¿Qué sesgos importan? Envenenando conjuntos de datos para clasificación de imágenes

Iris Dominguez-Catena

Noviembre 2024

Departamento de Estadística, Informática y Matemáticas,
Universidad Pública de Navarra(UPNA)

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

*iris.dominguez@unavarra.es

Outline

Outline

- **AI ethics**: multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.
  - **Algorithmic fairness**: Ensuring that algorithms make non-discriminatory decisions.

> *«Fairness is man's ability to rise above his prejudices.»*
>
> *Wes Fesler*

Algorithmic fairness

- **AI ethics**: multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.
  - **Algorithmic fairness**: Ensuring that algorithms make non-discriminatory decisions.

> *«Fairness is man's ability to rise above his prejudices.»*
> *Wes Fesler*

- **AI ethics**: multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.
  - **Algorithmic fairness**: Ensuring that algorithms make non-discriminatory decisions.

> Algorithmic machine's people's
> *«Fairness is ~~man's~~ ability to rise above ~~his~~ prejudices.»*
>
> *Wes Fesler*

AI IN HIRING: AMAZON

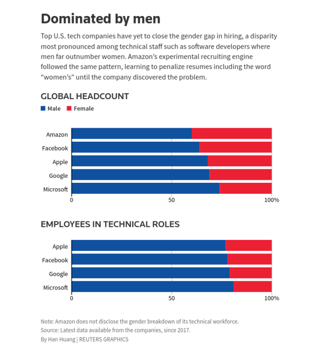
RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

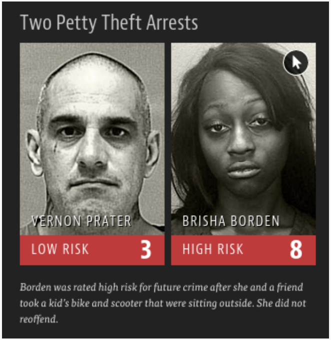
8 MIN READ

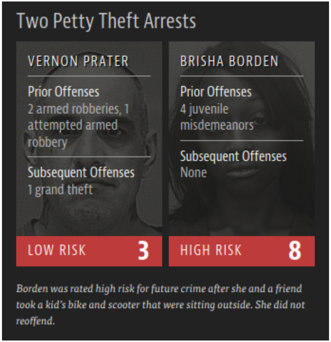
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



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

Recidivism prediction: COMPASS



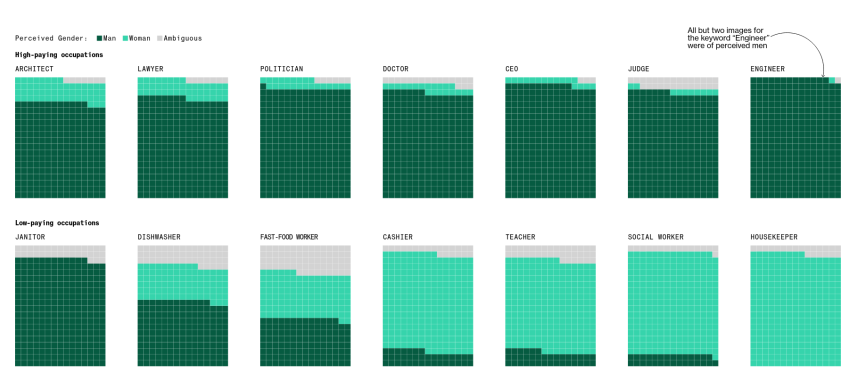


Prediction Fails Differently for Black Defendants

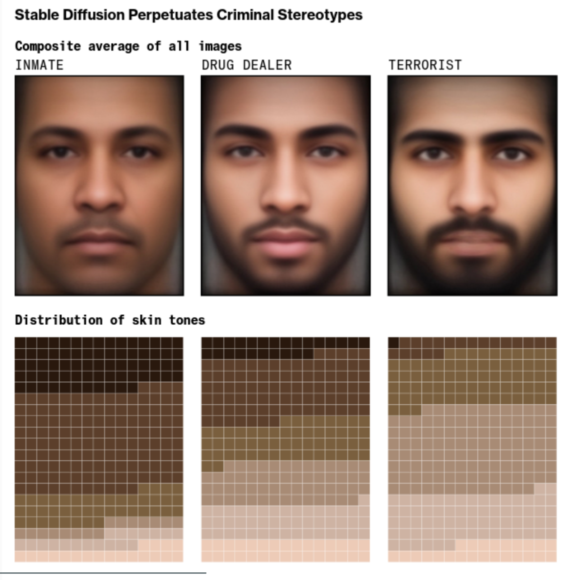
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

