Demographic Bias in Machine Learning: Measuring Transference from Dataset Bias to Model Predictions

Iris Dominguez-Catena

October 2024

Department of Statistics, Computer Science and Mathematics,
Public University of Navarre (UPNA)
Supervisors: Mikel Galar Idoate, Daniel Paternain Dallo

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

*iris.dominguez@unavarra.es

Outline

Outline

Introduction

Algorithmic fairness

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*

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*

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.

Algorithmic machine's people's

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

Introduction

Facial Expression Recognition

Facial Expression Recognition

Figure 1: A sample of FER2013/FER+, a popular FER dataset[1].

[1] 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.

FER modalities and applications

Modalities

- **Image** or video
- **RGB**, IR, Depth...
- **Discrete** (Ekman's basic emotions [2]) or continuous (NRC-VAD) labeling...

Applications

- Interactive multimedia
  - Emotional Films
- Healthcare [3]
- Assistive robotics [4]
- Public safety [5]

---

[3] 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

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

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

[6] 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

FER and FER-related known biases

- Age, race and gender biases:
  - In research models[7].
  - In commercial systems[8,9].
- But! All of them perform manual bias evaluation.

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

[8] 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.

[9] 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 and FER-related known biases

- Age, race and gender biases:
  - In research models[7].
  - In commercial systems[8,9].
- But! All of them perform manual bias evaluation.

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

[8] 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.

[9] 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[10] and Transformers[11].
  - They require large amounts of data!
- Shift to large datasets gathered from the Internet[12].
  - Datasets with little to no demographic metadata.

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

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

[12] 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

Initially: 55 candidates

Selection criteria

1. Image based datasets.
2. RGB images.
3. Ekman's basic emotions.
4. Publicly available.

3 data sources

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

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	—	—

Introduction

Fairness and Bias

Fairness definitions

- **Group fairness**[13]
  - Protected groups of people should be treated, on *average*, equally.

    e.g. Dem. Par.: $P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)$.
- **Individual fairness**[14]
  - *Similar* people should be treated equally.

    $$\forall x, x' \in \mathcal{X}, \quad d_X(x, x') < \epsilon \implies d_Y(f(x), f(x')) < \delta \ .$$
- **Causal fairness**[15]
  - No decision should be based, either directly or indirectly, on protected attributes.

    $$P(\hat{Y} \mid g, X) = P(\hat{Y} \mid X) \quad \forall g \in G \ .$$

[13] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[14] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[15] Niki Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *ArXiv* (2017).

Fairness definitions

- **Group fairness**[13]
  - Protected groups of people should be treated, on *average*, equally.

    e.g. Dem. Par.: $P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)$.
- **Individual fairness**[14]
  - *Similar* people should be treated equally.

$$\forall x, x' \in \mathcal{X}, \quad d_X(x, x') < \epsilon \implies d_{\hat{Y}}(f(x), f(x')) < \delta \ .$$

- **Causal fairness**[15]
  - No decision should be based, either directly or indirectly, on protected attributes.

$$P(\hat{Y} \mid g, X) = P(\hat{Y} \mid X) \quad \forall g \in G \ .$$

[13] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[14] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[15] Niki Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *ArXiv* (2017).

Fairness definitions

- **Group fairness**[13]
  - Protected groups of people should be treated, on *average*, equally.

    e.g. Dem. Par.: $P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)$.
- **Individual fairness**[14]
  - *Similar* people should be treated equally.

$$\forall x, x' \in \mathcal{X}, \quad d_X(x, x') < \epsilon \implies d_{\hat{Y}}(f(x), f(x')) < \delta \ .$$

- **Causal fairness**[15]
  - No decision should be based, either directly or indirectly, on protected attributes.

$$P(\hat{Y} \mid g, X) = P(\hat{Y} \mid X) \quad \forall g \in G \ .$$

[13] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[14] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[15] Niki Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *ArXiv* (2017).

Fairness definitions

- Group fairness[13]
  - Protected groups of people should be treated, on *average*, equally.

    e.g. Dem. Par.: $P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)$.
- Individual fairness[14]
  - *Similar* people should be treated equally.

$$\forall x, x' \in \mathcal{X}, \quad d_X(x, x') < \epsilon \implies d_{\hat{Y}}(f(x), f(x')) < \delta \ .$$

- Causal fairness[15]
  - No decision should be based, either directly or indirectly, on protected attributes.

$$P(\hat{Y} \mid g, X) = P(\hat{Y} \mid X) \quad \forall g \in G \ .$$

[13] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[14] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[15] Niki Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *ArXiv* (2017).

Mathematical fairness definitions

Fairness

- Defined as **requirements**.
- Binary (fair/unfair).
- Easy to check, but hard to reach.

Bias

- **Measures** unfairness.
- Quantitative.
- Can be incrementally improved.

Mathematical fairness definitions

Fairness

- Defined as **requirements**.
- Binary (fair/unfair).
- Easy to check, but hard to reach.

Bias

- **Measures** unfairness.
- Quantitative.
- Can be incrementally improved.

Introduction

Demographic bias

Bias sources

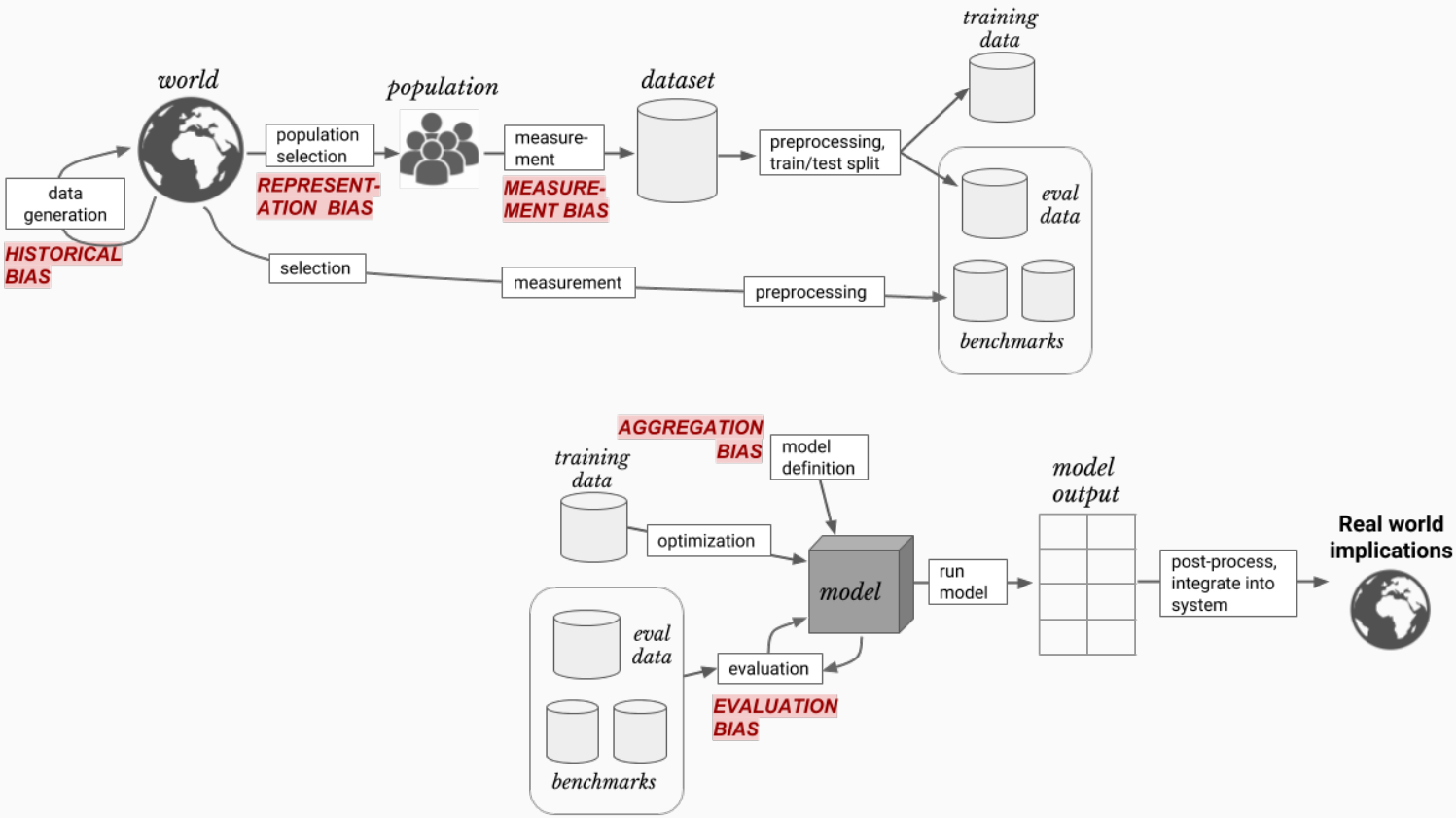


Figure 2: Bias source in the machine learning pipeline[16]

⚠ *There are almost no studies on the transference of bias.*

[16] 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

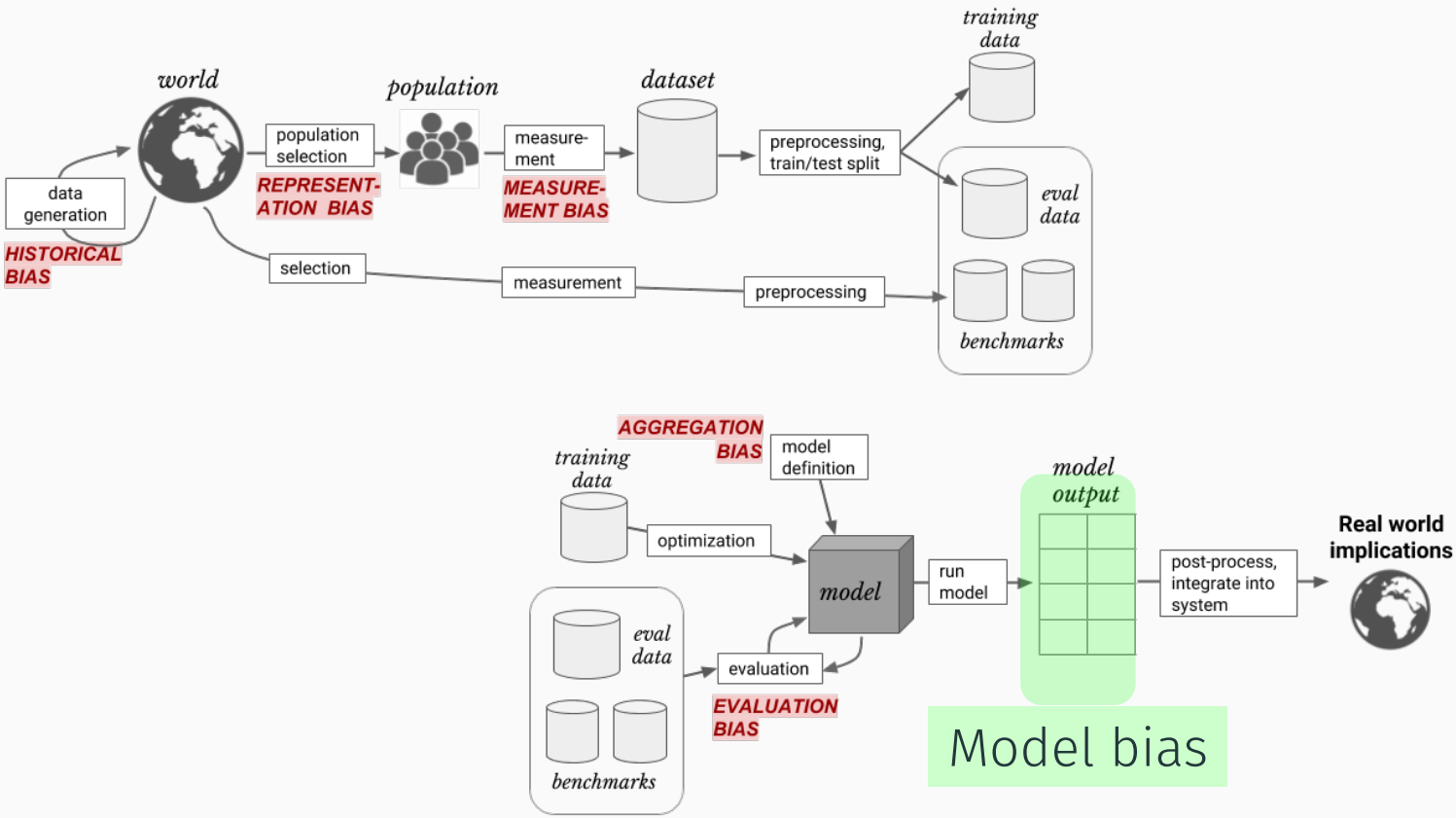


Figure 2: Bias source in the machine learning pipeline[16]

⚠ *There are almost no studies on the transference of bias.*

[16] 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

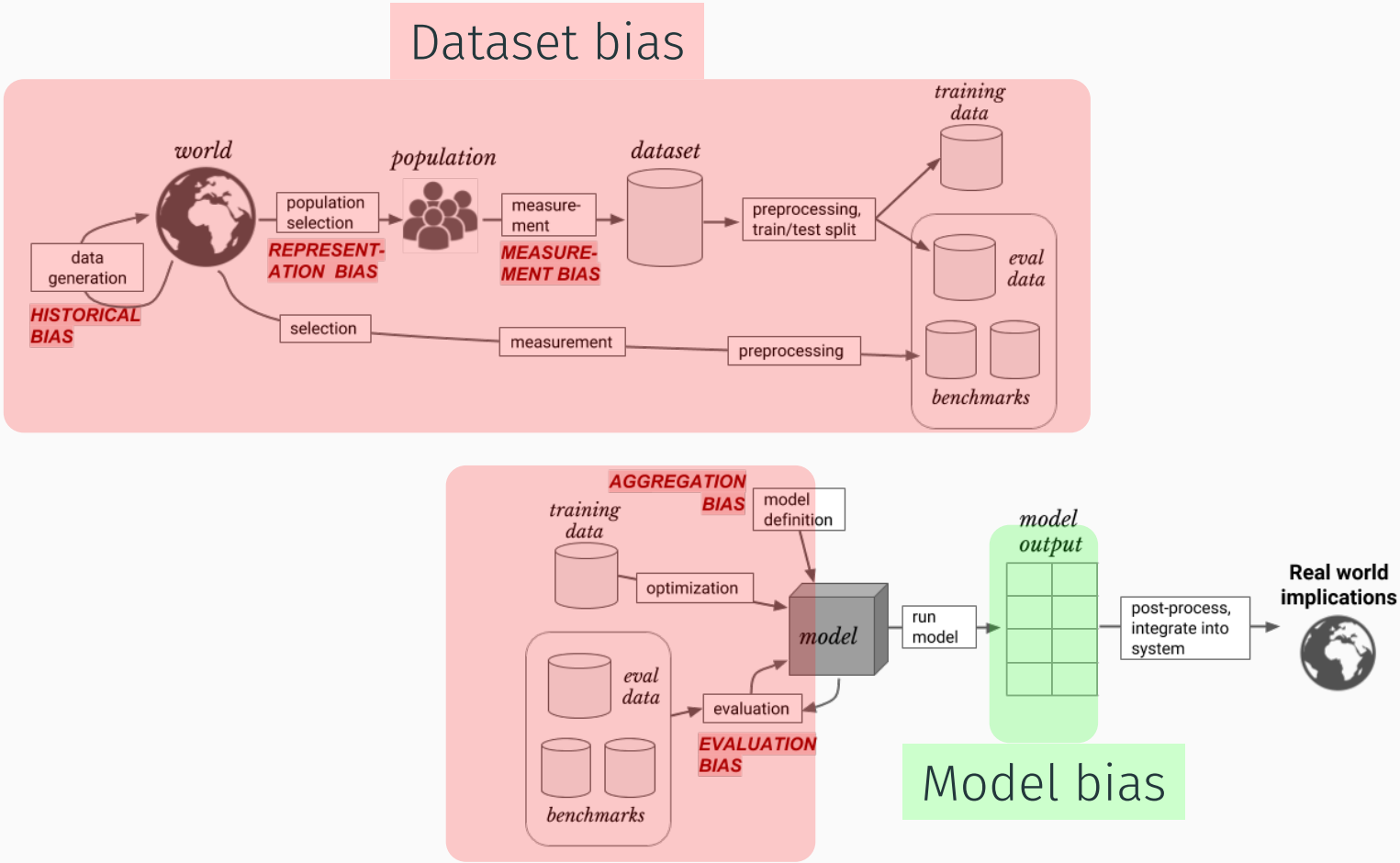


Figure 2: Bias source in the machine learning pipeline[16]

⚠ *There are almost no studies on the transference of bias.*

[16] 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

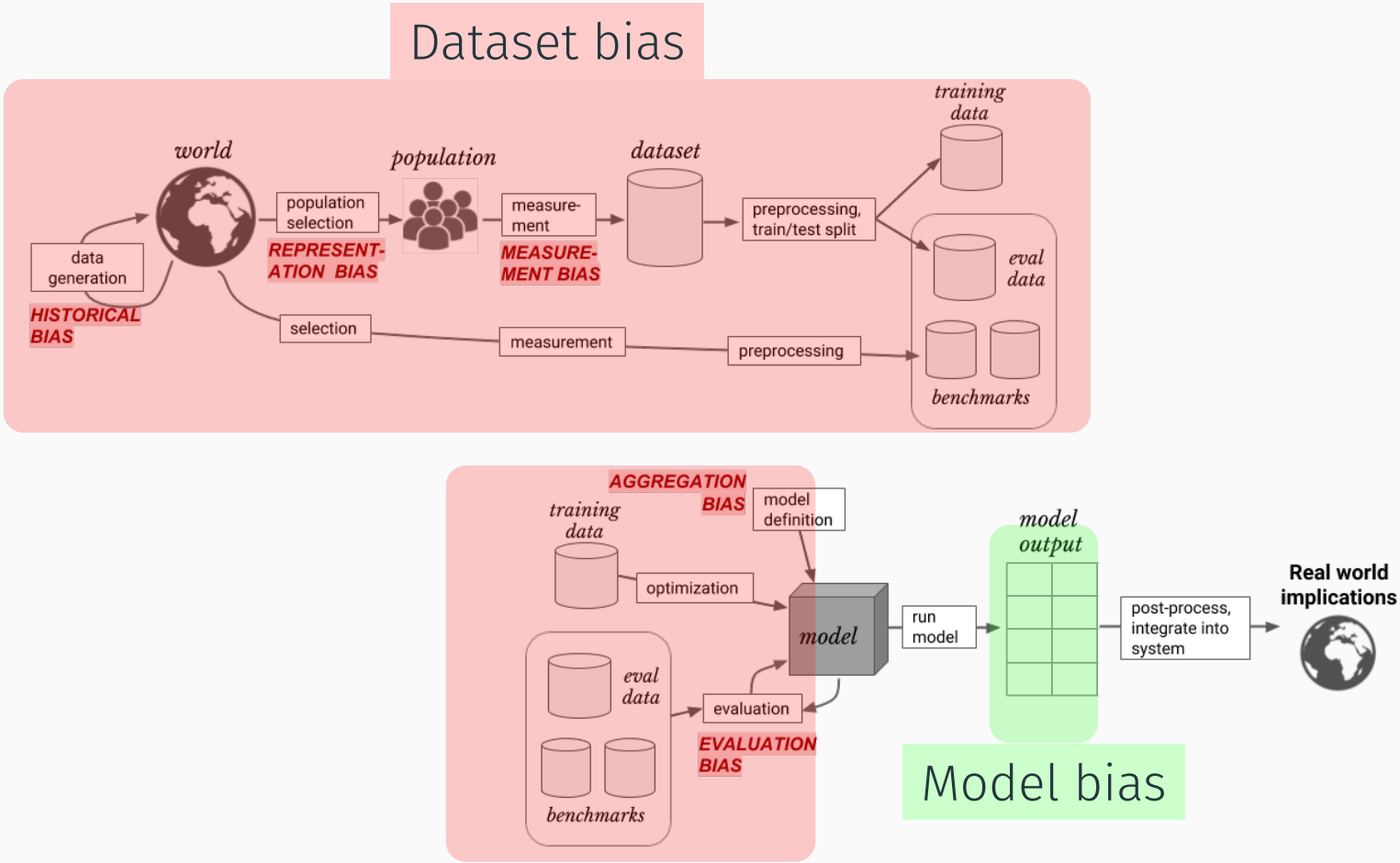


Figure 2: Bias source in the machine learning pipeline[16]

⚠ *There are almost no studies on the transference of bias.*

[16] 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.

