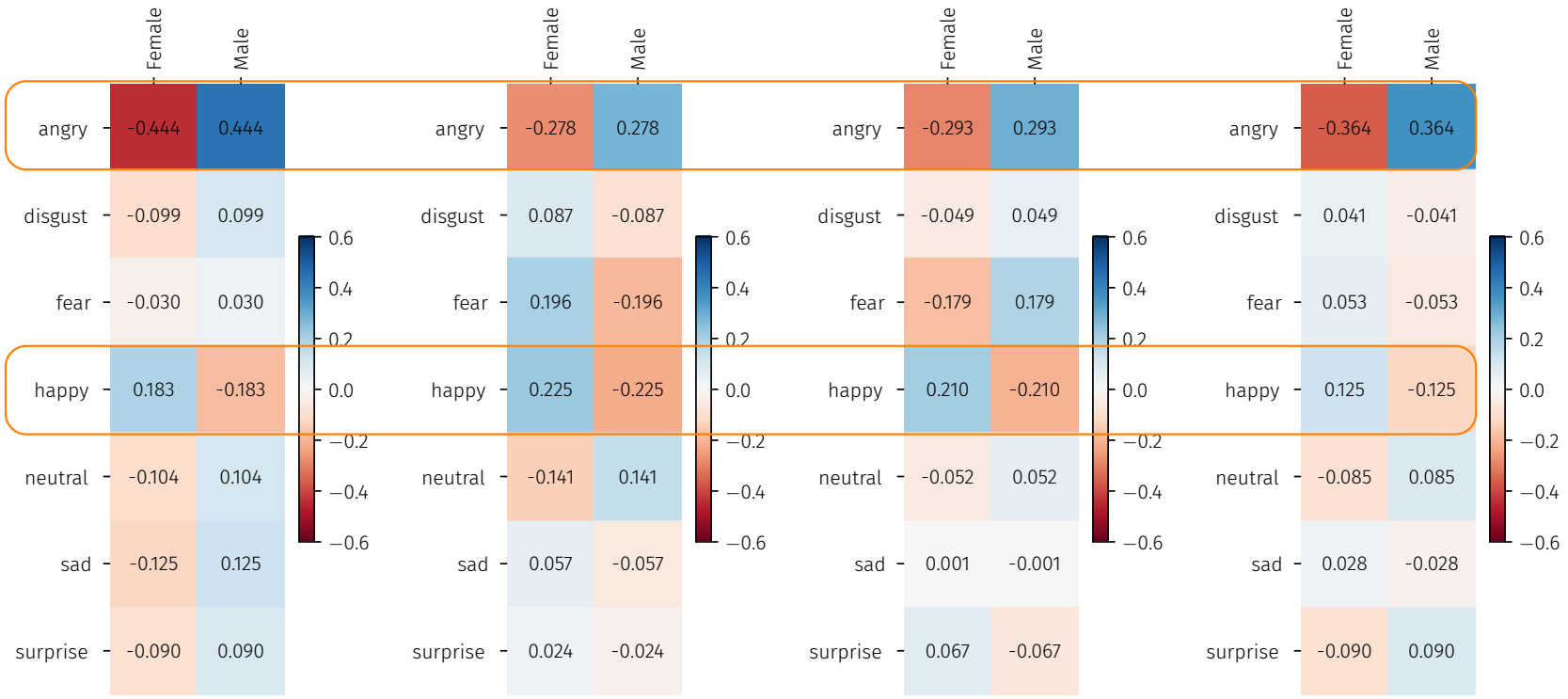