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

Generative AI: Stable Diffusion



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



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

That's only the beginning

- ChatGPT's political biases[1]
- AI agents and bias: Tay.ai [2]
- Military AI [3]
- Medical AI [4]
- Social networks and recommendation algorithms [5]

[1] Motoki2023.

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

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

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

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

BIAS SOURCES

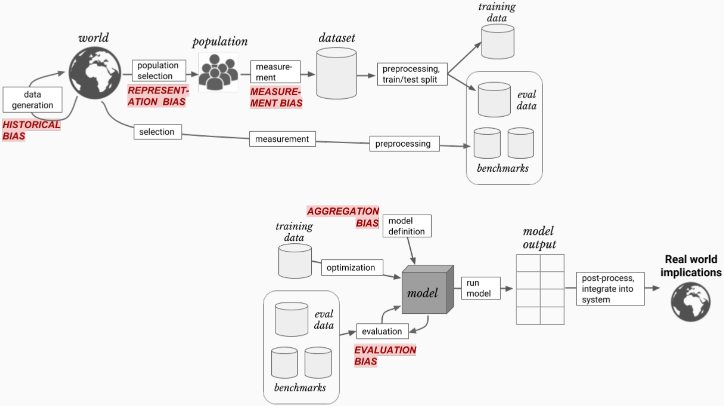


Figure 1: Bias source in the machine learning pipeline[6]

[6] 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. 2021, pp. 1–9.

Bias sources

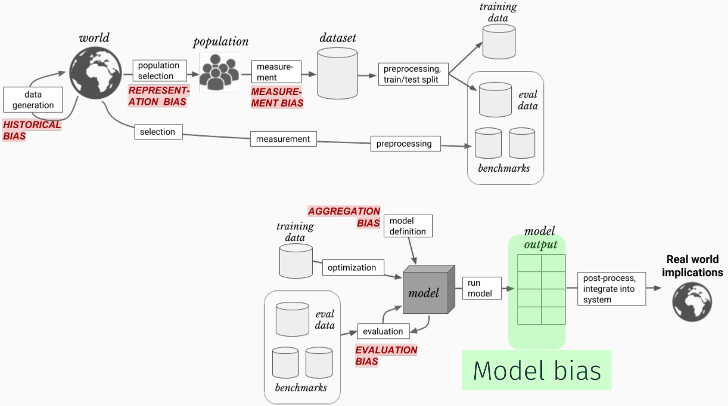


Figure 1: Bias source in the machine learning pipeline[6]

[6] 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. 2021, pp. 1–9.

BIAS SOURCES

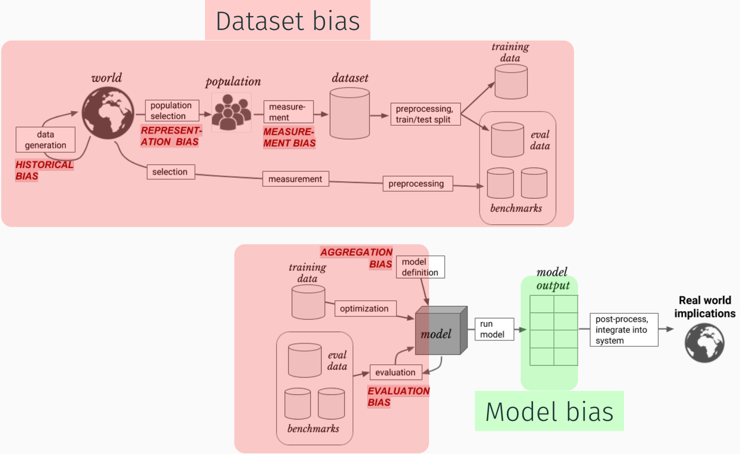


Figure 1: Bias source in the machine learning pipeline[6]

[6] 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. 2021, pp. 1–9.