Types of bias in image datasets

- **Selection bias** Number of samples for each group.
- **Label bias** Labeling scheme and assignation.
- **Framing bias** Image properties and context.

> **Selection bias** is our main focus:
>
> - Easiest to measure and quantify.
> - Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- **Selection bias** 20,000 examples of white people and 20 of black people.
- **Label bias** Labeling scheme and assignation.
- **Framing bias** Image properties and context.

> **Selection bias** is our main focus:
>
> - Easiest to measure and quantify.
> - Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- **Selection bias** 20,000 examples of white people and 20 of black people.
- **Label bias** Labeling scheme and assignation.
- **Framing bias** Image properties and context.

> **Selection bias** is our main focus:
>
> - Easiest to measure and quantify.
> - Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- **Selection bias** 20,000 examples of white people and 20 of black people.
- **Label bias** Not labeling *children* as *person*.
- **Framing bias** Image properties and context.

Selection bias is our main focus:

- Easiest to measure and quantify.
- Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- **Selection bias** 20,000 examples of white people and 20 of black people.
- **Label bias** Not labeling *children* as *person*.
- **Framing bias** Image properties and context.

> **Selection bias** is our main focus:
>
> - Easiest to measure and quantify.
> - Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- Selection bias: 20,000 examples of white people and 20 of black people.
- Label bias: Not labeling *children* as *person*.
- Framing bias: Lower image quality in images from certain countries.

Selection bias is our main focus:

- Easiest to measure and quantify.
- Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Types of bias in image datasets

- **Selection bias** Number of samples for each group.
- **Label bias** Labeling scheme and assignation.
- **Framing bias** Image properties and context.

> **Selection bias** is our main focus:
>
> - Easiest to measure and quantify.
> - Not limited to image datasets.

[16] Simone Fabbrizzi et al. "A Survey on Bias in Visual Datasets". In: *Computer Vision and Image Understanding* 223 (2022), p. 103552

Dataset bias, an example

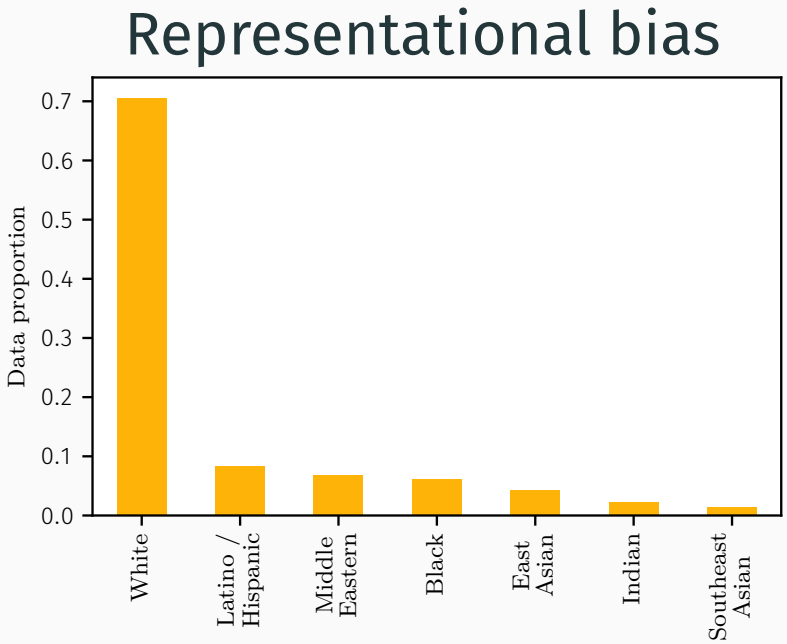


Figure 3: Apparent race distribution in FER+.

⚠ *There are no taxonomies of biases for these classes.*

⚠ *No studies on their impact.*

Dataset bias, an example

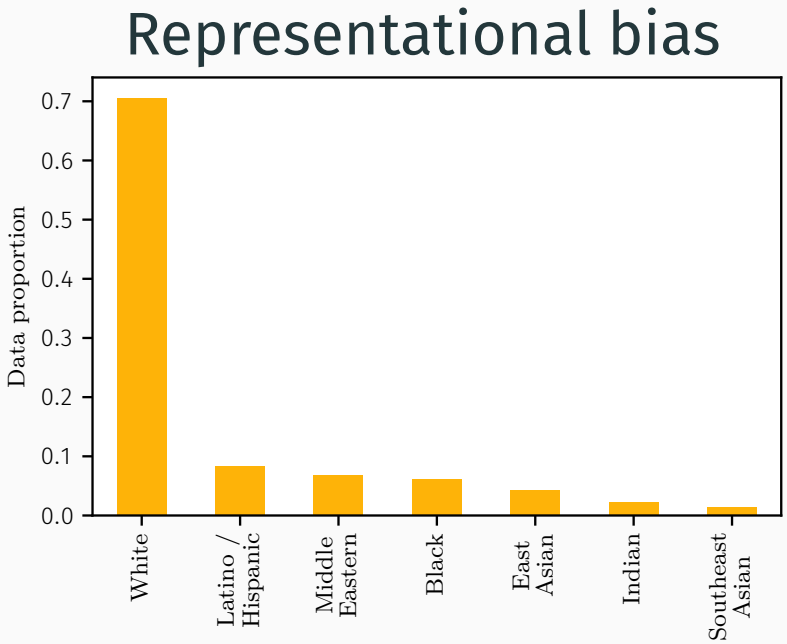


Figure 3: Apparent race distribution in FER+.

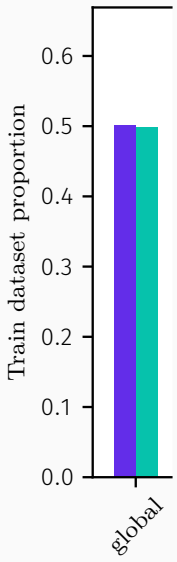


Figure 4: Apparent *per-label* gender distribution in FER+.

⚠ *There are no taxonomies or metrics for these biases.*

⚠ *No studies on their impact.*

Dataset bias, an example

Representational bias

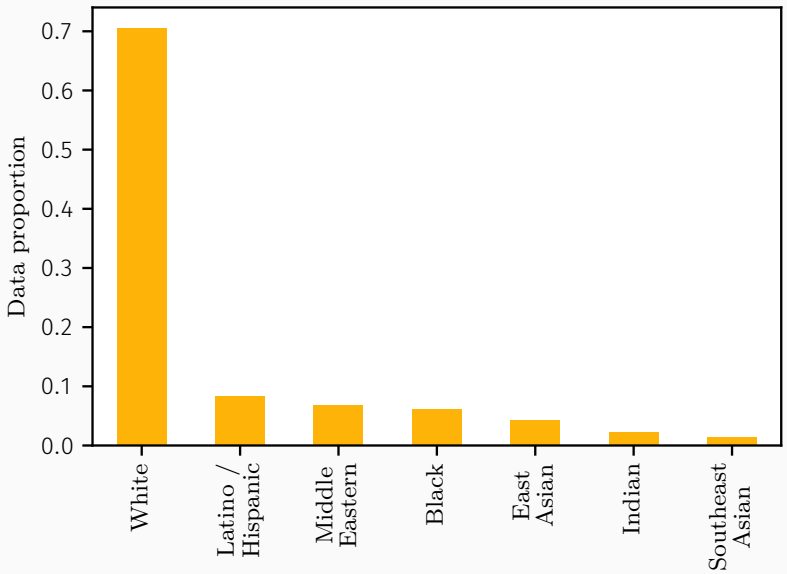


Figure 3: Apparent race distribution in FER+.

Stereotypical bias

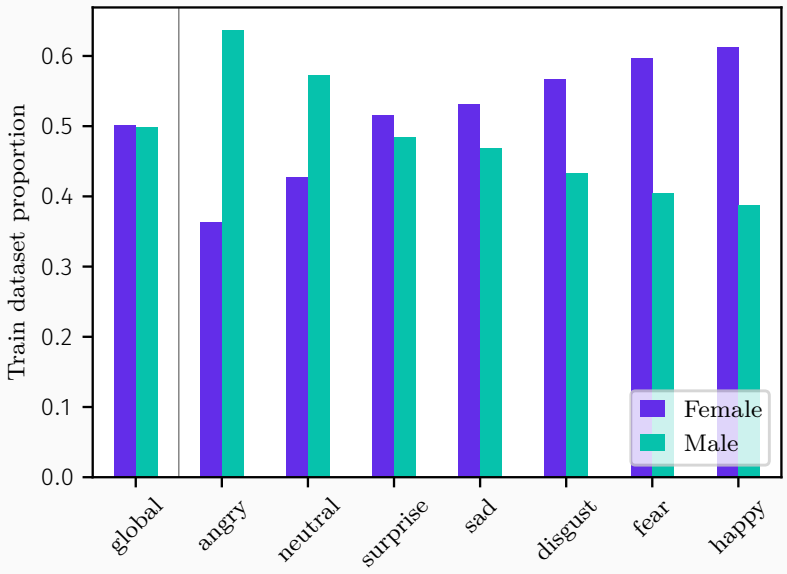


Figure 4: Apparent *per-label* gender distribution in FER+.

⚠ *There are no taxonomies or metrics for these biases.*

⚠ *No studies on their impact.*

Dataset bias, an example

Representational bias

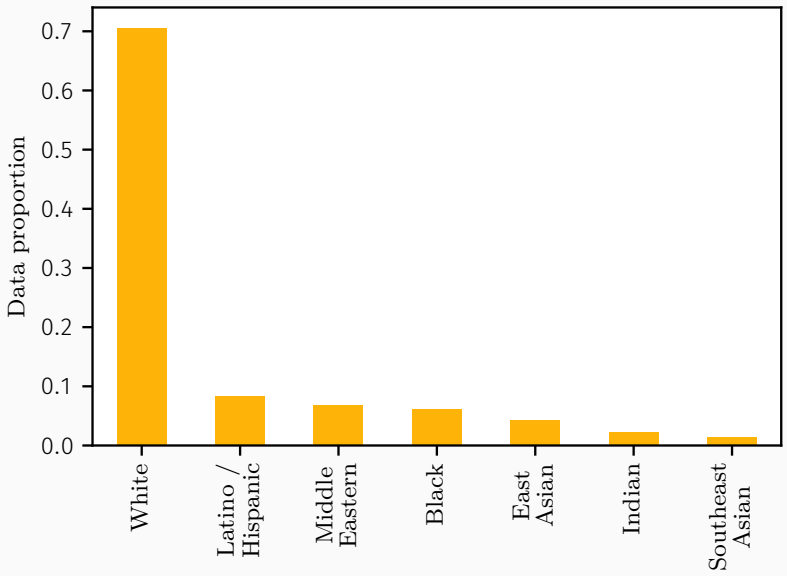


Figure 3: Apparent race distribution in FER+.

Stereotypical bias

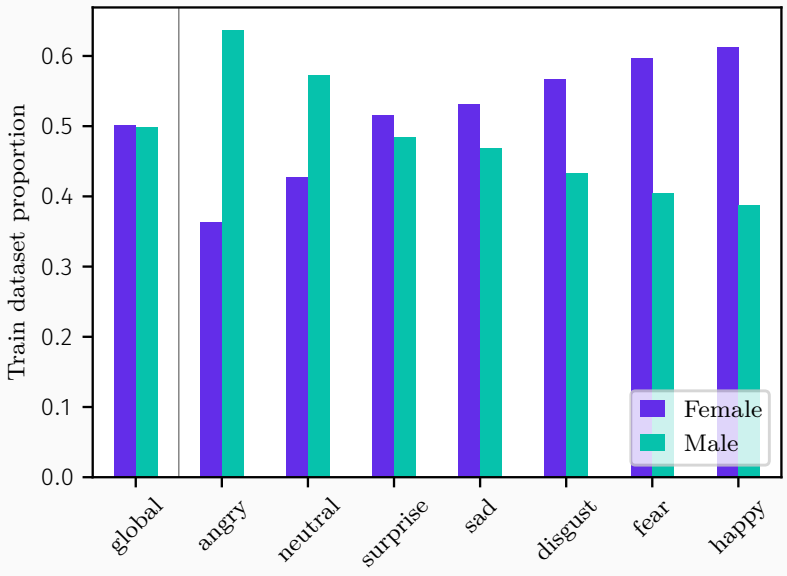


Figure 4: Apparent *per-label* gender distribution in FER+.

⚠ *There are no taxonomies or metrics for these biases.*

⚠ *No studies on their impact.*

Model bias

Model bias metrics:

- More studied[17] than dataset bias.
- Derived from fairness definitions.

(fairness) Demographic Parity[18,19]:

$$P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1) \,. \tag{1}$$

(bias) Disparate Impact[20]:

$$\mathrm{DI} = \frac{P(\hat{Y} = 1 \mid G = 0)}{P(\hat{Y} = 1 \mid G = 1)} \,. \tag{2}$$

[17] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proc. International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering. 2018, pp. 1–7.

[18] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[20] Michael Feldman et al. "Certifying and Removing Disparate Impact". 2015. arXiv: 1412.3756 [cs, stat].

Model bias

Model bias metrics:

- More studied[17] than dataset bias.
- Derived from fairness definitions.

(fairness) Demographic Parity[18,19]:

$$P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1) \,. \tag{1}$$

(bias) Disparate Impact[20]:

$$\mathrm{DI} = \frac{P(\hat{Y} = 1 \mid G = 0)}{P(\hat{Y} = 1 \mid G = 1)} \,. \tag{2}$$

[17] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proc. International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering. 2018, pp. 1–7.

[18] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[20] Michael Feldman et al. "Certifying and Removing Disparate Impact". 2015. arXiv: 1412.3756 [cs, stat].

Model bias

Model bias metrics:

- More studied[17] than dataset bias.
- Derived from fairness definitions.

(fairness) Demographic Parity[18,19]:

$$P(\hat{Y} = 1 \mid G = 0) = P(\hat{Y} = 1 \mid G = 1)\ . \tag{1}$$

(bias) Disparate Impact[20]:

$$\mathrm{DI} = \frac{P(\hat{Y} = 1 \mid G = 0)}{P(\hat{Y} = 1 \mid G = 1)}\ . \tag{2}$$

[17] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proc. International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering. 2018, pp. 1–7.

[18] Cynthia Dwork et al. "Fairness through Awareness". In: *Proc. 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012, pp. 214–226.

[19] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[20] Michael Feldman et al. "Certifying and Removing Disparate Impact". 2015. arXiv: `1412.3756 [cs, stat]`.

Mathematical fairness definitions

- Most model bias metrics[21,22]:
  - Specific for **binary classification** problems.
  - Only **two demographic groups**.
  - **Not symmetric**: *positive* class and a *protected* group.
- Extensions[23] tackle limitations individually.
  - ⚠ *Not complete for multiclass and multi-group problems, like FER.*

[21] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[22] Muhammad Bilal Zafar et al. "Fairness Constraints: Mechanisms for Fair Classification". In: *Proc. 20th International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. 2017, pp. 962–970.

[23] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

Mathematical fairness definitions

- Most model bias metrics[21,22]:
  - Specific for **binary classification** problems.
  - Only **two demographic groups**.
  - **Not symmetric**: *positive* class and a *protected* group.
- Extensions[23] tackle limitations individually.
  - ⚠ *Not complete for multiclass and multi-group problems, like FER.*

---

[21] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proc. 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, pp. 3323–3331.

[22] Muhammad Bilal Zafar et al. "Fairness Constraints: Mechanisms for Fair Classification". In: *Proc. 20th International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. 2017, pp. 962–970.

[23] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

Motivation

- **Dataset bias metrics**
  - There is no clear taxonomy of dataset bias, and no reviews of dataset bias metrics.
- *Better* dataset bias metrics
  - Metrics should be few, explainable and interpretable.
  - Applicable in the absence of demographic information.
- Effect of dataset bias
  - No research on whether the impact of different types of dataset bias is different.
- Bias transference measurement
  - There is no connection between dataset demographic bias metrics and model bias metrics.
  - There are no model bias metrics for multi-group and multiclass classification, such as FER.

Motivation

- Dataset bias metrics
  - There is no clear taxonomy of dataset bias, and no reviews of dataset bias metrics.
- *Better* dataset bias metrics
  - Metrics should be few, explainable and interpretable.
  - Applicable in the absence of demographic information.
- Effect of dataset bias
  - No research on whether the impact of different types of dataset bias is different.
- Bias transference measurement
  - There is no connection between dataset demographic bias metrics and model bias metrics.
  - There are no model bias metrics for multi-group and multiclass classification, such as FER.

Motivation

- Dataset bias metrics
  - There is no clear taxonomy of dataset bias, and no reviews of dataset bias metrics.
- *Better* dataset bias metrics
  - Metrics should be few, explainable and interpretable.
  - Applicable in the absence of demographic information.
- Effect of dataset bias
  - No research on whether the impact of different types of dataset bias is different.
- Bias transference measurement
  - There is no connection between dataset demographic bias metrics and model bias metrics.
  - There are no model bias metrics for multi-group and multiclass classification, such as FER.

Motivation

- Dataset bias metrics
  - There is no clear taxonomy of dataset bias, and no reviews of dataset bias metrics.
- *Better* dataset bias metrics
  - Metrics should be few, explainable and interpretable.
  - Applicable in the absence of demographic information.
- Effect of dataset bias
  - No research on whether the impact of different types of dataset bias is different.
- Bias transference measurement
  - There is no connection between dataset demographic bias metrics and model bias metrics.
  - There are no model bias metrics for multi-group and multiclass classification, such as FER.

Primary objective

- To **investigate the transference of bias from datasets to models**, offering a new perspective that emphasizes the measurement and understanding of dataset bias as the fundamental precursor to model bias.

Sepecific objectives

1. To develop **a taxonomy of dataset bias, as well as a review of its metrics**.
2. To **create better dataset bias metrics**, more interpretable and applicable to datasets without demographic information.
3. To **compare the impact of representational and stereotypical dataset biases** on models.
4. To analyze the **transference of bias** and to develop model bias metrics for multiclass and multigroup problems.

Sepecific objectives

1. To develop **a taxonomy of dataset bias, as well as a review of its metrics**.
2. To **create better dataset bias metrics**, more interpretable and applicable to datasets without demographic information.
3. To **compare the impact of representational and stereotypical dataset biases** on models.
4. To analyze the **transference of bias** and to develop model bias metrics for multiclass and multigroup problems.

Sepecific objectives

1. To develop **a taxonomy of dataset bias, as well as a review of its metrics**.
2. To **create better dataset bias metrics**, more interpretable and applicable to datasets without demographic information.
3. To **compare the impact of representational and stereotypical dataset biases** on models.
4. To analyze the **transference of bias** and to develop model bias metrics for multiclass and multigroup problems.

Sepecific objectives

1. To develop **a taxonomy of dataset bias, as well as a review of its metrics**.
2. To **create better dataset bias metrics**, more interpretable and applicable to datasets without demographic information.
3. To **compare the impact of representational and stereotypical dataset biases** on models.
4. To analyze the **transference of bias** and to develop model bias metrics for multiclass and multigroup problems.