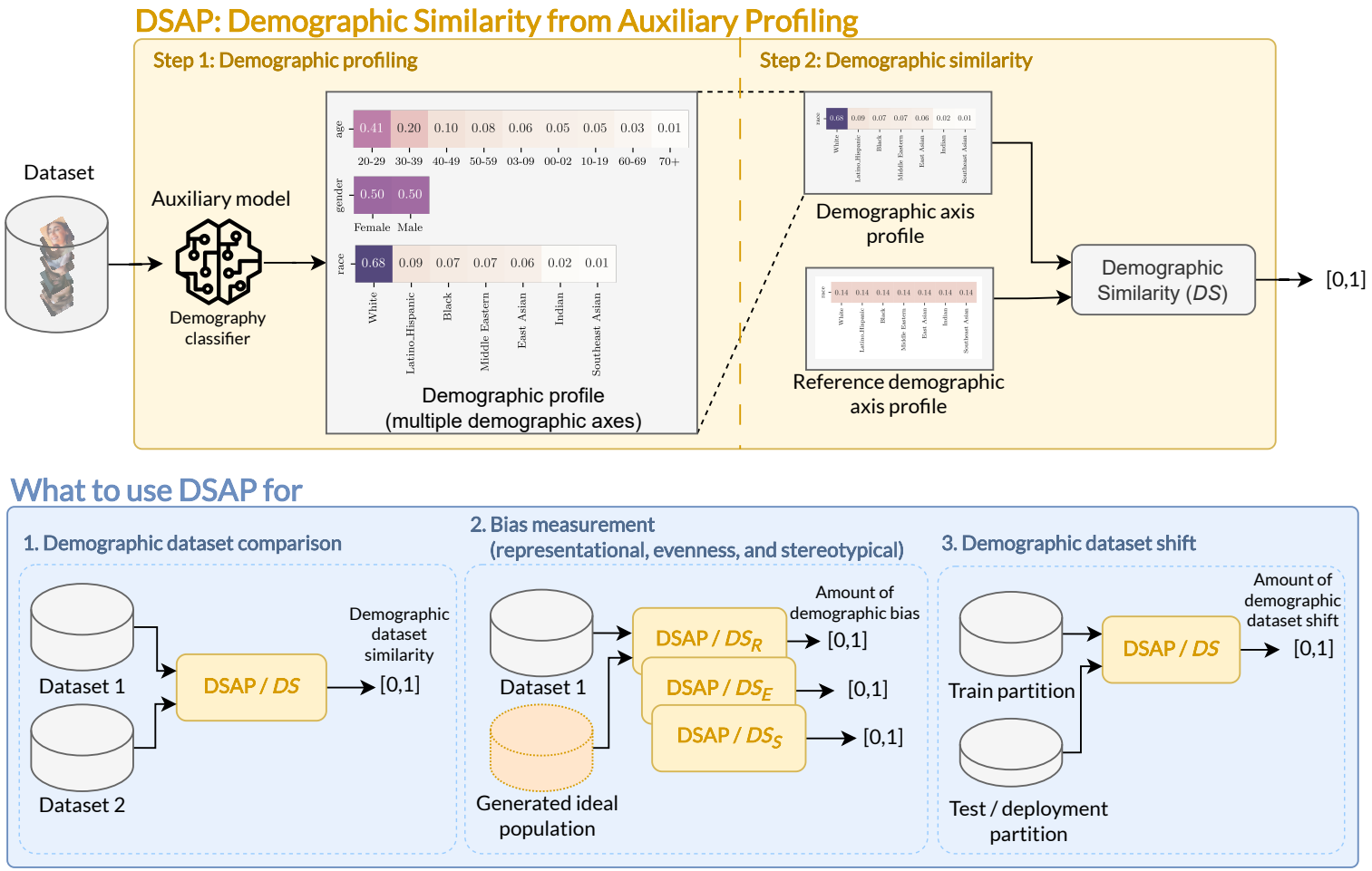