Facial Expression Recognition

Figure 2: A sample of FER2013/FER+, a popular FER dataset[7].

[7] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *Proc. 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 279–283.

Modalities

- **Image** or video
- **RGB**, IR, Depth...
- **Discrete** (Ekman's basic emotions [8]) or continuous (NRC-VAD) labeling...

Applications

- Interactive multimedia
    - Emotional Films
- Healthcare [9]
- Assistive robotics [10]
- Public safety [11]

[9] Paul Ekman and Wallace V. Friesen. "Constants across Cultures in the Face and Emotion.". In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129

[10] Philipp Werner et al. "Automatic Recognition Methods Supporting Pain Assessment: A Survey". In: *IEEE Trans. on Affective Computing* 13.1 (2022), pp. 530–552

[11] Ritvik Nimmagadda, Kritika Arora, and Miguel Vargas Martin. "Emotion Recognition Models for Companion Robots". In: *The Journal of Supercomputing* (2022)

[12] Luntian Mou et al. "Isotropic Self-Supervised Learning for Driver Drowsiness Detection With Attention-Based Multimodal Fusion". In: *IEEE Trans. on Multimedia* 25 (2023), pp. 529–542

- Age, race and gender biases:
    - In research models[13].
    - In commercial systems[14,15].

[13] Tian Xu et al. "Investigating Bias and Fairness in Facial Expression Recognition". In: *Computer Vision – ECCV 2020 Workshops*. 2020, pp. 506–523.

[14] Khurshid Ahmad et al. "Comparing the Performance of Facial Emotion Recognition Systems on Real-Life Videos: Gender, Ethnicity and Age". In: *Proc. Future Technologies Conference (FTC) 2021, Volume 1*. Vol. 358. 2022, pp. 193–210.

[15] Eugenia Kim et al. "Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults". In: *Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 638–644.

FER models and datasets

- Deep Learning approaches: CNN[16] and Transformers[17].
  - They require large amounts of data!
- Shift to large datasets gathered from the Internet[18].
  - Datasets with little to no demographic metadata.

[16] Shan Li and Weihong Deng. "Deep Facial Expression Recognition: A Survey". In: *IEEE Trans. on Affective Computing* (2020), pp. 1–1.

[17] Alexey Dosovitskiy et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs].

[18] Emily Denton et al. "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet". In: *Big Data & Society* 8.2 (2021).

Dataset selection

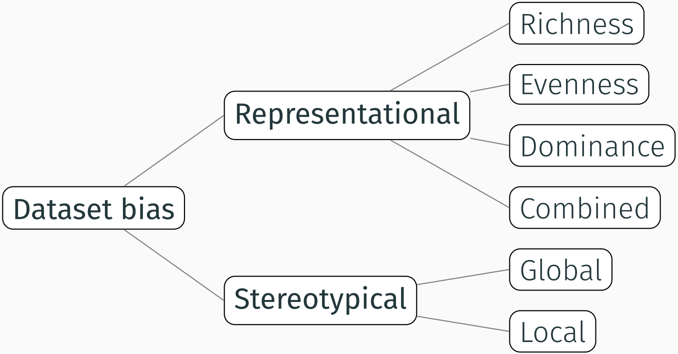
Table 1: 20 selected datasets.

Abbreviation	Year	Collection	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	32,298	—	—
WSEFEP	2014	LAB	210	—	30
ADFES	2016	LAB	—	648	22
FERPlus	2016	ITW	32,298	—	—
AffectNet	2017	ITW	291,652	—	—
ExpW	2017	ITW	91,793	—	—
RAF-DB	2017	ITW	29,672	—	—
CAER-S	2019	ITW-M	70,000	—	—
LIRIS-CSE	2019	LAB	26,000	208	12
iSAFE	2020	LAB	—	395	44
MMAFEDB	2020	ITW	128,000	—	—
NHFIER	2020	ITW	5,558	—	—

- Lab: Laboratory-gathered.
- ITW-I: From Internet.
- ITW-M: From motion pictures.