Proposals

Metrics for Dataset Demographic Bias

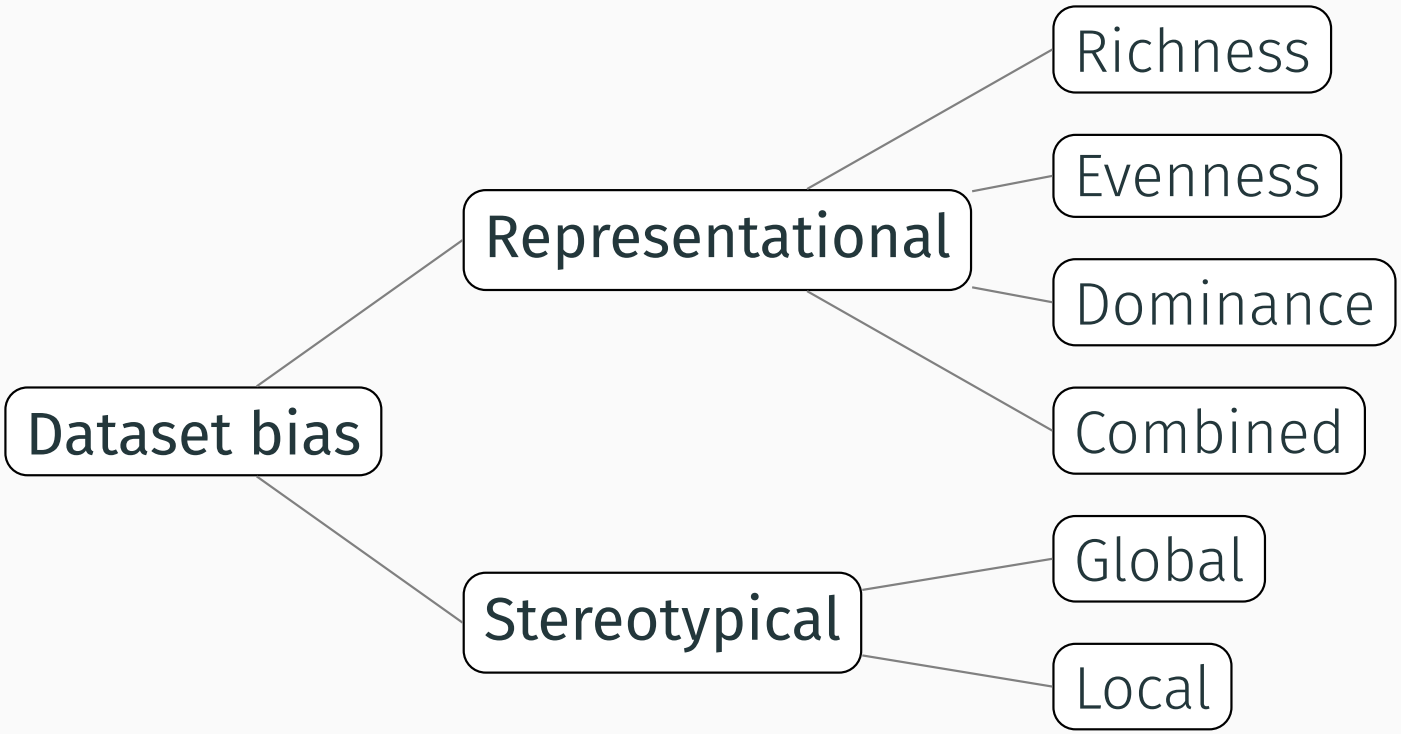
- There are no complete **dataset bias taxonomies** focused on selection bias.
- There are no **standardized metrics** for it.
  - Previous metrics focus on fairness or bias in the model predictions[24].
  - Similar problems in other fields, such as ecology[25,26].

[24] Dana Pessach and Erez Shmueli. "Algorithmic Fairness". 2020. arXiv: 2001.09784 [cs, stat].

[25] R. H. Whittaker. "Vegetation of the Siskiyou Mountains, Oregon and California". In: *Ecological Monographs* 30.3 (1960), pp. 279–338.

[26] 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.

Dataset bias taxonomy

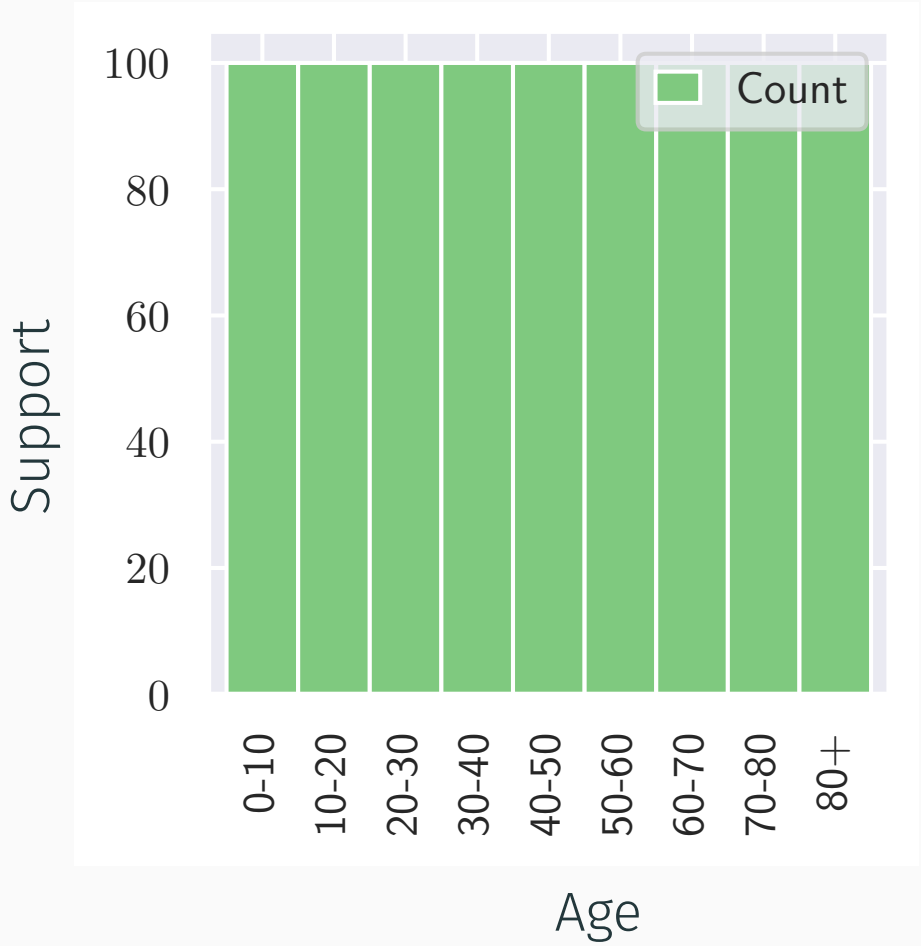


Representational bias

We identify four components:

- **Richness**. Raw number of groups represented in the dataset.
- **Evenness**. Homogeneity of group representation.
- **Dominance**. Population quota of the largest group in the dataset.
- **Combined**. Combinations of richness, evenness, and dominance.

Representational - The ideal

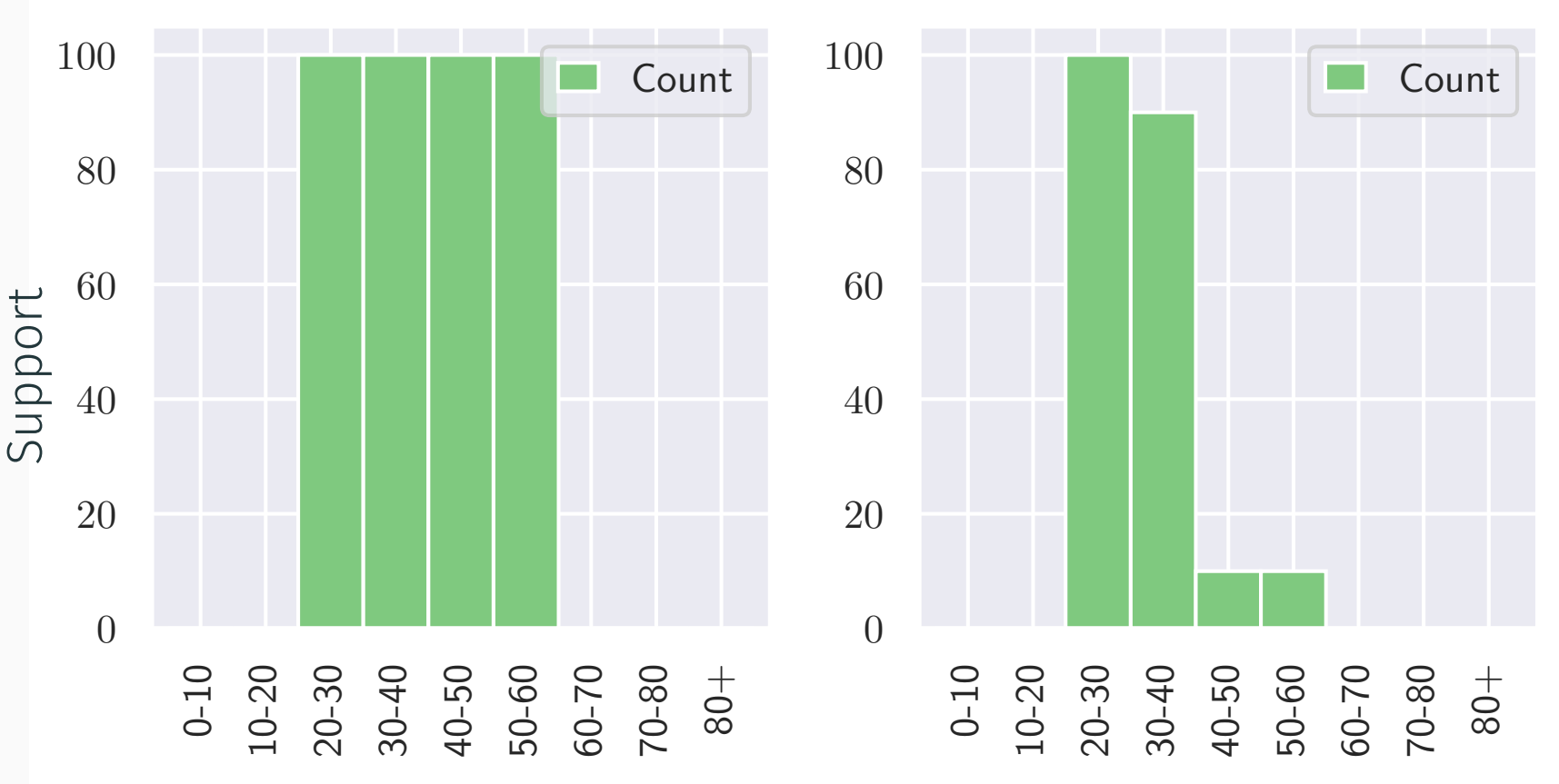


Representationally balanced dataset.

Representational - Richness

Richness. Raw number of groups represented in the dataset.

Representational - Evenness

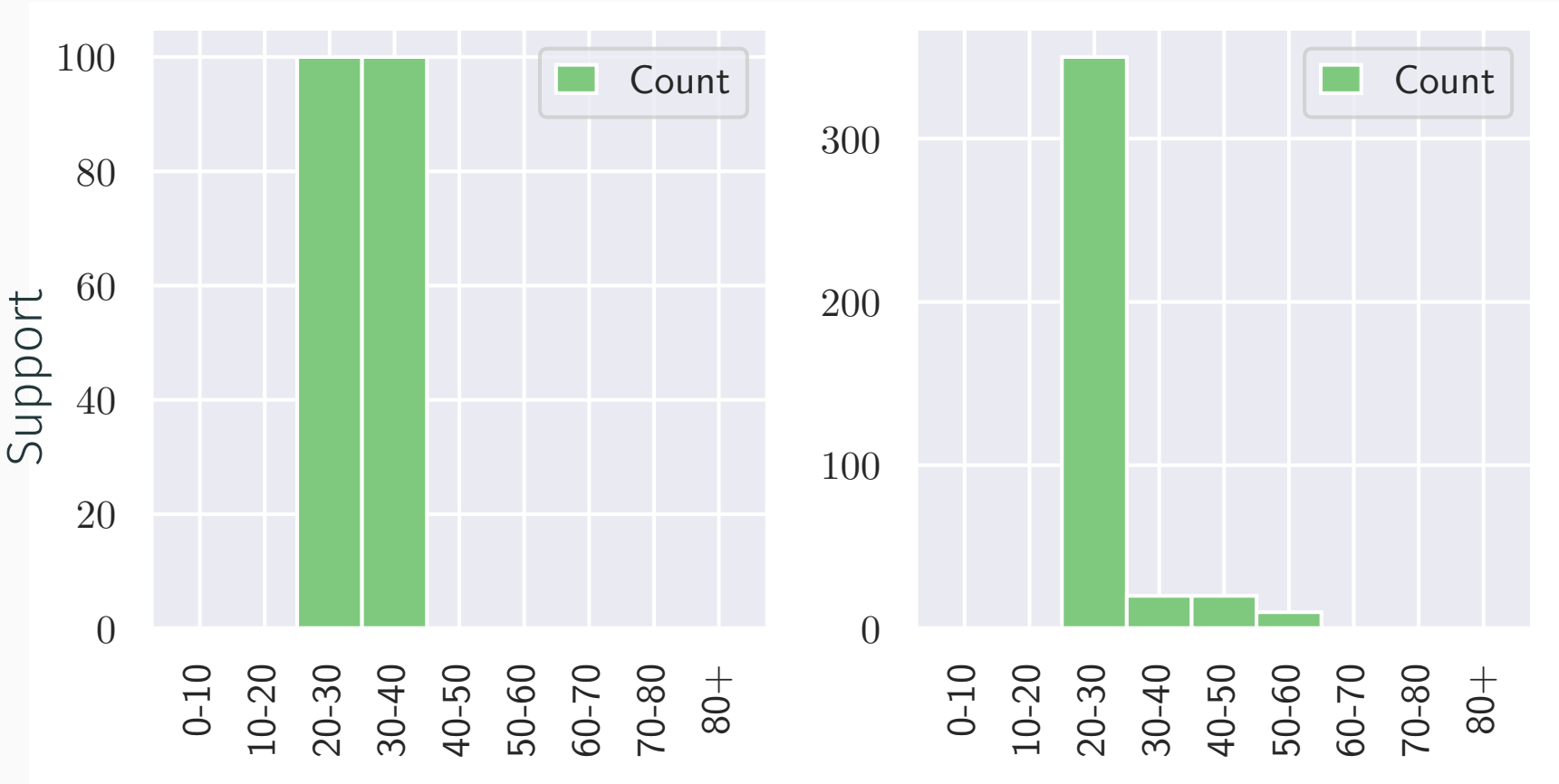


More homogeneous

Less homogeneous

Evenness. Homogeneity of group representation.

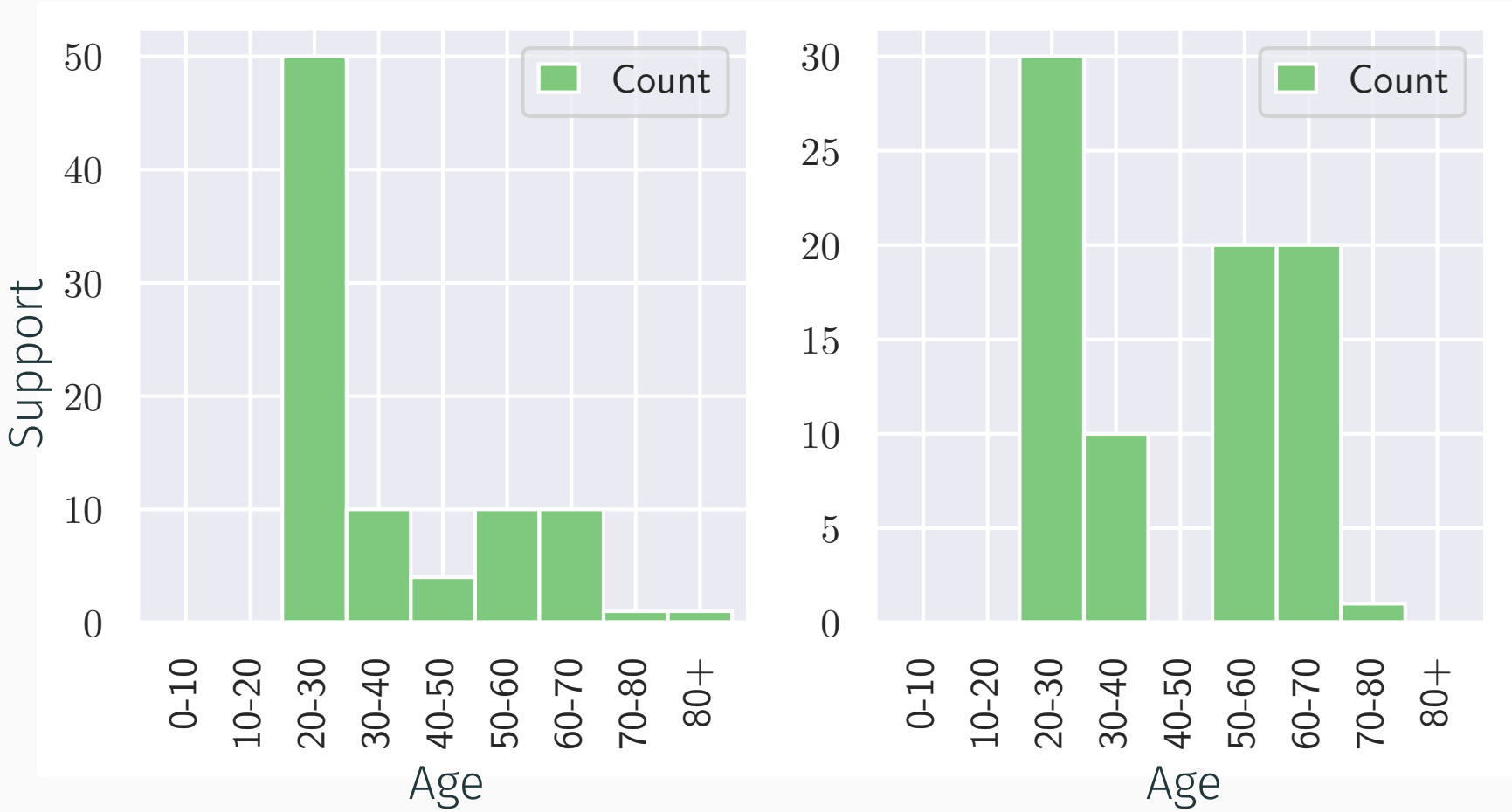
Representational - Dominance



Less dominance **More dominance**

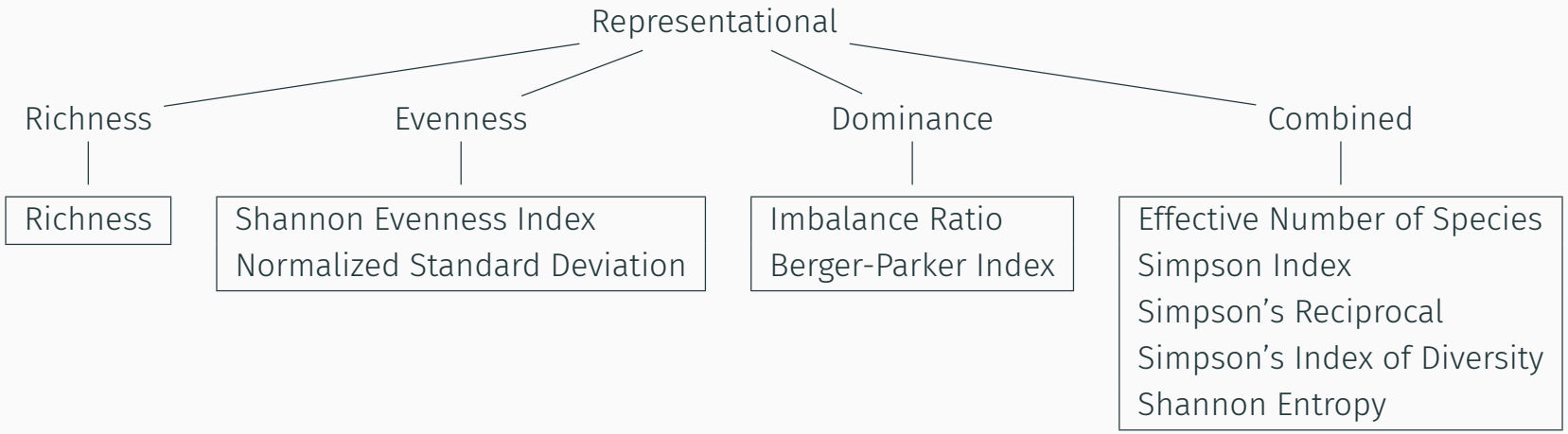
Dominance. Population quota of the largest group in the dataset.

Representational - Combined



Combined. Combinations of richness, evenness, and dominance.

Representational bias metrics

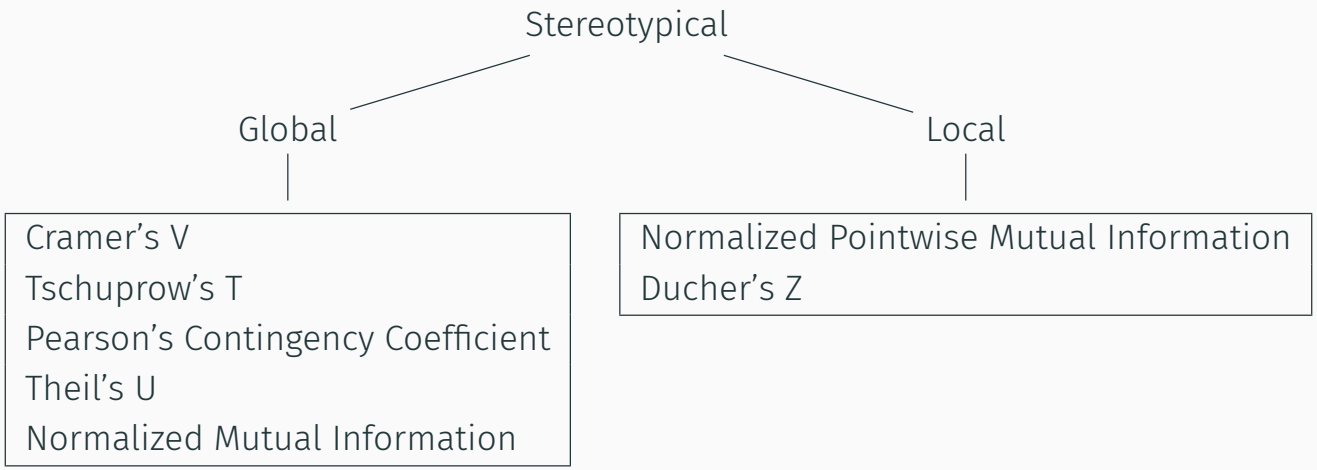


Stereotypical bias

There are two perspectives:

- **Global**.
  - One value for the **whole dataset**.
  - Overall association between a demographic component and the target classes.
- **Local**.
  - One value for **each demographic group and target class**.
  - Over- or underrepresentation of specific combinations of demographic group and target class.