Figure 12: Local stereotypical bias for gender in Affectnet, Fer+, NHFIER y Raf-DB (Ducher's Z). (F: Female, M: Male)

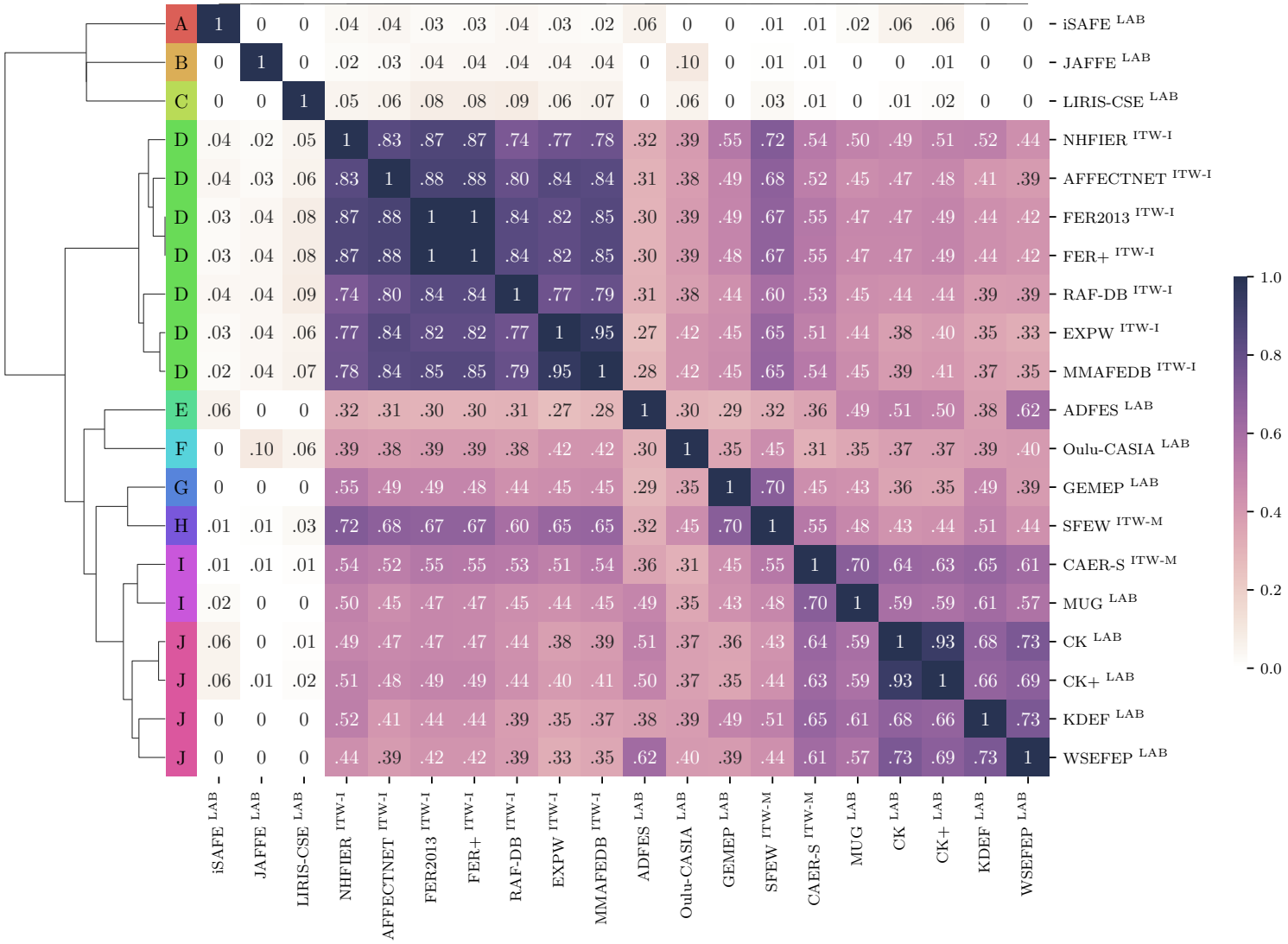
WHAT'S NEXT?

DATASET COMPARISON



[6] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *DSAP: Analyzing Bias Through Demographic Comparison of Datasets*. Dec. 22, 2023. DOI: 10.48550/arXiv.2312.14626. arXiv: 2312.14626 [cs]. URL: http://arxiv.org/abs/2312.14626 (visited on 01/24/2024). preprint