Outline

Dataset bias taxonomy



Representational bias

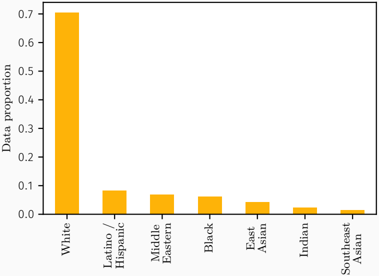


Figure 3: Apparent race distribution in FER+.

Representational bias

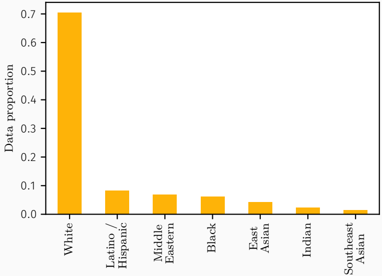


Figure 3: Apparent race distribution in FER+.

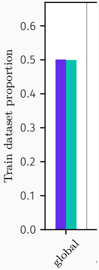


Figure 4: Apparent *per-label* gender distribution in FER+.

Representational bias

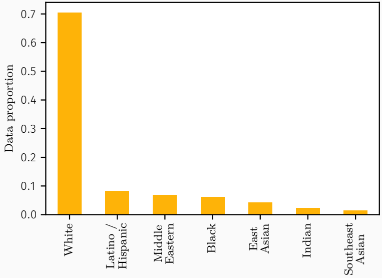


Figure 3: Apparent race distribution in FER+.

Stereotypical bias

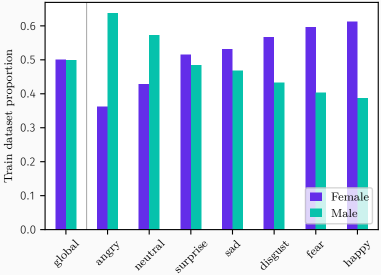


Figure 4: Apparent *per-label* gender distribution in FER+.

Measured bias

		Laboratory	ITW-M	ITW-I
Age	$9 - \text{ENS}$	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	$1 - \text{SEI}$	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	$7 - \text{ENS}$	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	$1 - \text{SEI}$	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	$2 - \text{ENS}$	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	$1 - \text{SEI}$	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Average representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of Lab, ITW-M and ITW-I datasets.

		Laboratory	ITW-M	ITW-I
Age	9 – ENS	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	1 – SEI	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	7 – ENS	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	1 – SEI	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	2 – ENS	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	1 – SEI	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Average representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of Lab, ITW-M and ITW-I datasets.

		Laboratory	ITW-M	ITW-I
Age	$9 - \text{ENS}$	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	$1 - \text{SEI}$	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	$7 - \text{ENS}$	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	$1 - \text{SEI}$	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	$2 - \text{ENS}$	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	$1 - \text{SEI}$	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Average representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C) of Lab, ITW-M and ITW-I datasets.

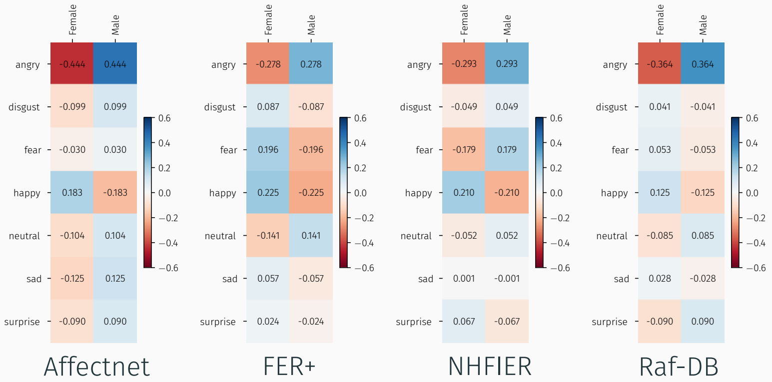


Figure 6: Local stereotypical bias (Ducher's Z) for some ITW-I datasets.