Stereotypical bias metrics



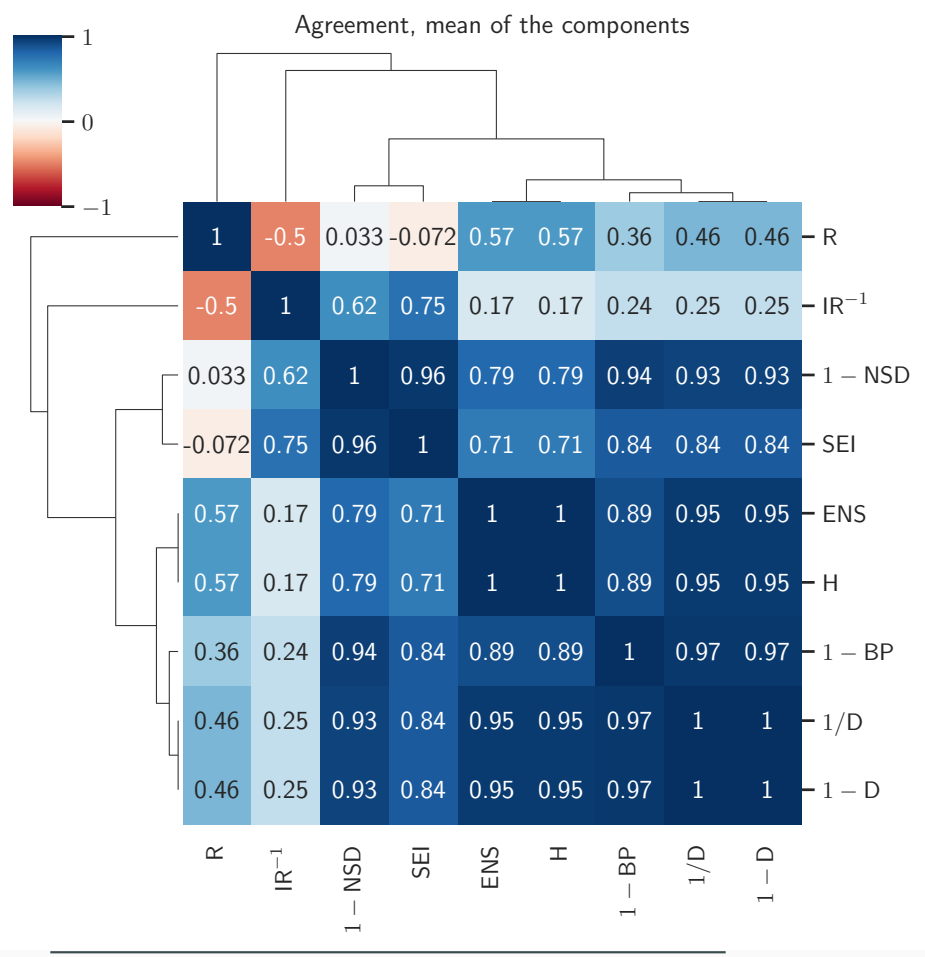
Properties of the dataset bias metrics

Name	Symbol	Type	Subtype	Range	Relation to bias
Richness	R	representational	richness	$[0, \infty)$	inverse
Shannon Evenness Index	SEI	representational	evenness	$[0, 1]$	inverse
Normalized Standard Deviation	NSD	representational	evenness	$[0, 1]$	direct
Imbalance Ratio	IR	representational	dominance	$[1, \infty)$	direct
Berger-Parker Index	BP	representational	dominance	$[1/R, 1]$	direct
Effective Number of Species	ENS	representational	combined	$[1, R]$	inverse
Simpson Index	D	representational	combined	$(0, 1]$	direct
Simpson's Reciprocal	1 / D	representational	combined	$[1, R]$	inverse
Simpson's Index of Diversity	1-D	representational	combined	$[0, 1)$	inverse
Shannon Entropy	H	representational	combined	$[0, \ln R]$	direct
Cramer's V	ϕ_c	stereotypical	global	$[0, 1]$	direct
Tschuprow's T	T	stereotypical	global	$[0, 1]$	direct
Pearson's Contingency Coefficient	C	stereotypical	global	$[0, 1]$	direct
Theil's U	U	stereotypical	global	$[0, 1]$	direct
Normalized Mutual Information	NMI	stereotypical	global	$[0, 1]$	direct
Normalized Mutual Pointwise Information	NPMI	stereotypical	local	$[-1, 1]$	direct
Ducher's Z	Z	stereotypical	local	$[-1, 1]$	direct

1. **Demographic analysis** of the 20 datasets.
   - FairFace[27].
2. **Bias measurement** with all metrics.
   - 3 demographic axis: age, gender, and race.
3. **Correlation analysis** between metrics.
4. **Dataset bias metric selection**.
   - Discard *redundant* metrics, prioritizing *interpretability*.

[27] 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, pp. 1547–1557.

Representational bias metrics coherence



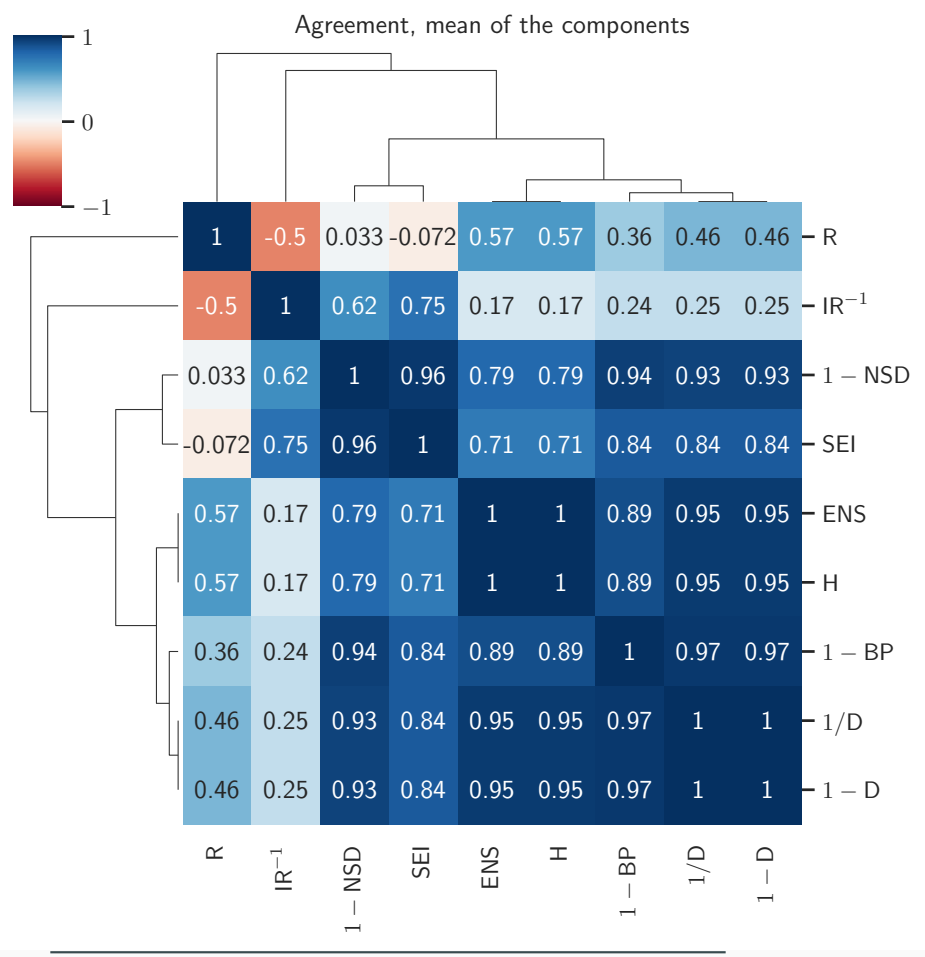
- General: **Effective Number of Species** (ENS) [28]
- Evenness: **Shannon Evenness Index** (SEI) [29]
- Dominance: **Berger-Parker Index** (BP) [30]

[28] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375

[29] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144

[30] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347

Representational bias metrics coherence



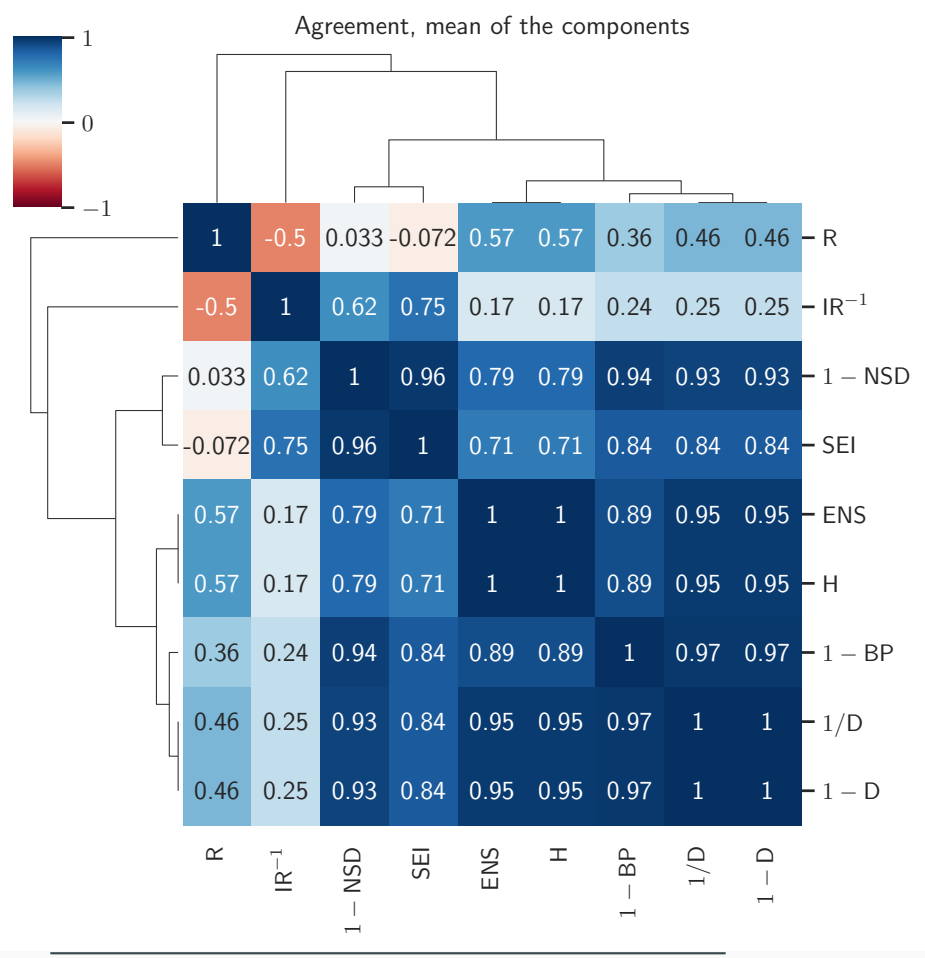
- General: **Effective Number of Species** (ENS) [28]
- Evenness: **Shannon Evenness Index** (SEI) [29]
- Dominance: **Berger-Parker Index** (BP) [30]

[28] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375

[29] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144

[30] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347

Representational bias metrics coherence



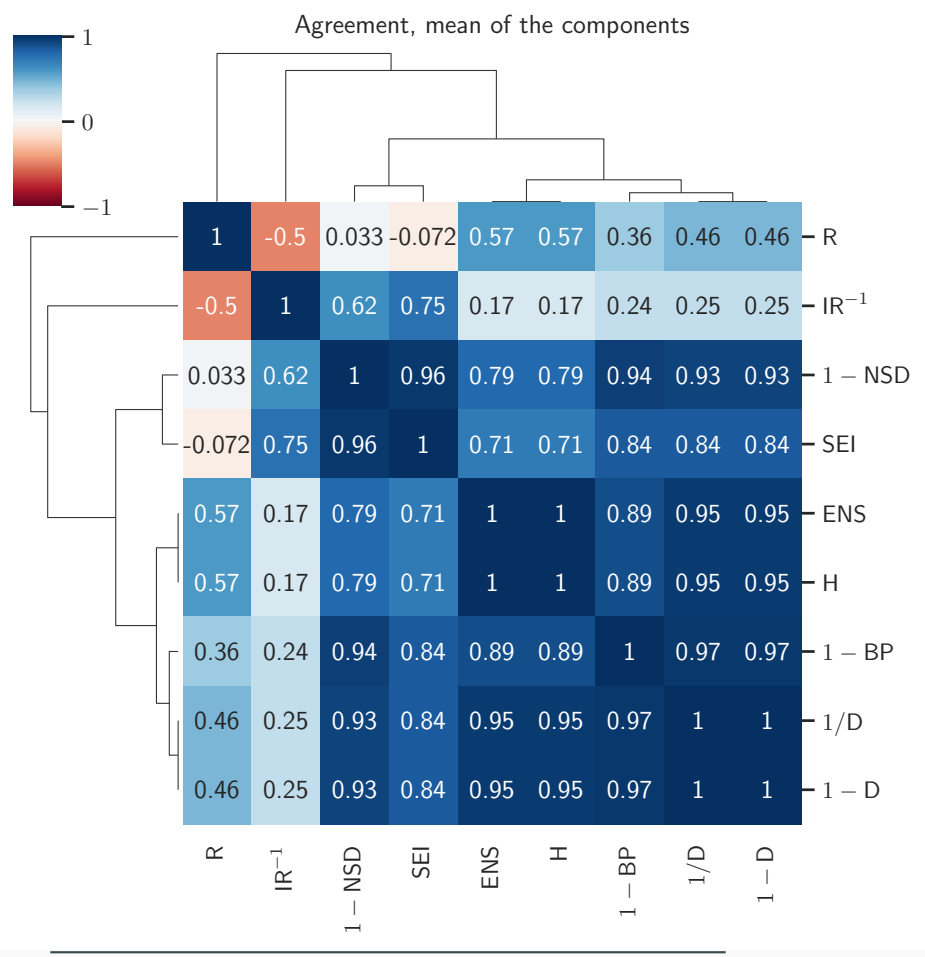
- General: **Effective Number of Species** (ENS) [28]
- Evenness: **Shannon Evenness Index** (SEI) [29]
- Dominance: **Berger-Parker Index** (BP) [30]

[28] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375

[29] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144

[30] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347

Representational bias metrics coherence



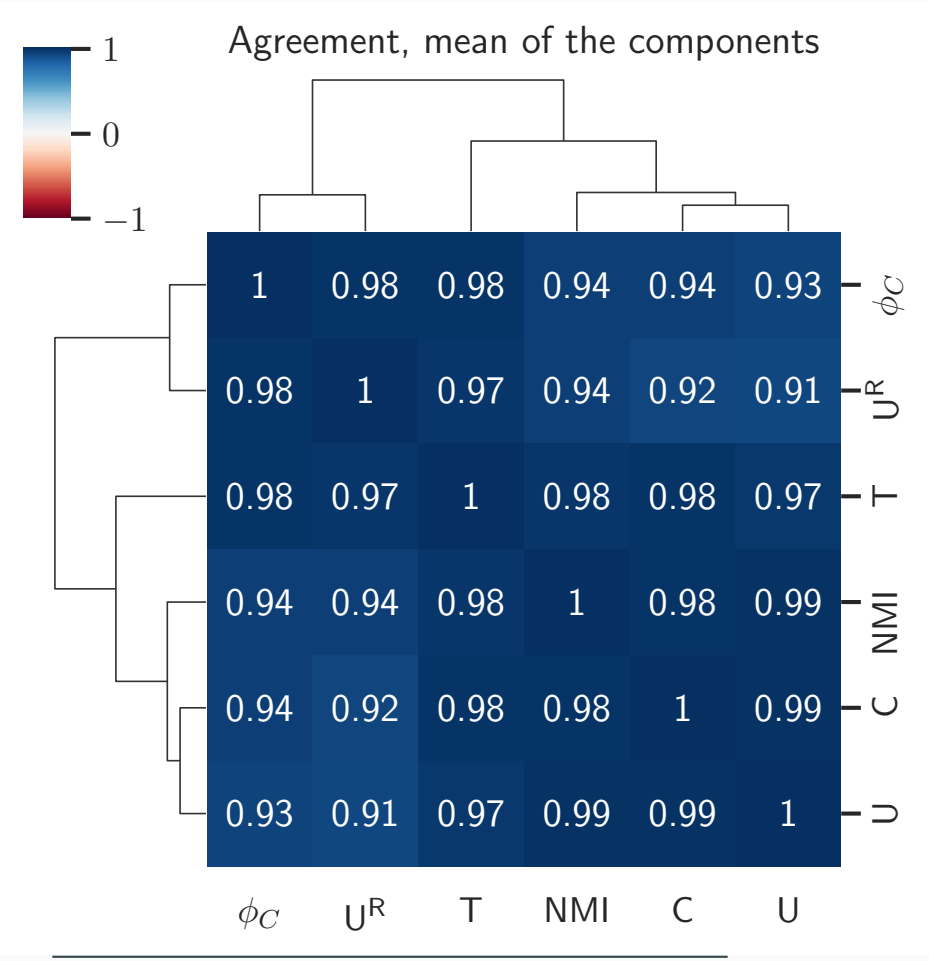
- General: **Effective Number of Species** (ENS) [28]
- Evenness: **Shannon Evenness Index** (SEI) [29]
- Dominance: **Berger-Parker Index** (BP) [30]

[28] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375

[29] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144

[30] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347

Stereotypical bias metrics coherence

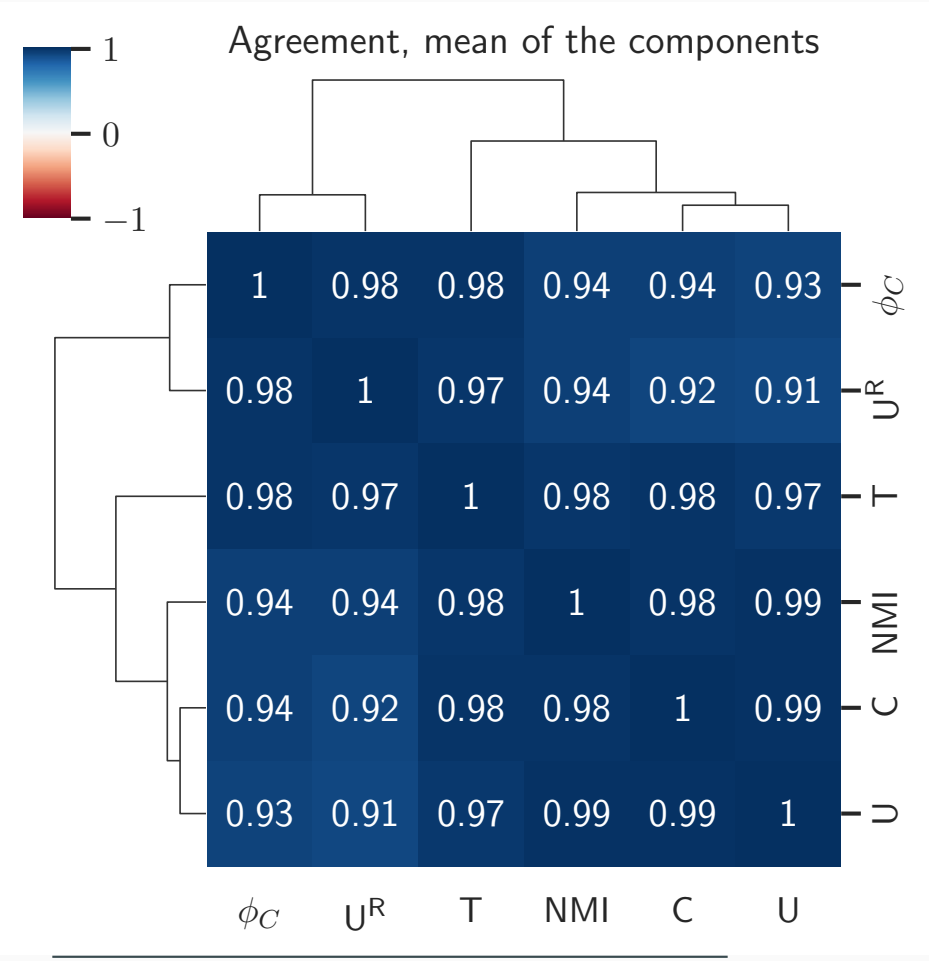


- Global: **Cramer's V** (ϕ_C) [31]
- Local: **Ducher's Z** (Z) [32]