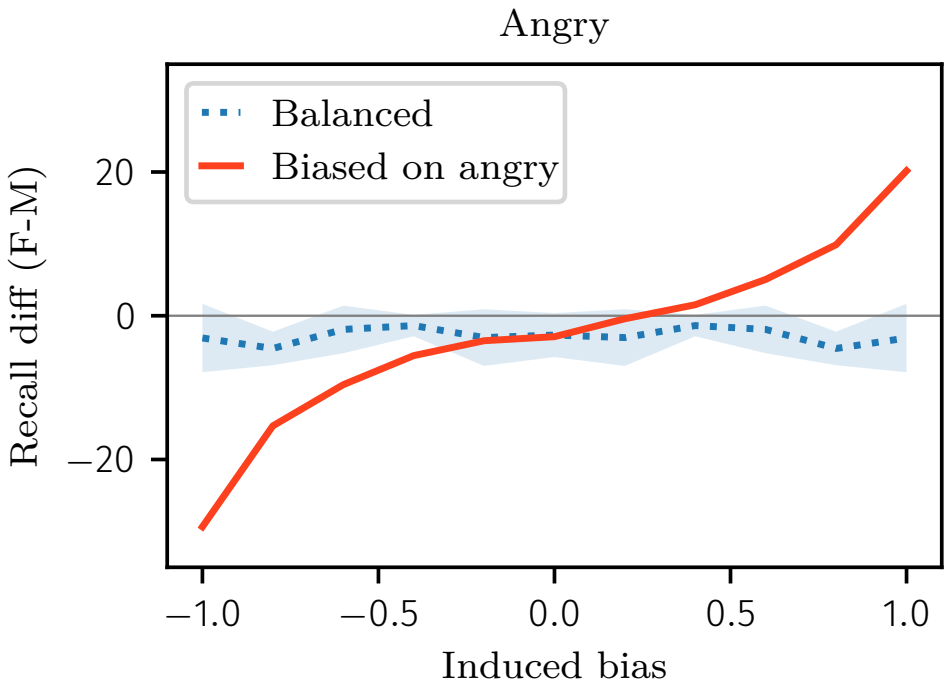
DATASET COMPARISON



[6] Dominguez-Catena, Paternain, and Galar, *DSAP*

STEREOTYPICAL BIAS TRANSFERENCE

		Biased on angry		
		Female −1.00	Female 0.00	Female +1.00
angry	Male	**81.25 ± 2.61**	76.30 ± 3.40	55.87 ± 2.24
	Female	51.89 ± 3.77	73.21 ± 3.33	**76.04 ± 1.99**
	Diff	**−29.36 ± 4.59**	−3.10 ± 4.76	20.17 ± 2.99



[6] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Gender Stereotyping Impact in Facial Expression Recognition". In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Vol. 1752. Cham: Springer Nature Switzerland, 2023, pp. 9–22. ISBN: 978-3-031-23617-4 978-3-031-23618-1. DOI: 10.1007/978-3-031-23618-1_1

BIAS IN OTHER PROBLEMS

AI IN HIRING: AMAZON

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

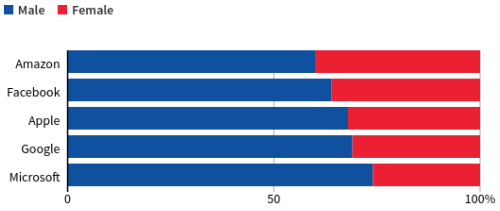
By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

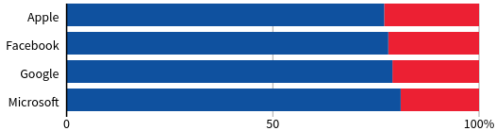
Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT