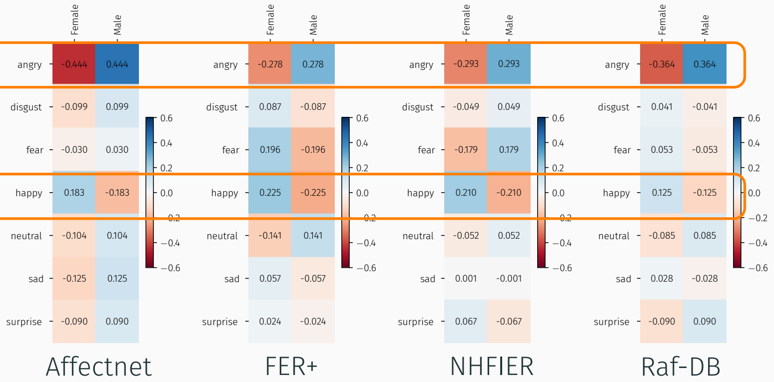


Figure 6: Local stereotypical bias (Ducher's Z) for some ITW-I datasets.

Motivation

- Dataset bias metrics still lack **interpretability**.
    - Issues of comparability and range disparities.
    - Lack of clear meaning.
- Most datasets lack demographic information.
    - Especially modern ITW datasets.
- There is no way to study the evolution and changes in bias across datasets.
    - Do equally biased dataset represent the same populations?
    - Analogous problems in archaeology[19] and ecology[20,21].

[19] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *American Antiquity* 16.4 (1951), pp. 293–301.

[20] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

[21] C. Ricotta and J. Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (2017), pp. 201–205.

Motivation

- Dataset bias metrics still lack **interpretability**.
    - Issues of comparability and range disparities.
    - Lack of clear meaning.
- Most datasets lack demographic information.
    - Especially modern ITW datasets.
- There is no way to study the evolution and changes in bias across datasets.
    - Do equally biased dataset represent the same populations?
    - Analogous problems in archaeology[19] and ecology[20,21].

[19] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *American Antiquity* 16.4 (1951), pp. 293–301.

[20] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

[21] C. Ricotta and J. Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (2017), pp. 201–205.

Motivation

- Dataset bias metrics still lack **interpretability**.
  - Issues of comparability and range disparities.
  - Lack of clear meaning.
- Most datasets lack demographic information.
  - Especially modern ITW datasets.
- There is no way to study the evolution and changes in bias across datasets.
  - Do equally biased dataset represent the same **populations**?
  - Analogous problems in archaeology[19] and ecology[20,21].

[19] W. S. Robinson. “A Method for Chronologically Ordering Archaeological Deposits”. In: *American Antiquity* 16.4 (1951), pp. 293–301.

[20] M. V. Wilson and A. Shmida. “Measuring Beta Diversity with Presence-Absence Data”. In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

[21] C. Ricotta and J. Podani. “On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning”. In: *Ecological Complexity* 31 (2017), pp. 201–205.

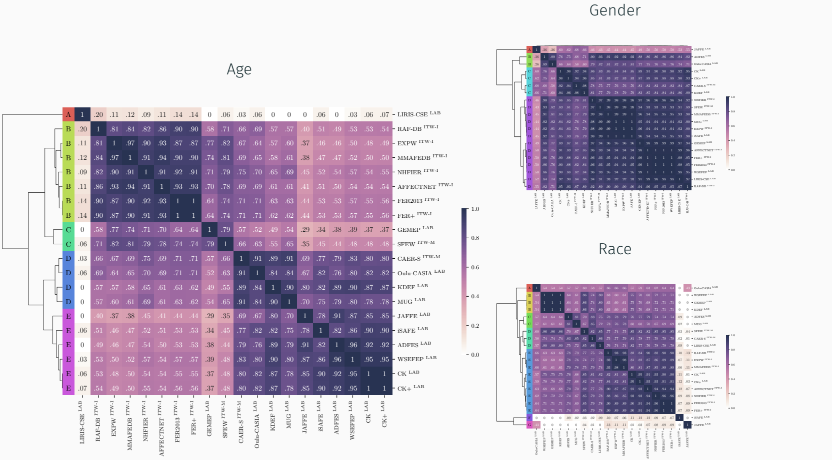


Figure 7: DSAP based comparison of datasets (age, gender and race axis).