[31] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. 1991, p. 282

[32] 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 (1994), pp. 649–655

Stereotypical bias metrics coherence

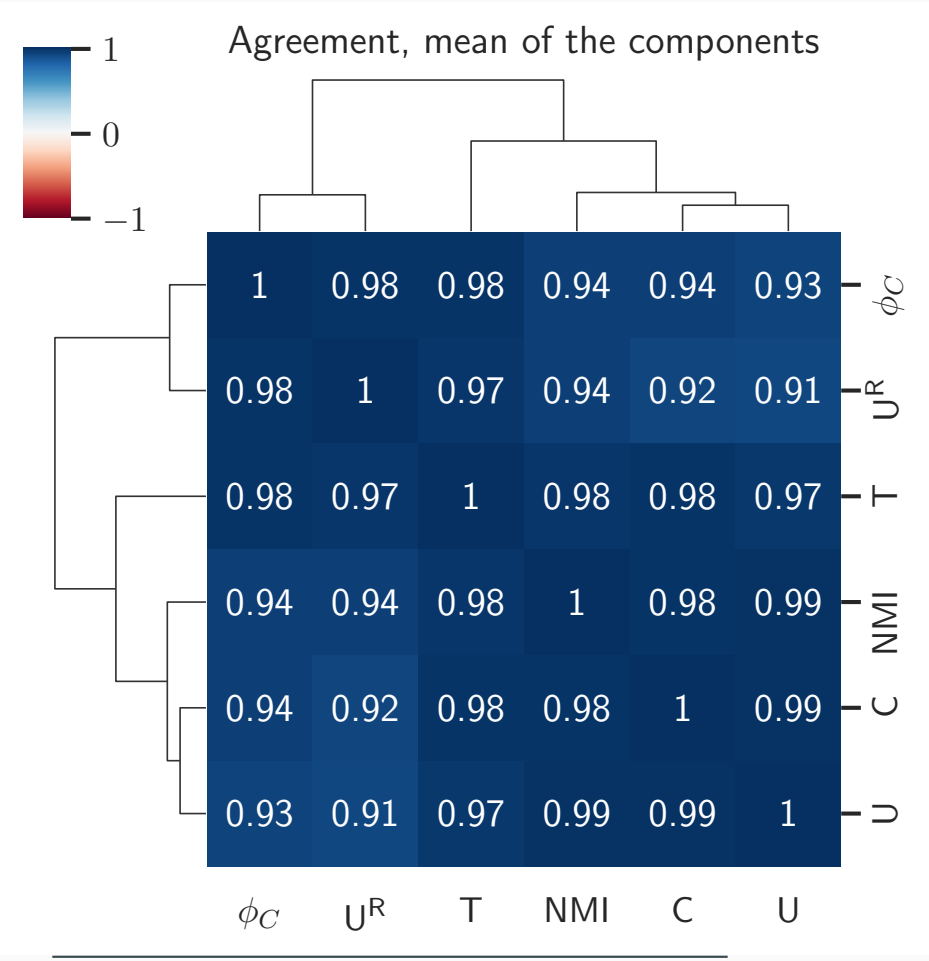


- Global: **Cramer's V** (ϕ_C) [31]
- Local: **Ducher's Z** (Z) [32]

[31] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. 1991, p. 282

[32] 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 (1994), pp. 649–655

Stereotypical bias metrics coherence



- Global: **Cramer's V** (ϕ_C) [31]
- Local: **Ducher's Z** (Z) [32]

[31] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. 1991, p. 282

[32] 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 (1994), pp. 649–655

Measured bias

		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.

Measured bias

		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.

Measured bias

		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.

Local stereotypical bias

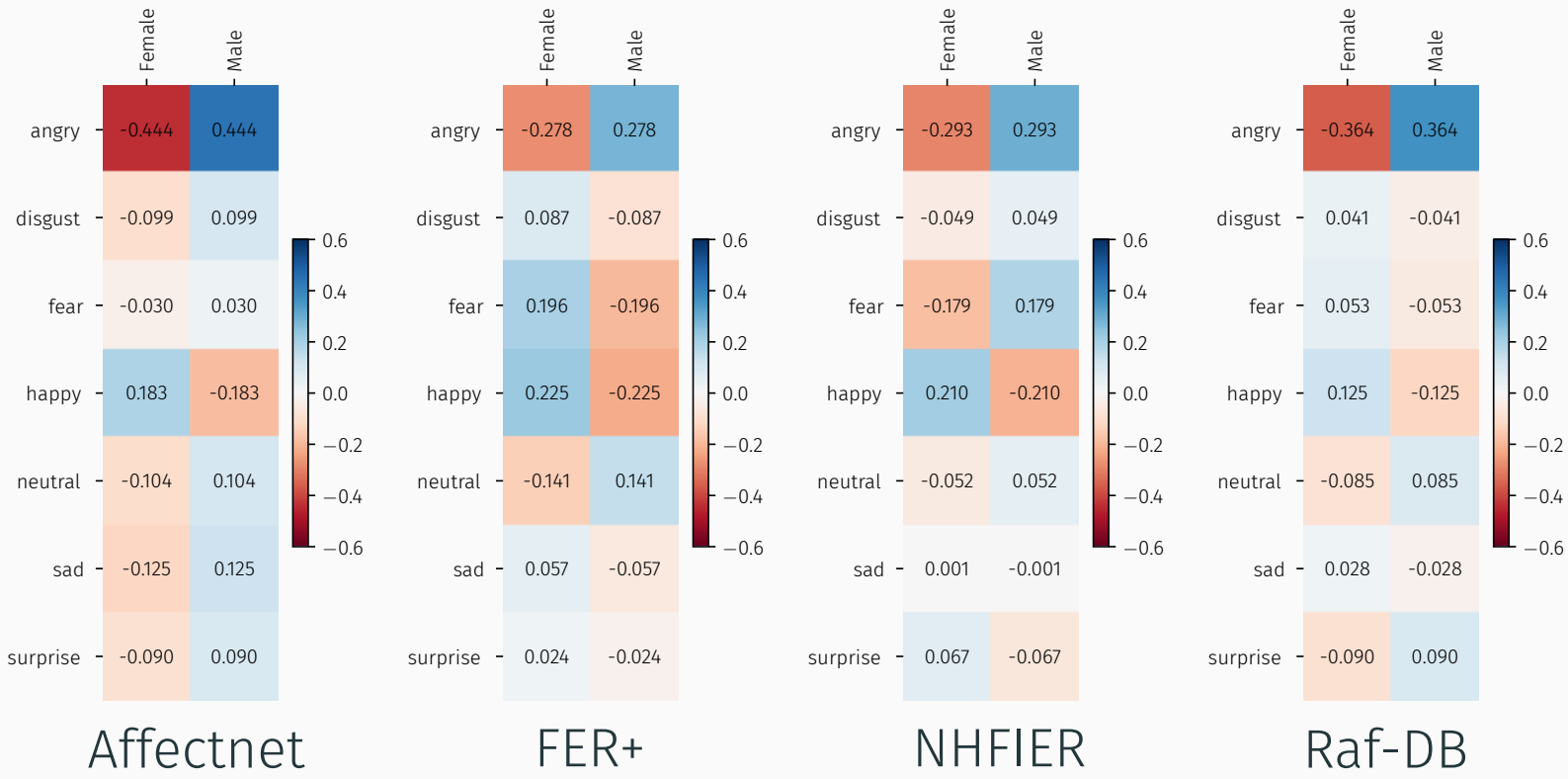


Figure 6: Local stereotypical bias (Ducher's Z) for some ITW-I datasets.

Local stereotypical bias

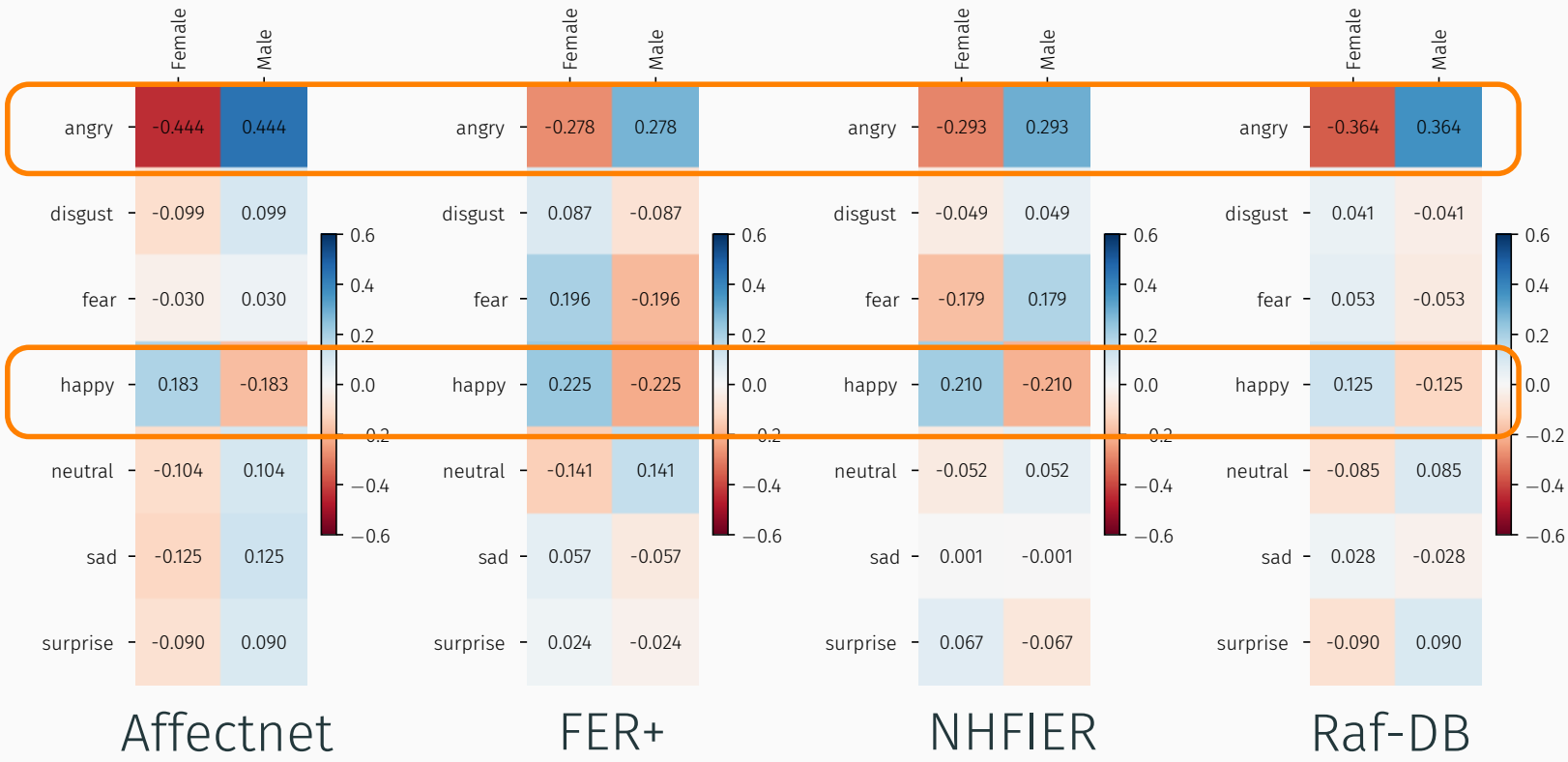
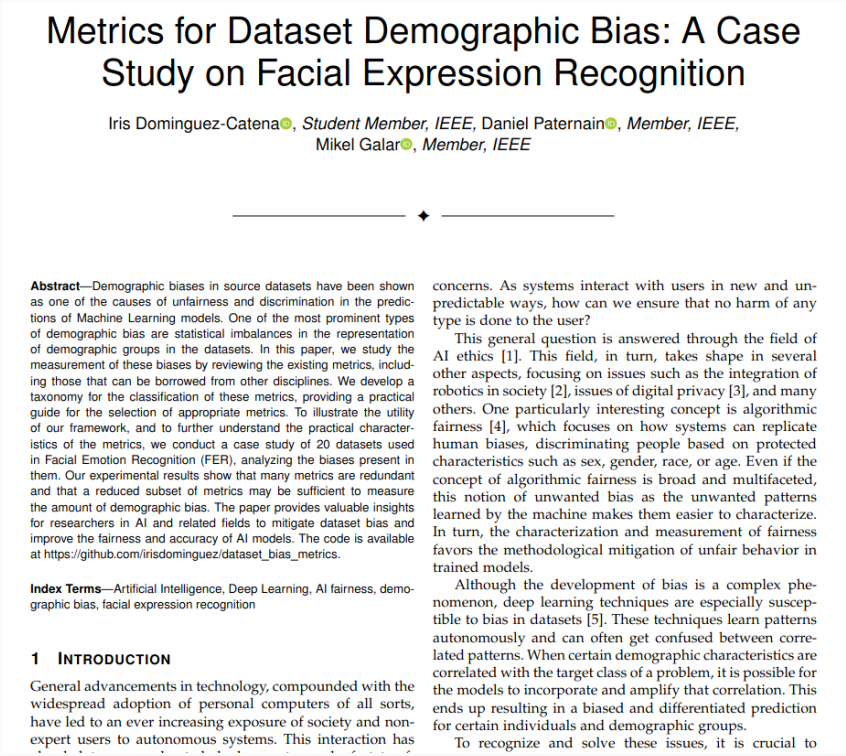


Figure 6: Local stereotypical bias (Ducher's Z) for some ITW-I datasets.

Conclusions

- We have proposed a **taxonomy** of demographic biases in datasets and its metrics.
  - We have included metrics both existing metrics and adapted new ones from other fields.
  - We proposed a selection of metrics based on their interpretability.
- Biases seem to be strongly associated with the **source of data**.
  - Newer datasets seem to correct representational bias at the cost of stereotypical bias.



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, 2nd /197, Impact Factor: 23.6)

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

Code output

Available on Github.

https://github.com/irisdominguez/Dataset_Bias_Metrics

Available as a PyPI package.

https://pypi.org/project/dataset-bias-metrics/

Proposals

Analyzing Bias Through Demographic Comparison of Datasets

- 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[33] and ecology[34,35].

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

[34] 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.

[35] 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.

- 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[33] and ecology[34,35].

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

[34] 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.

[35] 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

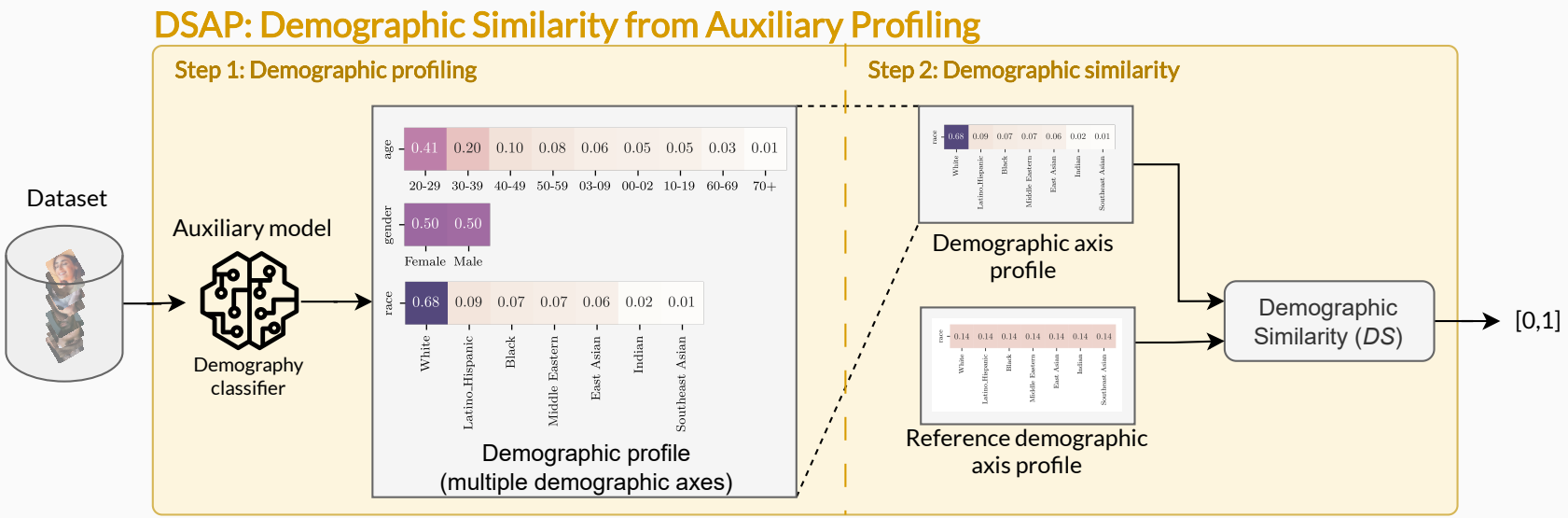
- Dataset bias metrics still lack **interpretability**.
  - Issues of comparability and range disparities.
  - Lack of clear meaning.
- Most datasets lack demographic information.
  - Especially modern ITW datasets.
- There is no way to study the evolution and changes in bias across datasets.
  - Do equally biased dataset represent the same **populations**?
  - Analogous problems in archaeology[33] and ecology[34,35].

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

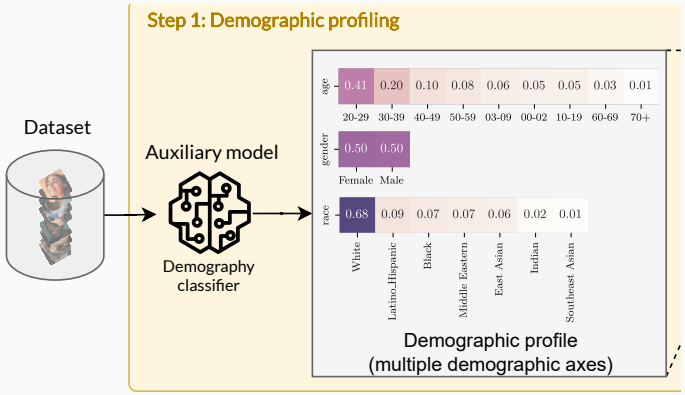
[34] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

[35] 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.

DSAP methodology

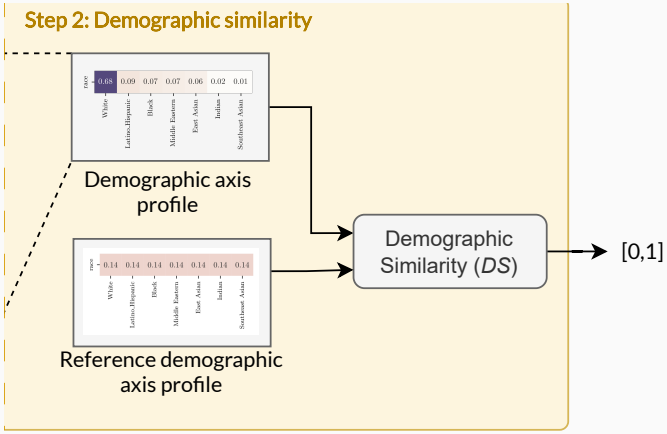


Step 1: Demographic profiling



- Based on **auxiliary models**.
  - Apparent / approximated demographies levels the playing field for different datasets.
- **Demographic profile**: for each demographic axis, proportion of the dataset in each group.
  - Total size varies a lot and is less relevant.