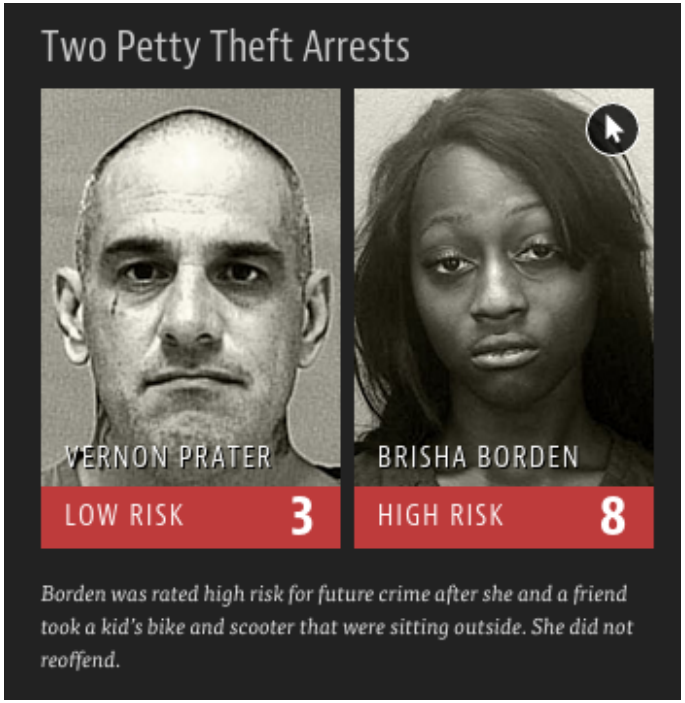
EMPLOYEES IN TECHNICAL ROLES

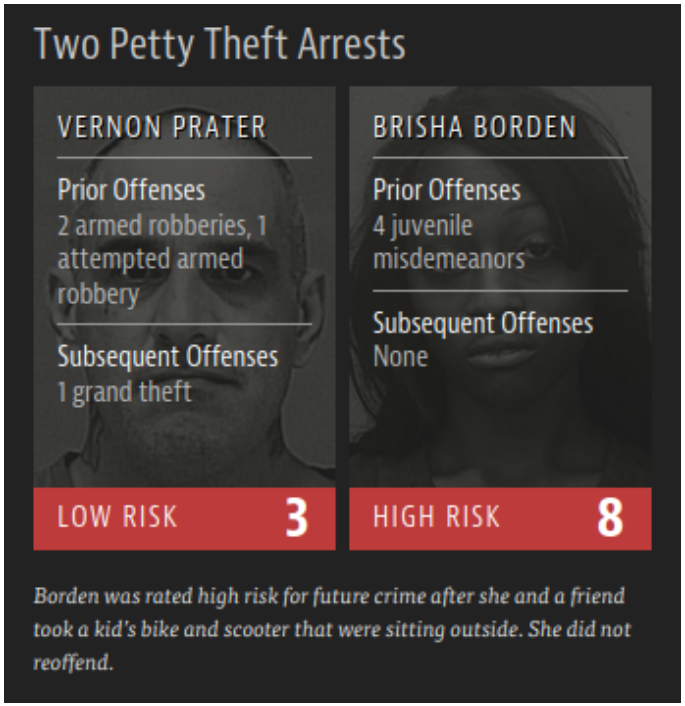


Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

[6] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