Results: Clustering II

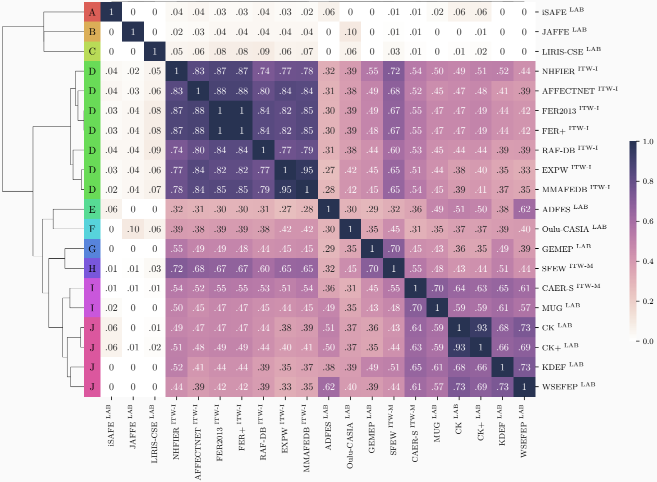


Figure 8: DSAP based comparison of datasets (combination axis, 126 subgroups).

Outline

Methodology

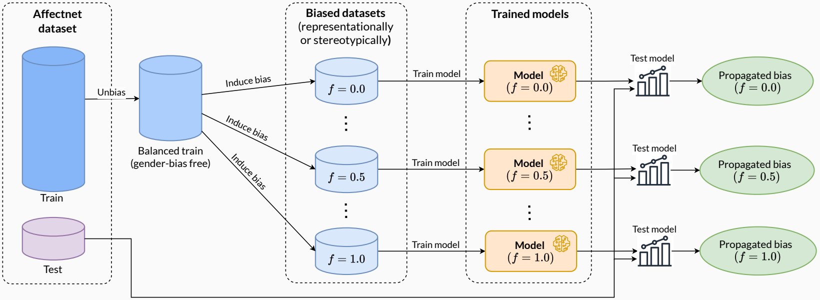


Figure 9: Summary of the methodology

Representational bias

Gender	Female	Male	% F
angry	6, 962	0	100%
disgust	1, 733	0	100%
fear	3, 117	0	100%
happy	54, 241	0	100%
neutral	33, 708	0	100%
sad	11, 175	0	100%
surprise	6, 439	0	100%
Total	117, 375	0	100%

Original

Gender	Female	Male	% F
angry	6, 962	17, 803	28.11%
disgust	1, 733	2, 054	45.76%
fear	3, 117	3, 222	49.17%
happy	79, 797	54, 241	59.53%
neutral	33, 708	40, 858	45.21%
sad	11, 175	14, 156	44.12%
surprise	6, 439	7, 588	45.90%
Total	142, 931	139, 922	50.53%

Balanced

Gender	Female	Male	% F
angry	6, 962	6, 962	50%
disgust	1, 733	1, 733	50%
fear	3, 117	3, 117	50%
happy	54, 241	54, 241	50%
neutral	33, 708	33, 708	50%
sad	11, 175	11, 175	50%
surprise	6, 439	6, 439	50%
Total	117, 375	117, 375	50%

Stereotypical bias

Gender	Female	Male	% F
angry	0	6, 962	0%
disgust	866	866	50%
fear	1, 558	1, 558	50%
happy	27, 120	27, 120	50%
neutral	16, 854	16, 854	50%
sad	5, 587	5, 587	50%
surprise	3, 219	3, 219	50%
Total	55, 204	62, 166	47.4%

Bias induction

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanced

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Representational bias

Gender	Female	Male	% F
angry	6,962	0	100%
disgust	1,733	0	100%
fear	3,117	0	100%
happy	54,241	0	100%
neutral	33,708	0	100%
sad	11,175	0	100%
surprise	6,439	0	100%
Total	117,375	0	100%

Stereotypical bias

Gender	Female	Male	% F
angry	0	6,962	0%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,204	62,166	47.4%

Bias induction

Original

Gender	Female	Male	% F
angry	6, 962	17, 803	28.11%
disgust	1, 733	2, 054	45.76%
fear	3, 117	3, 222	49.17%
happy	79, 797	54, 241	59.53%
neutral	33, 708	40, 858	45.21%
sad	11, 175	14, 156	44.12%
surprise	6, 439	7, 588	45.90%
Total	142, 931	139, 922	50.53%