Step 2: Demographic similarity



From multiple similarity indexes[36,37,38,39]:

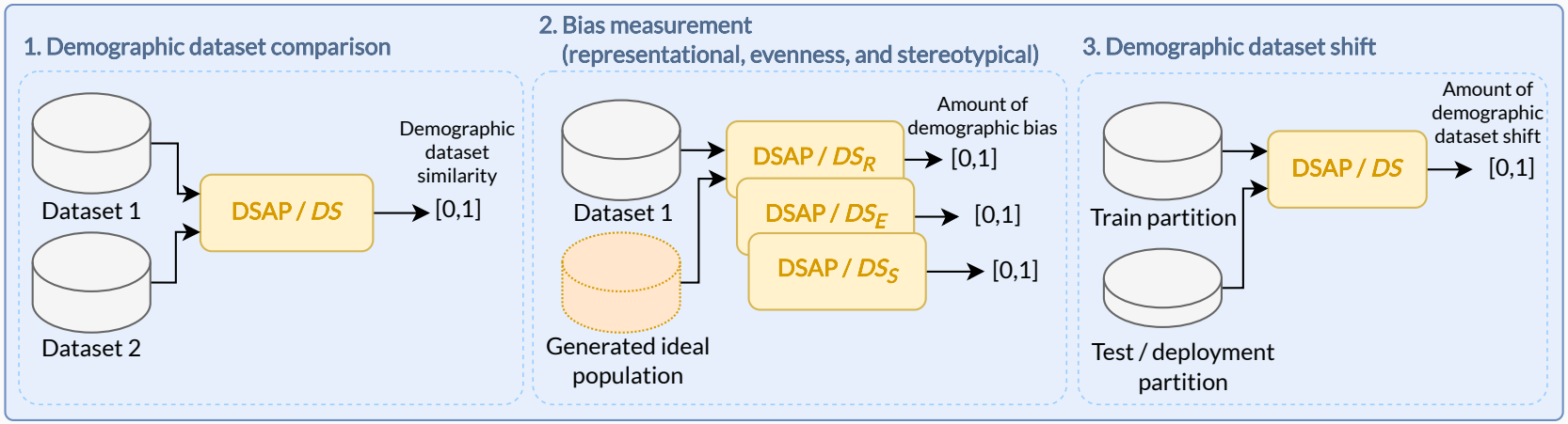
$$DS(X, X') = \mathrm{R}(X, X') = 1 - 0.5 \sum_{g \in G} \left| p_g - p'_g \right| \ . \tag{3}$$

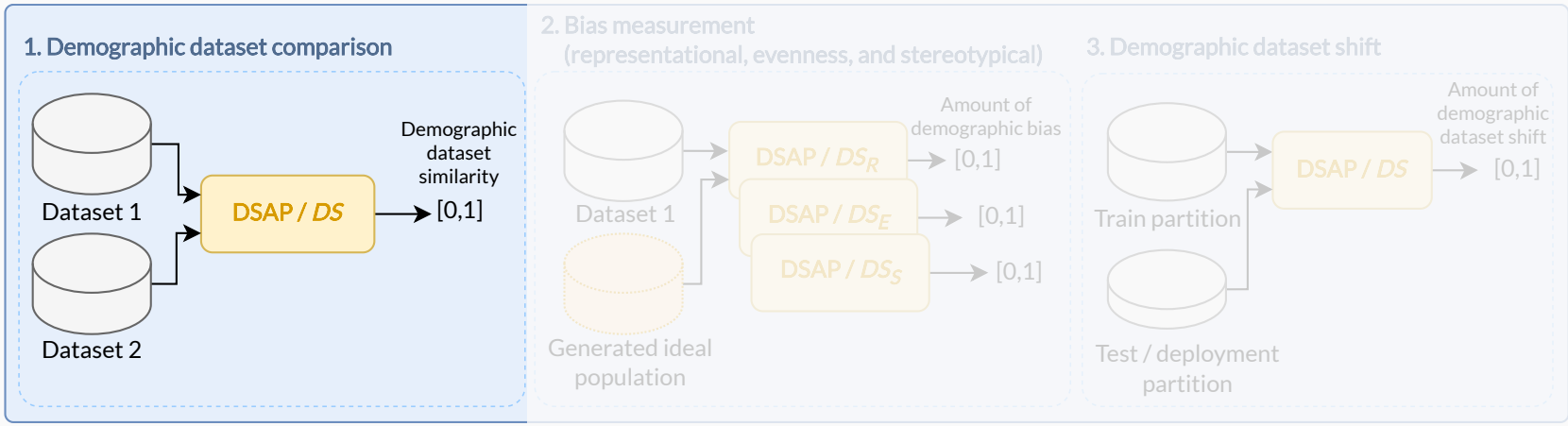
[36] O. Renkonen. *Statistisch-Ökologische Untersuchungen Über Die Terrestrische Käferwelt Der Finnischen Buchmoore*. Vol. 6. Annales Zoologici Societatis Zoologicae-Botanicae Fennicae Vanamo. 1938.

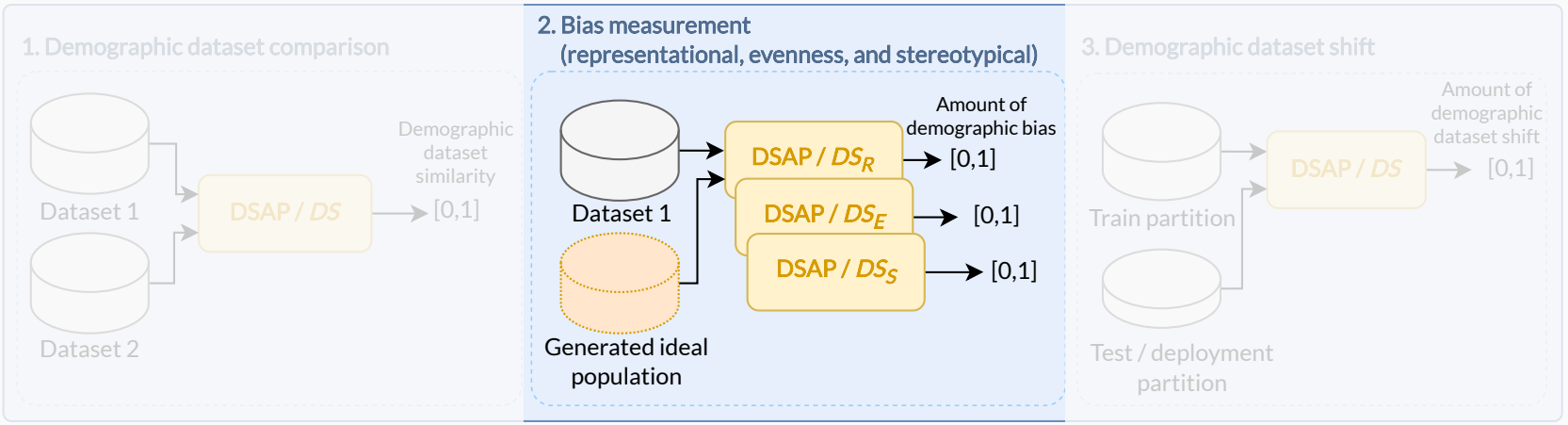
[37] R. H. Whittaker. "A Study of Summer Foliage Insect Communities in the Great Smoky Mountains". In: *Ecological Monographs* 22.1 (1952), pp. 2–44. JSTOR: 1948527.

[38] M Ružička. "Anwendung Mathematisch–Statistischer Methoden in Der Geobotanik (Synthetische Bearbeitung von Aufnahmen)". In: *Biologia, Bratislava* 13 (1958), p. 647.

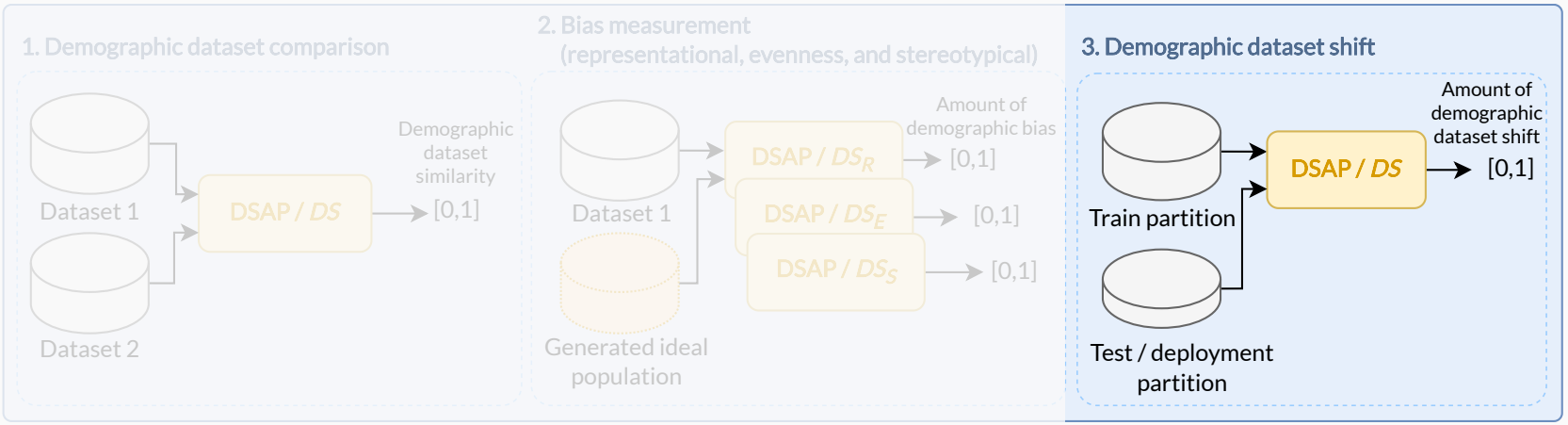
[39] David A. Brock. "Comparison of Community Similarity Indexes". In: *Journal (Water Pollution Control Federation)* 49.12 (1977), pp. 2488–2494. JSTOR: 25039481.


1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]


1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]


1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]

Applications



- **Representational bias** (DS_R).
  - Ideal profile: $p_g^{\text{rep}} = \frac{1}{|G|}$.
  - Metric: $DS_R(X) = DS(X, X^{\text{rep}})$.
- **Evenness** (DS_E).
  - Ideal profile:

$$p_g^{\text{even}} = \begin{cases} \frac{1}{|\{g \in G | p_g > 0\}|} & \text{if } p_g > 0 \\ 0 & \text{otherwise .} \end{cases}$$

  - Metric: $DS_E(X) = DS(X, X^{\text{even}})$.
- **Stereotypical bias** (DS_S).
  - Metric:

$$DS_S(X) = \frac{\sum_{y \in Y} DS(X_y, X_{\hat{y}})}{|Y|} .$$

Experiments:

- **Direct dataset comparison**: Similarity based clustering analysis.
- **Bias measurement**: Comparison with previous dataset bias metrics.
- **Demographic dataset shift**: Train/test comparison in datasets.

Experiment details:

- 20 FER datasets.
- Fairface[40] as auxiliary model.

[40] 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, pp. 1547–1557.

Results: Clustering I

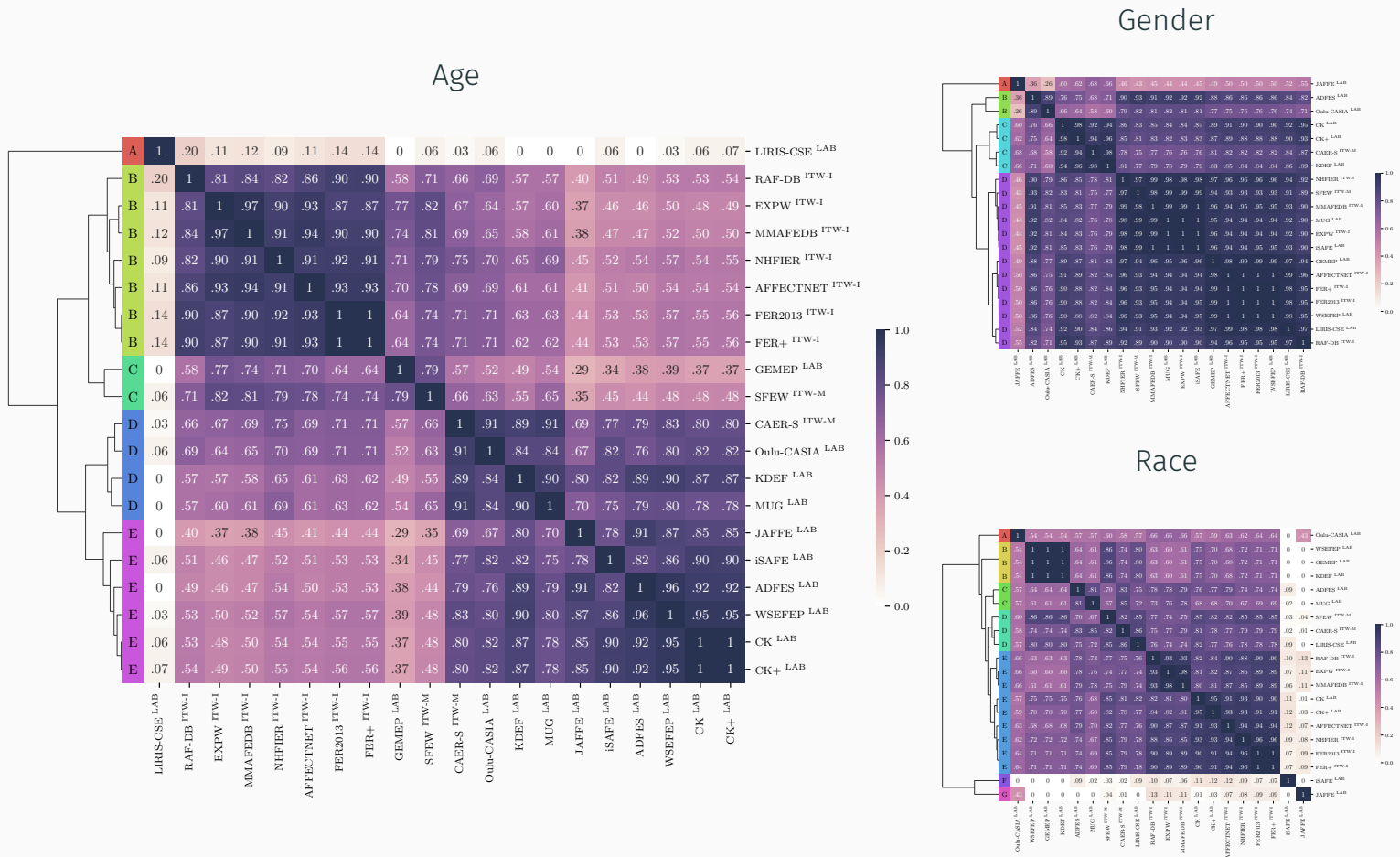


Figure 7: DSAP based comparison of datasets (age, gender and race axis).

Results: Clustering II

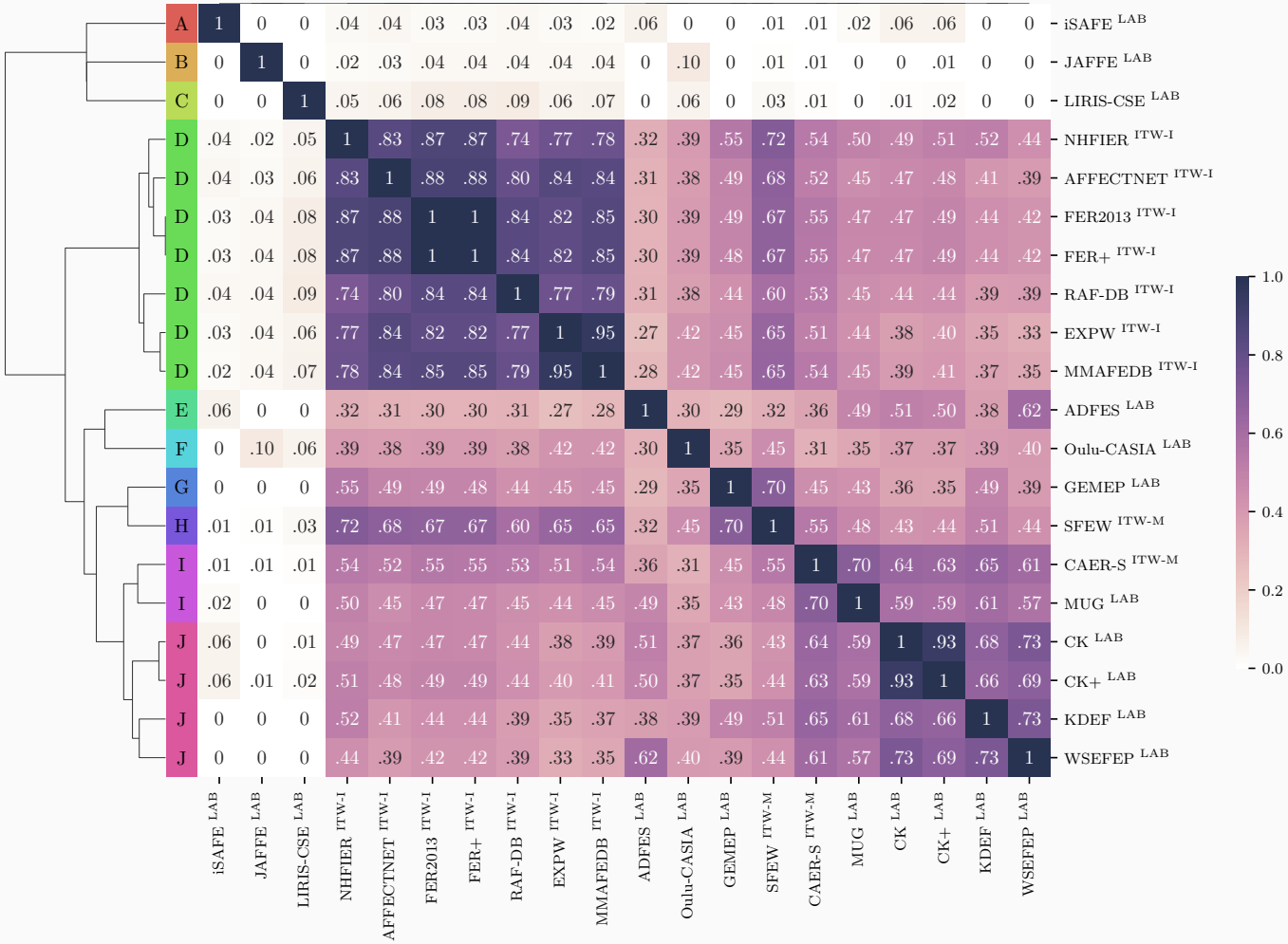


Figure 8: DSAP based comparison of datasets (combination axis, 126 subgroups).

Results: Bias measurement

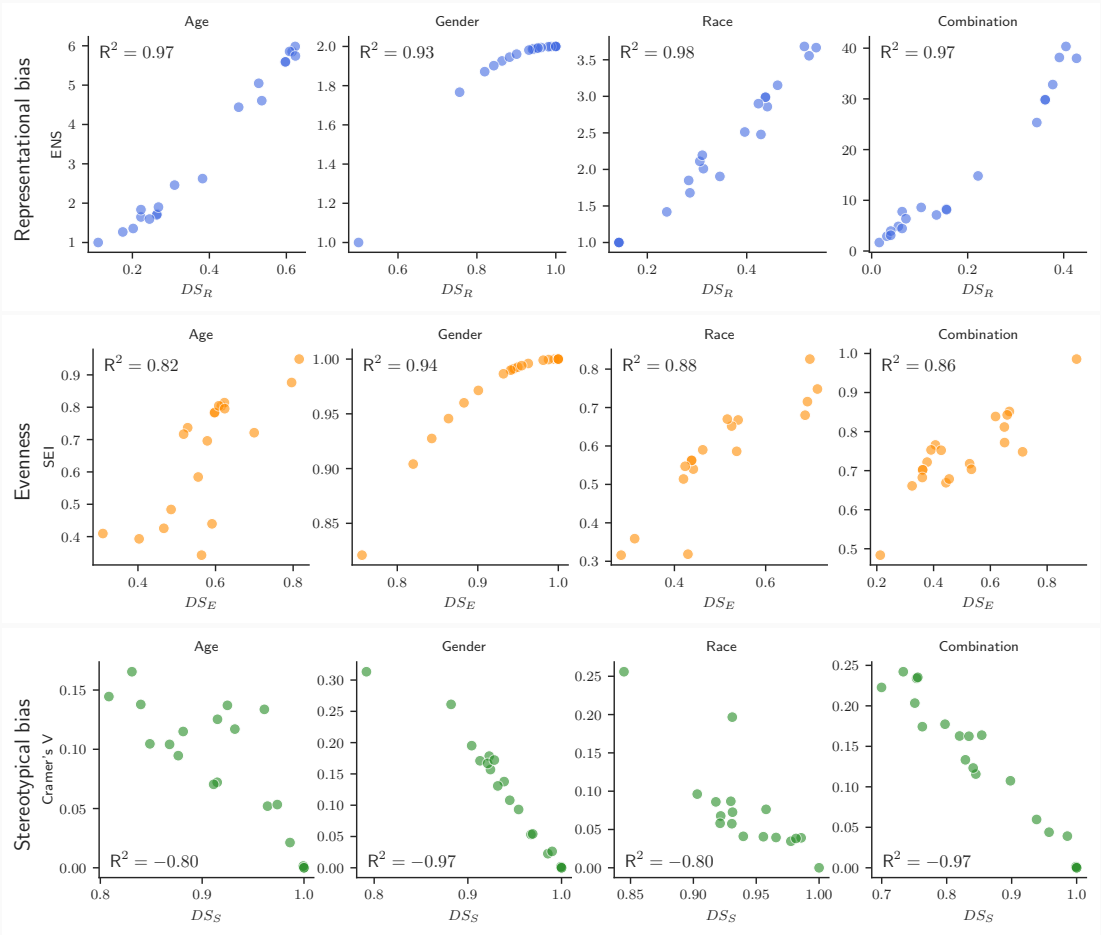


Figure 9: DSAP-based bias metrics (x-axis) compared to their classical counterparts (y-axis).