RECIDIVISM PREDICTION: COMPASS



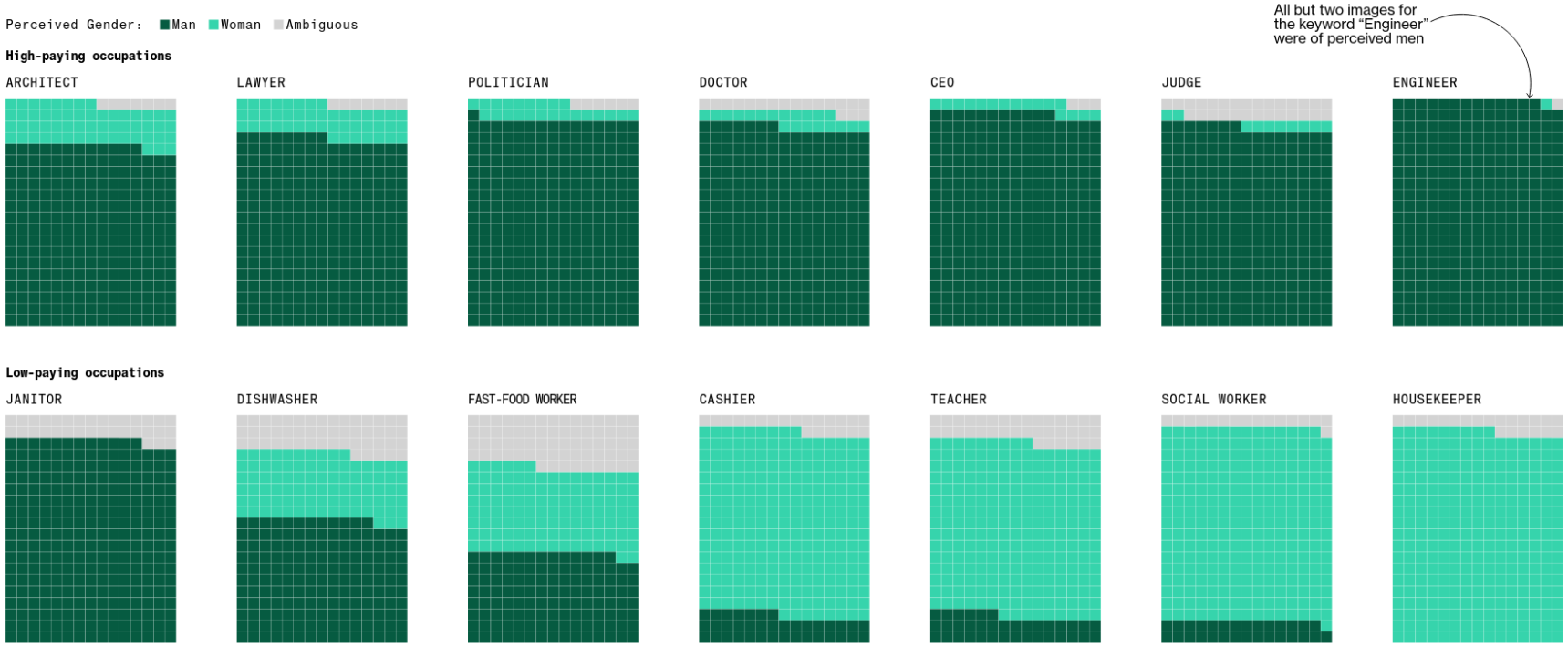


Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

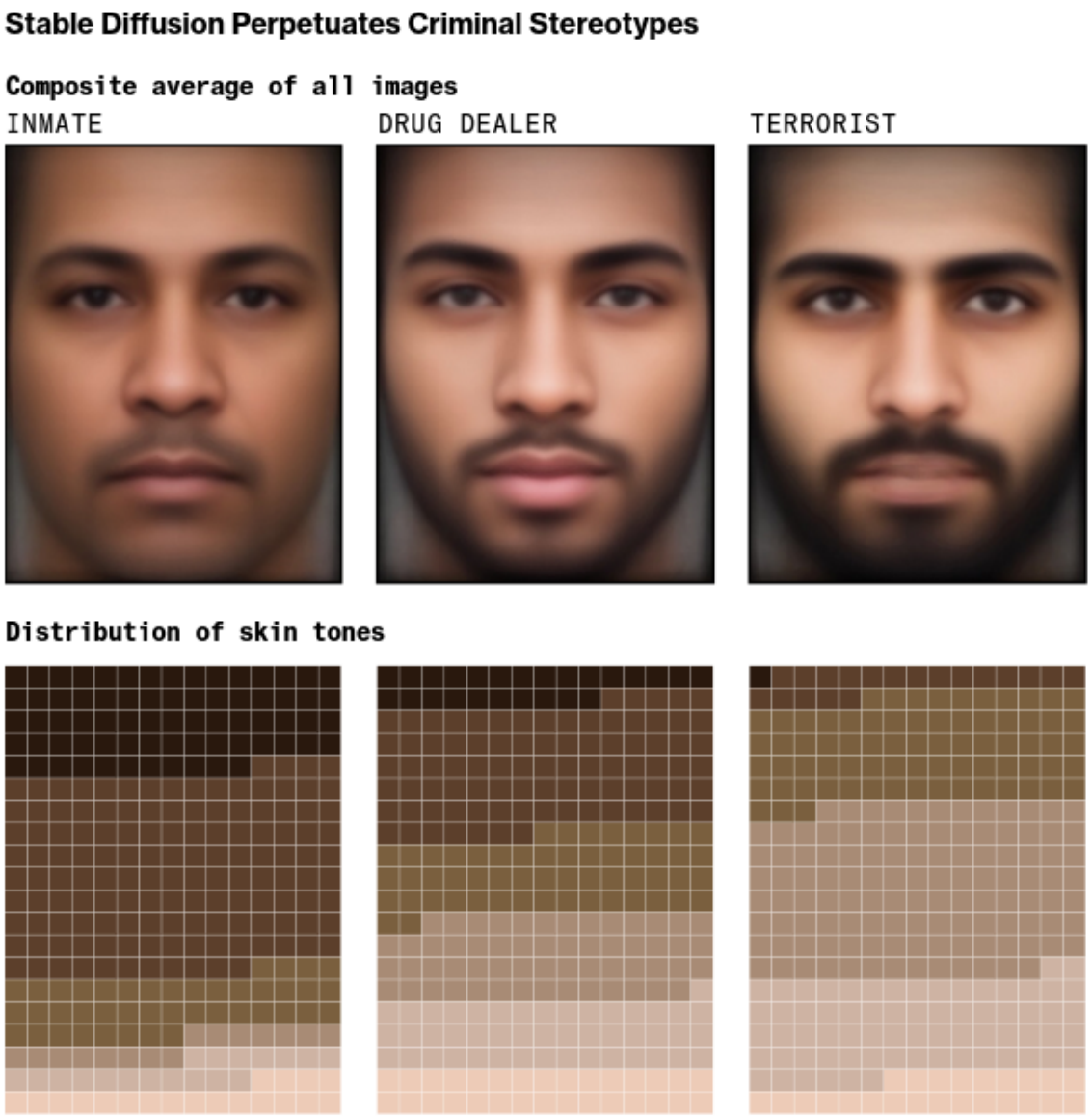
[6] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