Balanced

Gender	Female	Male	% F
angry	6, 962	6, 962	50%
disgust	1, 733	1, 733	50%
fear	3, 117	3, 117	50%
happy	54, 241	54, 241	50%
neutral	33, 708	33, 708	50%
sad	11, 175	11, 175	50%
surprise	6, 439	6, 439	50%
Total	117, 375	117, 375	50%

Representational bias

Gender	Female	Male	% F
angry	6, 962	0	100%
disgust	1, 733	0	100%
fear	3, 117	0	100%
happy	54, 241	0	100%
neutral	33, 708	0	100%
sad	11, 175	0	100%
surprise	6, 439	0	100%
Total	117, 375	0	100%

Stereotypical bias

Gender	Female	Male	% F
angry	0	6, 962	0%
disgust	866	866	50%
fear	1, 558	1, 558	50%
happy	27, 120	27, 120	50%
neutral	16, 854	16, 854	50%
sad	5, 587	5, 587	50%
surprise	3, 219	3, 219	50%
Total	55, 204	62, 166	47.4%

Bias induction

Original

Gender	Female	Male	% F
angry	6, 962	17, 803	28.11%
disgust	1, 733	2, 054	45.76%
fear	3, 117	3, 222	49.17%
happy	79, 797	54, 241	59.53%
neutral	33, 708	40, 858	45.21%
sad	11, 175	14, 156	44.12%
surprise	6, 439	7, 588	45.90%
Total	142, 931	139, 922	50.53%

Balanced

Gender	Female	Male	% F
angry	6, 962	6, 962	50%
disgust	1, 733	1, 733	50%
fear	3, 117	3, 117	50%
happy	54, 241	54, 241	50%
neutral	33, 708	33, 708	50%
sad	11, 175	11, 175	50%
surprise	6, 439	6, 439	50%
Total	117, 375	117, 375	50%

Representational bias

Gender	Female	Male	% F
angry	6, 962	0	100%
disgust	1, 733	0	100%
fear	3, 117	0	100%
happy	54, 241	0	100%
neutral	33, 708	0	100%
sad	11, 175	0	100%
surprise	6, 439	0	100%
Total	117, 375	0	100%

Stereotypical bias

Gender	Female	Male	% F
angry	0	6, 962	0%
disgust	866	866	50%
fear	1, 558	1, 558	50%
happy	27, 120	27, 120	50%
neutral	16, 854	16, 854	50%
sad	5, 587	5, 587	50%
surprise	3, 219	3, 219	50%
Total	55, 204	62, 166	47.4%

BIAS INDUCTION

Original

Gender	Female	Male	% F
angry	6, 962	17, 803	28.11%
disgust	1, 733	2, 054	45.76%
fear	3, 117	3, 222	49.17%
happy	79, 797	54, 241	59.53%
neutral	33, 708	40, 858	45.21%
sad	11, 175	14, 156	44.12%
surprise	6, 439	7, 588	45.90%
Total	142, 931	139, 922	50.53%

Balanced

Gender	Female	Male	% F
angry	6, 962	6, 962	50%
disgust	1, 733	1, 733	50%
fear	3, 117	3, 117	50%
happy	54, 241	54, 241	50%
neutral	33, 708	33, 708	50%
sad	11, 175	11, 175	50%
surprise	6, 439	6, 439	50%
Total	117, 375	117, 375	50%

Representational bias

Gender	Female	Male	% F
angry	1, 392	5, 570	20%
disgust	347	1, 386	20%
fear	623	2, 494	20%
happy	10, 848	43, 393	20%
neutral	6, 742	26, 966	20%
sad	2, 235	8, 940	20%
surprise	1, 288	5, 151	20%
Total	23, 475	93, 900	20%

Stereotypical bias

Gender	Female	Male	% F
angry	1, 392	5, 570	20%
disgust	866	866	50%
fear	1, 558	1, 558	50%
happy	27, 120	27, 120	50%
neutral	16, 854	16, 854	50%
sad	5, 587	5, 587	50%
surprise	3, 219	3, 219	50%
Total	55, 596	60, 774	47.7%

- **Dataset original:** Affectnet[22].
- **Modelos:** ResNet50[23] y ViT-Base[24].
- **Configuraciones de sesgo:** 1 representacional, 7 estereotípicas.
  - **Proportions:** $[0\%, 10\%, \ldots, 100\%]$.