Results: Dataset shift

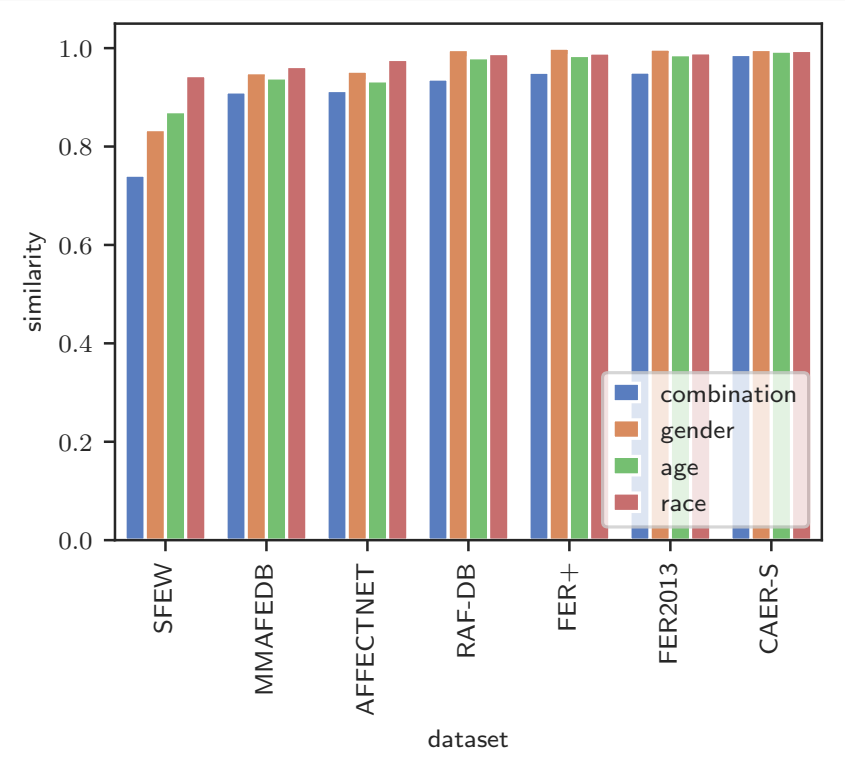
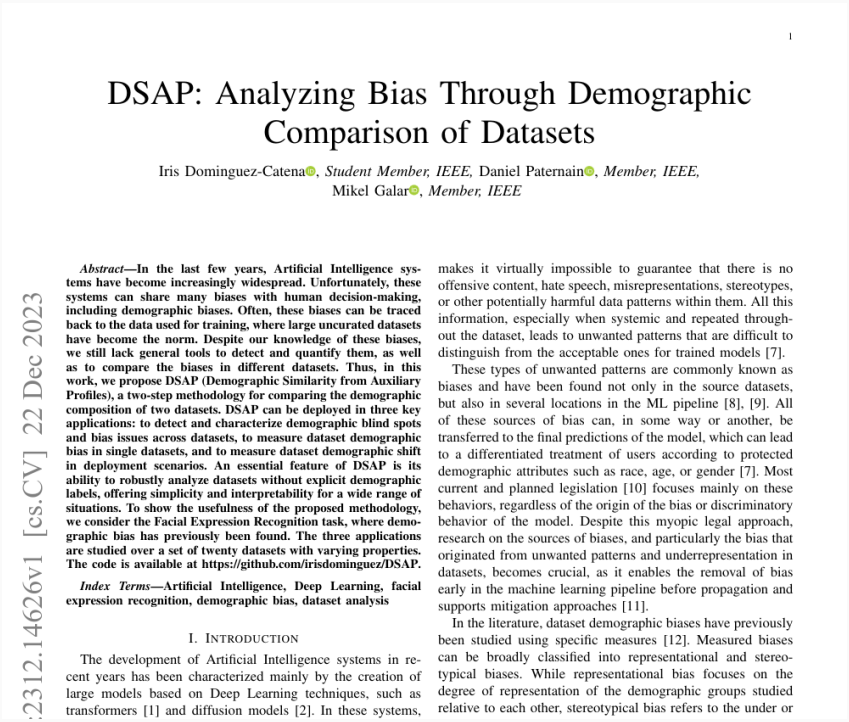


Figure 10: DSAP as a demographic dataset shift detector, comparing train and test partitions.

Conclusion

- DSAP can be applied as a **versatile and interpretable bias metric**, correlating with previously known metrics.
- Confirmed previously found demographic biases in FER datasets, adding novel perspectives:
  - **Dangerous homogeneity** across ITW datasets.
- **Demographic dataset shift** detected in certain datasets:
  - Variations between train and test partitions.



DSAP: Analyzing Bias Through Demographic Comparison of Datasets

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

***Abstract*—In the last few years, Artificial Intelligence systems have become increasingly widespread. Unfortunately, these systems can share many biases with human decision-making, including demographic biases. Often, these biases can be traced back to the data used for training, where large uncurated datasets have become the norm. Despite our knowledge of these biases, we still lack general tools to detect and quantify them, as well as to compare the biases in different datasets. Thus, in this work, we propose DSAP (Demographic Similarity from Auxiliary Profiles), a two-step methodology for comparing the demographic composition of two datasets. DSAP can be deployed in three key applications: to detect and characterize demographic blind spots and bias issues across datasets, to measure dataset demographic bias in single datasets, and to measure dataset demographic shift in deployment scenarios. An essential feature of DSAP is its ability to robustly analyze datasets without explicit demographic labels, offering simplicity and interpretability for a wide range of situations. To show the usefulness of the proposed methodology, we consider the Facial Expression Recognition task, where demographic bias has previously been found. The three applications are studied over a set of twenty datasets with varying properties. The code is available at https://github.com/irisdominguez/DSAP.**

***Index Terms*—Artificial Intelligence, Deep Learning, facial expression recognition, demographic bias, dataset analysis**

I. Introduction

The development of Artificial Intelligence systems in recent years has been characterized mainly by the creation of large models based on Deep Learning techniques, such as transformers [1] and diffusion models [2]. In these systems, makes it virtually impossible to guarantee that there is no offensive content, hate speech, misrepresentations, stereotypes, or other potentially harmful data patterns within them. All this information, especially when systemic and repeated throughout the dataset, leads to unwanted patterns that are difficult to distinguish from the acceptable ones for trained models [7].

These types of unwanted patterns are commonly known as biases and have been found not only in the source datasets, but also in several locations in the ML pipeline [8], [9]. All of these sources of bias can, in some way or another, be transferred to the final predictions of the model, which can lead to a differentiated treatment of users according to protected demographic attributes such as race, age, or gender [7]. Most current and planned legislation [10] focuses mainly on these behaviors, regardless of the origin of the bias or discriminatory behavior of the model. Despite this myopic legal approach, research on the sources of biases, and particularly the bias that originated from unwanted patterns and underrepresentation in datasets, becomes crucial, as it enables the removal of bias early in the machine learning pipeline before propagation and supports mitigation approaches [11].

In the literature, dataset demographic biases have previously been studied using specific measures [12]. Measured biases can be broadly classified into representational and stereotypical biases. While representational bias focuses on the degree of representation of the demographic groups studied relative to each other, stereotypical bias refers to the under or

Under review in

Information Fusion

(Q1, Impact Factor: 14.7)

https://doi.org/10.48550/arXiv.2312.14626

Code output

Available on Github.

https://github.com/irisdominguez/DSAP

Proposals

Representational vs. Stereotypical Bias Transference

- We have identified and measured **two types of dataset bias**.
  - Increase in **gender stereotypical** bias in ITW-I datasets.
- But there is **no research on the transference** of the different types of dataset bias to the model predictions.

- We have identified and measured **two types of dataset bias**.
  - Increase in **gender stereotypical** bias in ITW-I datasets.
- But there is **no research on the transference** of the different types of dataset bias to the model predictions.

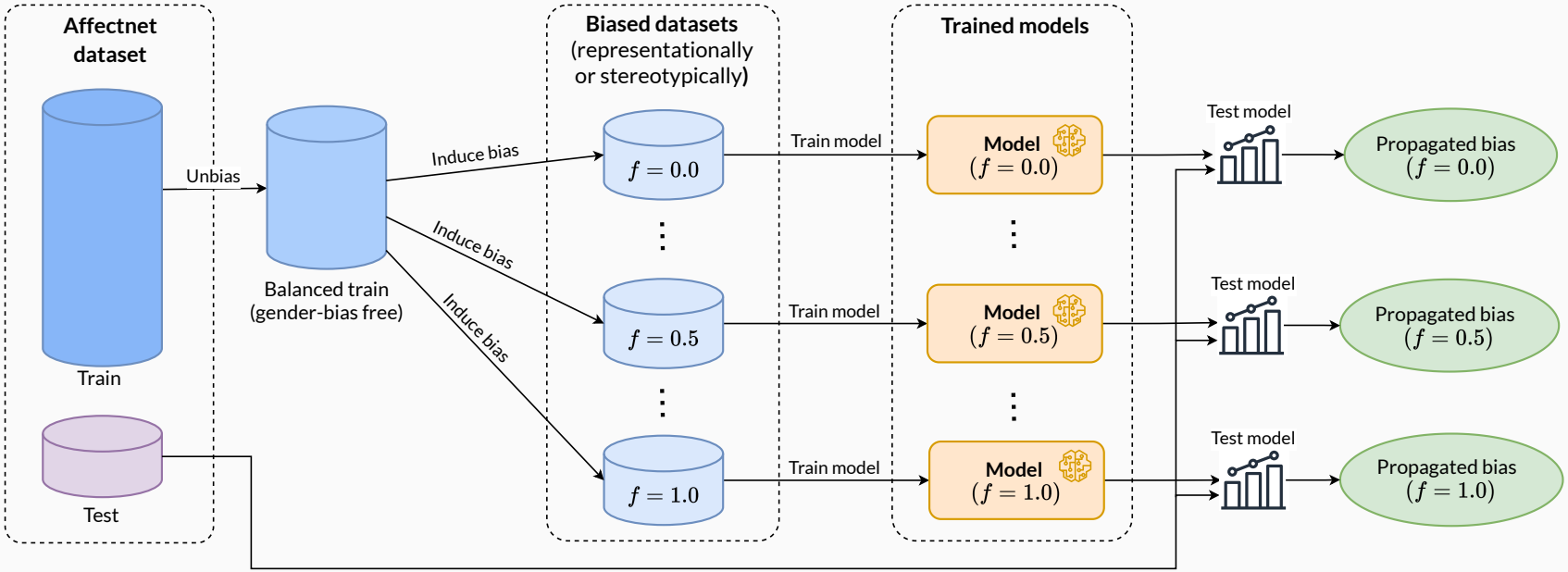


Figure 11: Summary of the methodology

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	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%

Experimental framework

- **Original dataset:** Affectnet[41].
- **Network:** pretrained ResNet50[42] and ViT-Base[43].
- **Training:** 20 epochs, 1cycle policy, learning rate $1e^{-4}$.
- **Bias proportions:** $[0\%, 10\%, \ldots, 100\%]$.
  - 3 repetitions per configuration.

[41] 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.

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

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

Results on propagation

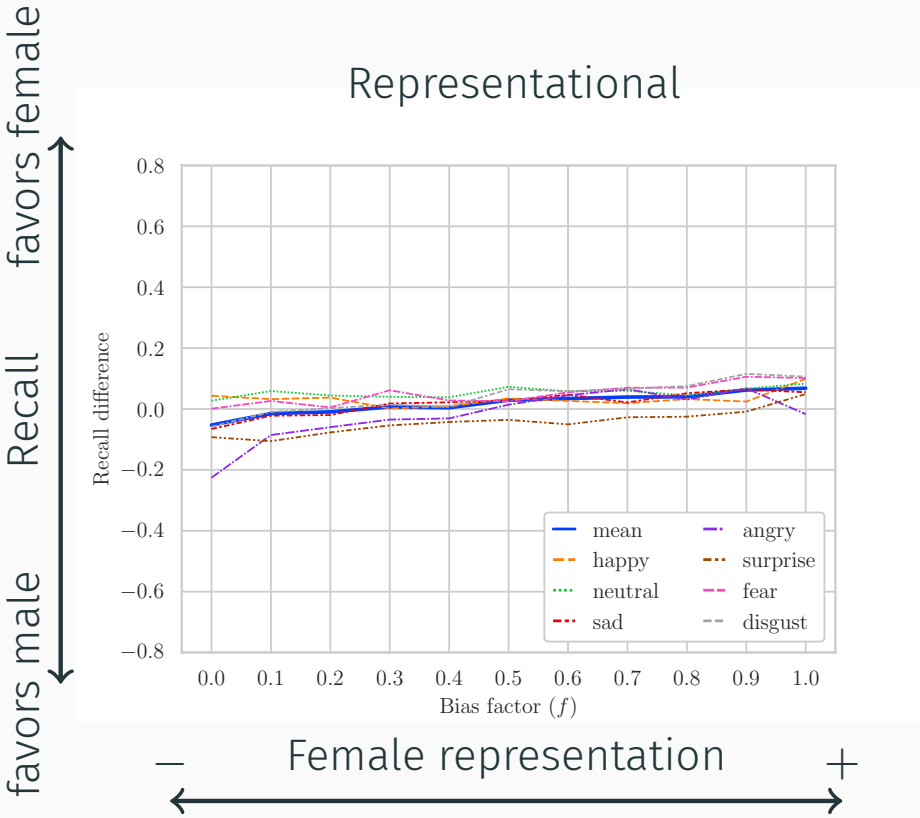


Figure 12: Recall difference (female recall minus male recall).

Results on propagation

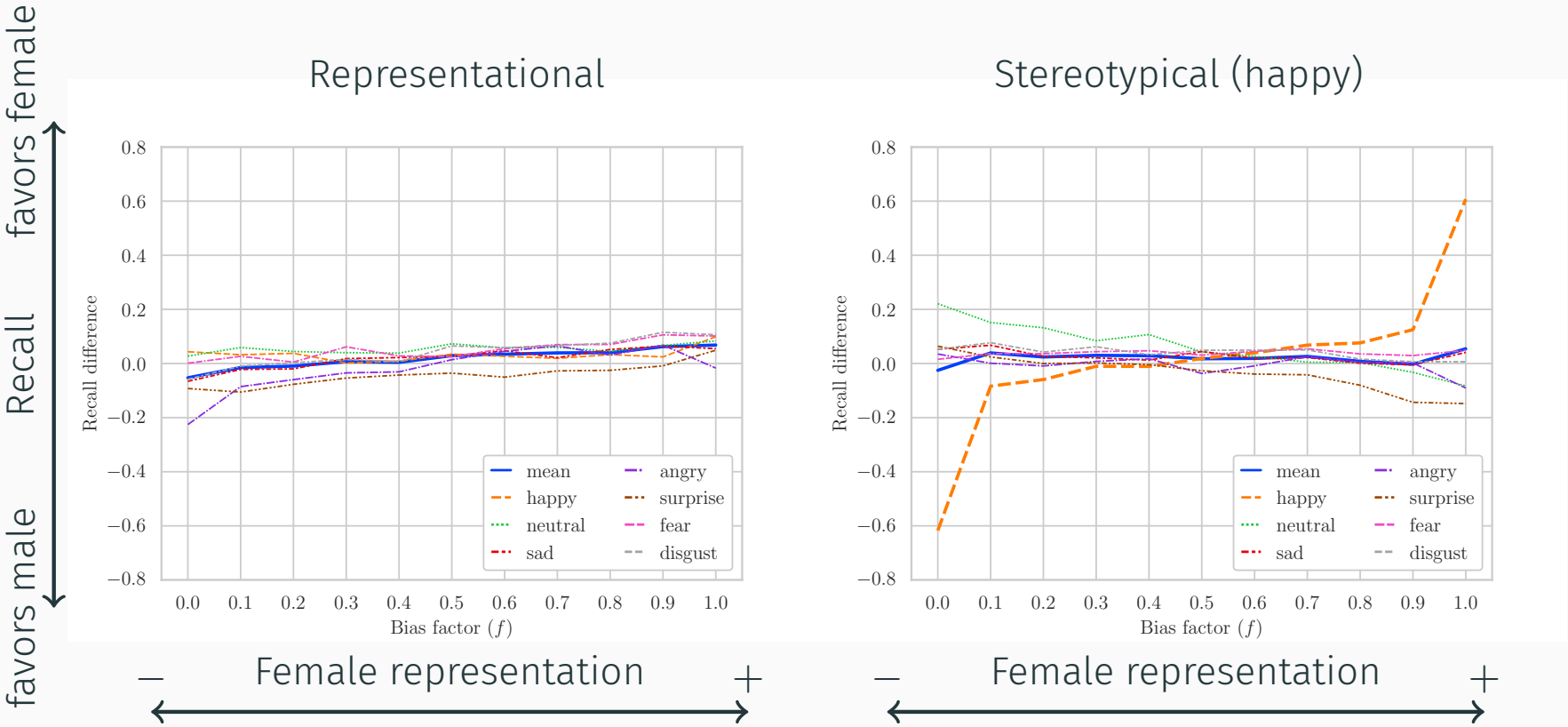


Figure 12: Recall difference (female recall minus male recall).