GENERATIVE AI: STABLE DIFFUSION



[6] https://www.bloomberg.com/graphics/2023-generative-ai-bias/

GENERATIVE AI: STABLE DIFFUSION



[6] https://www.bloomberg.com/graphics/2023-generative-ai-bias/

THAT'S ONLY THE BEGINNING

- ChatGPT's political biases[7]
- AI agents and bias: Tay.ai [8]
- Military AI [9]
- Medical AI [10]
- Social networks and recommendation algorithms [11]

[7] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. "More Human than Human: Measuring ChatGPT Political Bias". In: *Public Choice* (Aug. 2023). ISSN: 1573-7101. DOI: 10.1007/s11127-023-01097-2.

[8] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[9] https://www.euronews.com/next/2022/10/17/israel-deploys-ai-powered-robot-guns-that-can-track-targets-in-the-west-bank

[10] https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

[11] https://www.adl.org/resources/report/exposure-alternative-extremist-content-youtube

CONCLUSION

CONCLUSION

- IAs **are not perfect**, ”neutral” or fair
- They can replicate and worsen our biases, especially through data
- **Measuring** these biases is vital to tackle them

¿Questions?

✉ iris.dominguez@unavarra.es

https://irisai.neocities.org

FORMULAS

Effective Number of Species (ENS)[12]:

$$\mathrm{ENS}(X) = \exp\left(-\sum_{g \in G} p_g \ln p_g\right) . \tag{1}$$

Adjusted entropy. *Effective* number of represented group.

[12] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (May 2006), pp. 363–375. ISSN: 00301299. DOI: 10.1111/j.2006.0030-1299.14714.x.

Shannon Evenness Index (SEI)[13]:

$$\mathrm{SEI}(X) = \frac{\mathrm{H}(X)}{\ln(\mathrm{R}(X))}\ , \tag{2}$$

where $\mathrm{H}(X)$ is Shannon entropy.

Group evenness.

[13] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (Dec. 1966), pp. 131–144. ISSN: 00225193. DOI: 10.1016/0022-5193(66)90013-0.

Berger-Parker Index (BP)[14]:

$$\mathrm{BP}(X) = \frac{\max\limits_{g \in G} n_g}{n} \,. \tag{3}$$

Ratio between the most represented group and the whole population.

[14] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (June 12, 1970), pp. 1345–1347. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.168.3937.1345.

Cramer's V (ϕ_C)[15]:

$$\chi^2(X) = \sum_{g \in G} \sum_{y \in Y} \frac{(n_{g \wedge y} - \frac{n_g n_y}{n})^2}{\frac{n_g n_y}{n}} , \tag{4}$$

$$\phi_C(X) = \sqrt{\frac{\chi^2(X)/n}{\min(|G| - 1, |Y| - 1)}} , \tag{5}$$

[15] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. Princeton: Princeton university press, 1991, p. 282. ISBN: 978-0-691-08004-8.

Ducher's Z (Z)[16]:

$$Z(X, g, y) = \begin{cases} \frac{p_{g \wedge y} - p_g p_y}{\mathbf{min}[p_g, p_y] - p_g p_y} & \text{if } p_{g \wedge y} - p_g p_y > 0 \\ \frac{p_{g \wedge y} - p_g p_y}{p_g p_y - \mathbf{max}[0, p_g + p_y - 1]} & \text{if } p_{g \wedge y} - p_g p_y < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

[16] M. Ducher et al. "Statistical Relationships between Systolic Blood Pressure and Heart Rate and Their Functional Significance in Conscious Rats". In: *Medical & Biological Engineering & Computing* 32.6 (Nov. 1994), pp. 649–655. ISSN: 0140-0118, 1741-0444. DOI: 10.1007/BF02524241.