[22]Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Trans. on Affective Computing* 10.1 (2019), pp. 18–31. arXiv: 1708.03985.

[23]Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs].

[24]Alexey Dosovitskiy et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs].

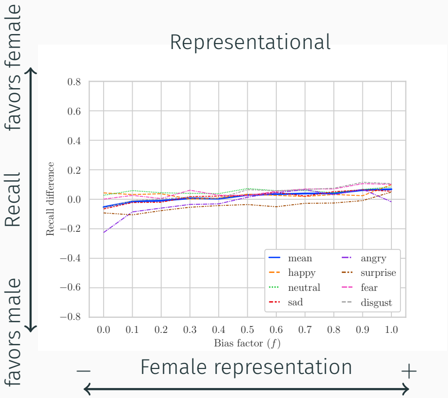


Figure 10: Recall difference (female recall minus male recall).

Results on propagation

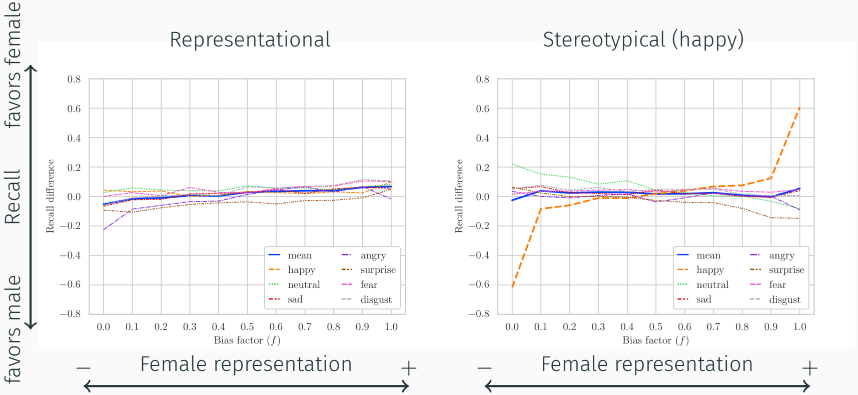


Figure 10: Recall difference (female recall minus male recall).

Results on propagation

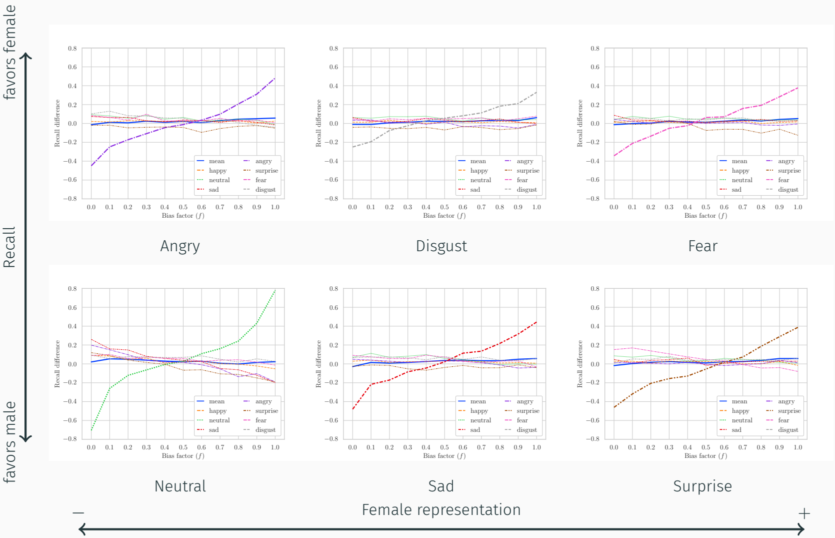


Figure 11: Effect of stereotypical bias.

Outline

General conclusions I

- Los nuevos datasets ITW-I han cambiado de perfil de sesgo. Predomina el sesgo estereotípico.
- Los datasets ITW-I tienden a ser extremadamente homogéneos.
- Los modelos entrenados son mucho más sensibles a sesgo estereotípico.

Thank you for the attention.

¿Questions?

✉ iris.dominguez@unavarra.es

https://irisai.neocities.org