Results on propagation

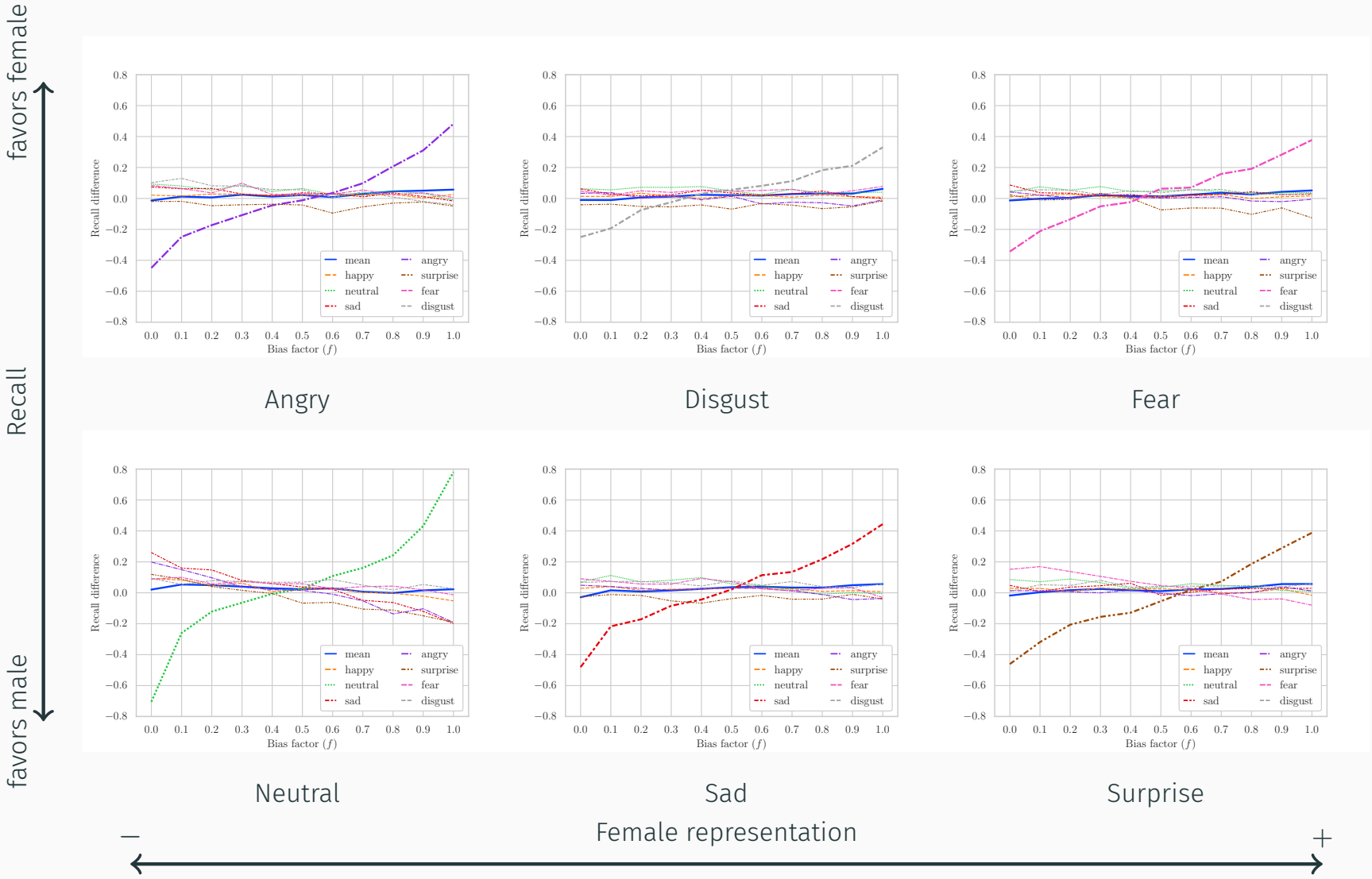
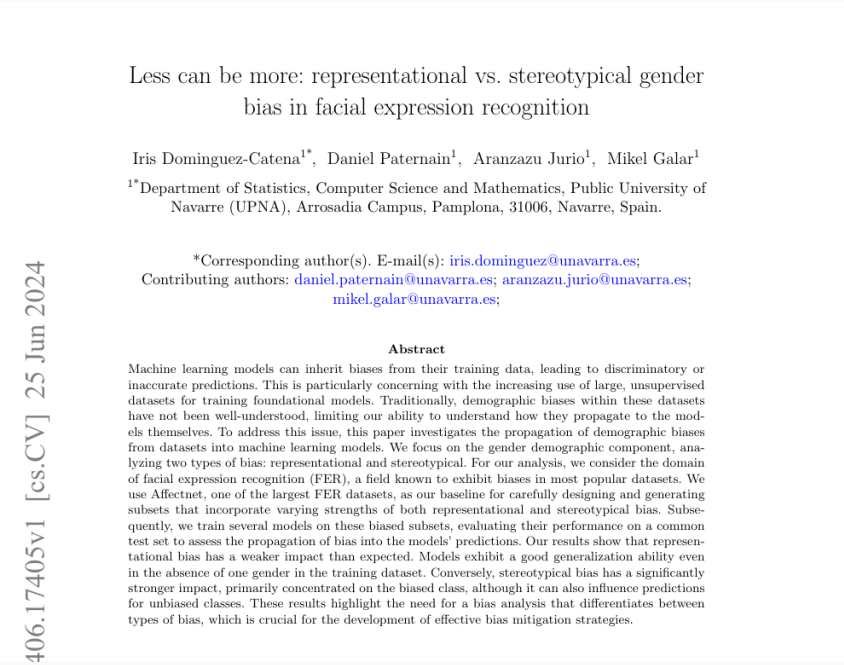


Figure 13: Effect of stereotypical bias.

Conclusion

- **Different impact** of dataset bias:
  - Weaker for representational bias, stronger for stereotypical.
  - Localized to the biased class.
  - Classes show different behaviors.



406.17405v1 [cs.CV] 25 Jun 2024

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

Iris Dominguez-Catena[1*], Daniel Paternain[1], Aranzazu Jurio[1], Mikel Galar[1]

[1*]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.

Accepted for publication in

Progress in Artificial Intelligence

(Q3, Impact Factor: 2.0)

https://doi.org/10.48550/arXiv.2406.17405

Proposals

Measuring Transference from Dataset Bias to Model Predictions

- **Different effects** of different types of dataset bias.
- We have developed **metrics for dataset bias**.
- However, there is **no direct way to connect** dataset bias metrics and model bias metrics.
  - In some cases, we don't have model bias metrics.
  - Multi-group and multiclass classification, like FER.

- **Different effects** of different types of dataset bias.
- We have developed **metrics for dataset bias**.
- However, there is **no direct way to connect** dataset bias metrics and model bias metrics.
  - In some cases, we don't have model bias metrics.
  - Multi-group and multiclass classification, like FER.

- **Different effects** of different types of dataset bias.
- We have developed **metrics for dataset bias**.
- However, there is **no direct way to connect** dataset bias metrics and model bias metrics.
  - In some cases, we don't have model bias metrics.
  - Multi-group and multiclass classification, like FER.

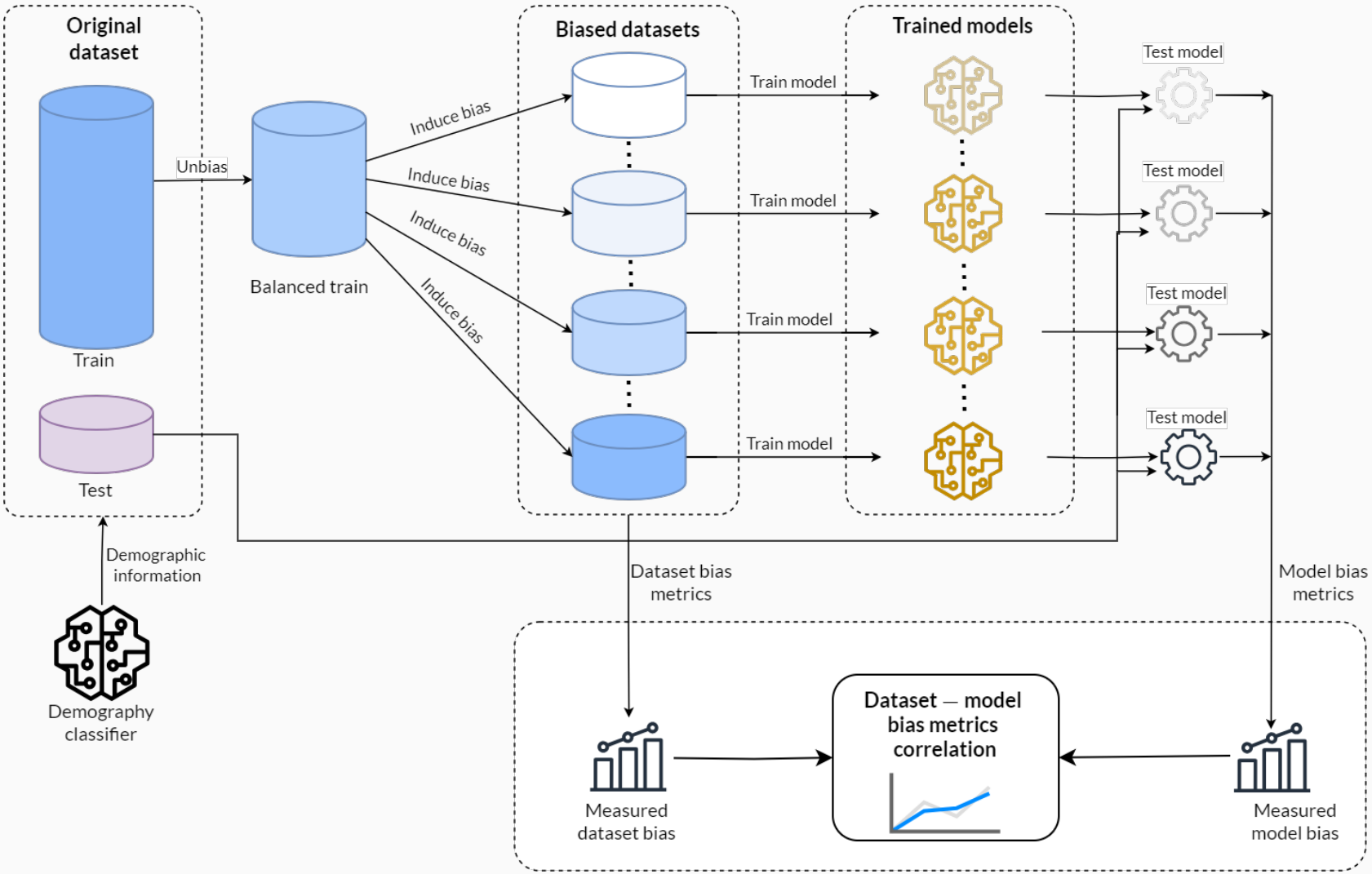


Figure 14: Summary of the methodology

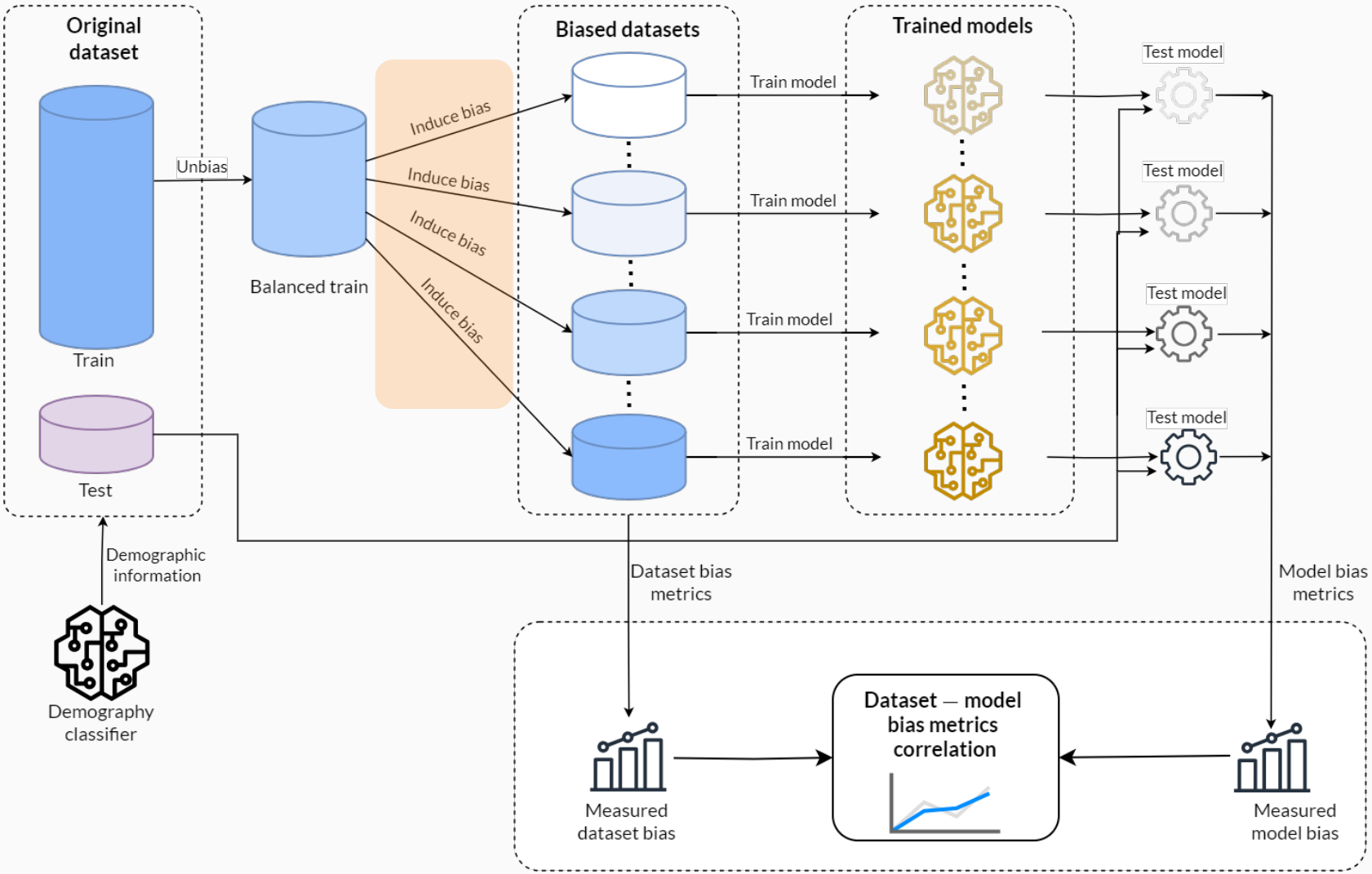


Figure 14: Summary of the methodology

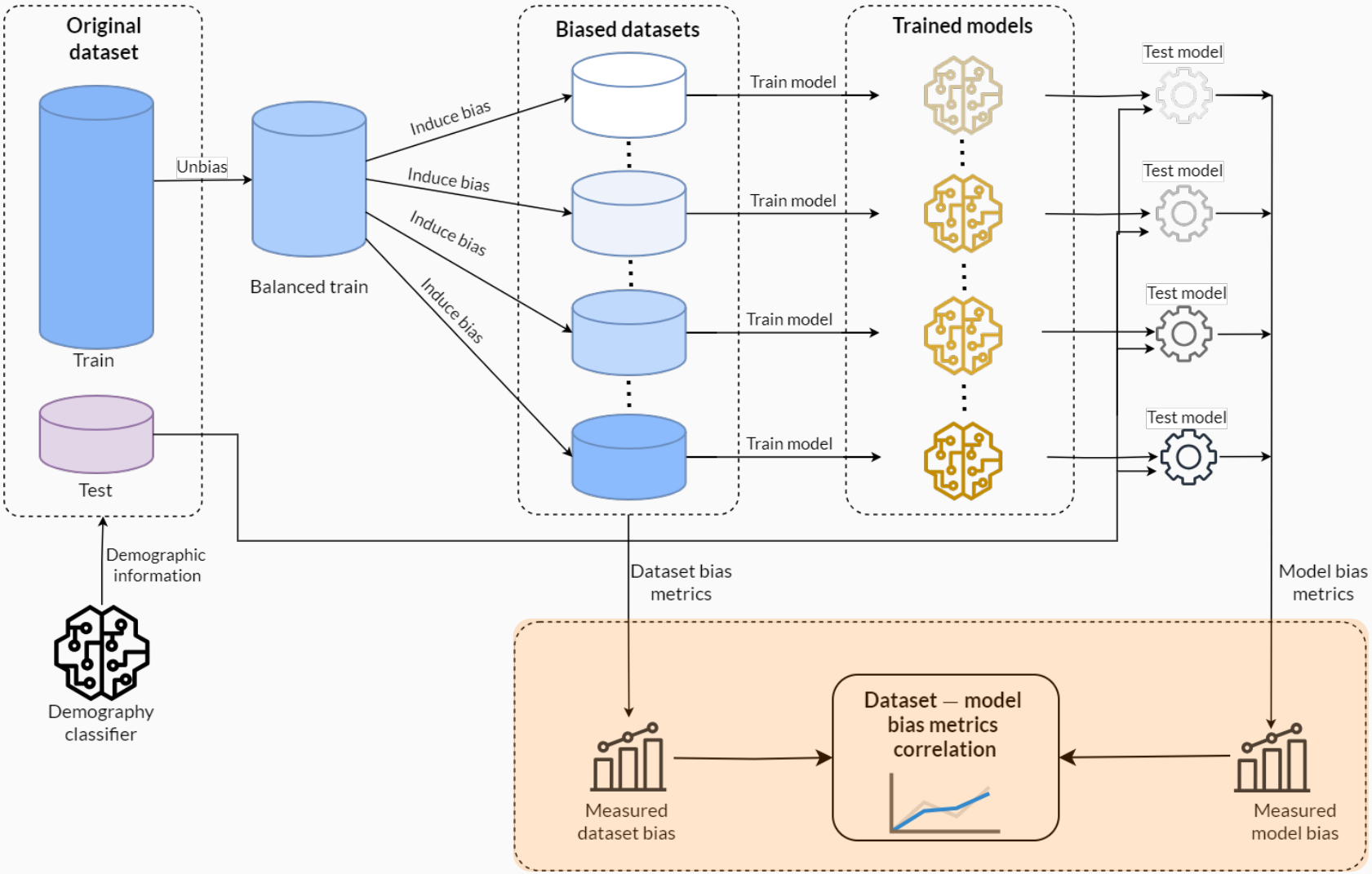


Figure 14: Summary of the methodology

Bias scenarios

Gender

Class	Gender	
	F	M
angry	p	$1-p$
disgust	p	$1-p$
fear	p	$1-p$
happy	p	$1-p$
neutral	p	$1-p$
sad	p	$1-p$
surprise	p	$1-p$

(a) Representational bias

Class	Gender	
	F	M
angry	0.5	0.5
disgust	0.5	0.5
fear	0.5	0.5
happy	p	$1-p$
neutral	0.5	0.5
sad	0.5	0.5
surprise	0.5	0.5

(b) Single class stereotypical bias

Class	Gender	
	F	M
angry	p^+	$1-p^+$
disgust	0.5	0.5
fear	0.5	0.5
happy	0.5	0.5
neutral	0.5	0.5
sad	p^-	$1-p^-$
surprise	0.5	0.5

(c) Multiclass stereotypical bias

Race

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p)/4$ ..				p
disgust	.. $(1-p)/4$ ..				p
fear	.. $(1-p)/4$ ..				p
happy	.. $(1-p)/4$ ..				p
neutral	.. $(1-p)/4$ ..				p
sad	.. $(1-p)/4$ ..				p
surprise	.. $(1-p)/4$ ..				p

(d) Representational bias

Class	Race				
	B	EA	LH	ME	W
angry	0.2	0.2	0.2	0.2	0.2
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	... $(1-p)/4$ ...				p
neutral	0.2	0.2	0.2	0.2	0.2
sad	0.2	0.2	0.2	0.2	0.2
surprise	0.2	0.2	0.2	0.2	0.2

(e) Single class stereotypical bias

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p^+)/4$ ..				p^+
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	0.2	0.2	0.2	0.2	0.2
neutral	0.2	0.2	0.2	0.2	0.2
sad	.. $(1-p^-)/4$ ..				p^-
surprise	0.2	0.2	0.2	0.2	0.2

(f) Multiclass stereotypical bias

Bias scenarios

Gender

Class	Gender	
	F	M
angry	p	$1-p$
disgust	p	$1-p$
fear	p	$1-p$
happy	p	$1-p$
neutral	p	$1-p$
sad	p	$1-p$
surprise	p	$1-p$

(a) Representational bias

Class	Gender	
	F	M
angry	0.5	0.5
disgust	0.5	0.5
fear	0.5	0.5
happy	p	$1-p$
neutral	0.5	0.5
sad	0.5	0.5
surprise	0.5	0.5

(b) Single class stereotypical bias

Class	Gender	
	F	M
angry	p^+	$1-p^+$
disgust	0.5	0.5
fear	0.5	0.5
happy	0.5	0.5
neutral	0.5	0.5
sad	p^-	$1-p^-$
surprise	0.5	0.5

(c) Multiclass stereotypical bias

Race

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p)/4$ ..				p
disgust	.. $(1-p)/4$ ..				p
fear	.. $(1-p)/4$ ..				p
happy	.. $(1-p)/4$ ..				p
neutral	.. $(1-p)/4$ ..				p
sad	.. $(1-p)/4$ ..				p
surprise	.. $(1-p)/4$ ..				p

(d) Representational bias

Class	Race				
	B	EA	LH	ME	W
angry	0.2	0.2	0.2	0.2	0.2
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	... $(1-p)/4$ ...				p
neutral	0.2	0.2	0.2	0.2	0.2
sad	0.2	0.2	0.2	0.2	0.2
surprise	0.2	0.2	0.2	0.2	0.2

(e) Single class stereotypical bias

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p^+)/4$ ..				p^+
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	0.2	0.2	0.2	0.2	0.2
neutral	0.2	0.2	0.2	0.2	0.2
sad	.. $(1-p^-)/4$ ..				p^-
surprise	0.2	0.2	0.2	0.2	0.2

(f) Multiclass stereotypical bias

Gender

Class	Gender	
	F	M
angry	p	$1-p$
disgust	p	$1-p$
fear	p	$1-p$
happy	p	$1-p$
neutral	p	$1-p$
sad	p	$1-p$
surprise	p	$1-p$

(a) Representational bias

Class	Gender	
	F	M
angry	0.5	0.5
disgust	0.5	0.5
fear	0.5	0.5
happy	p	$1-p$
neutral	0.5	0.5
sad	0.5	0.5
surprise	0.5	0.5

(b) Single class stereotypical bias

Class	Gender	
	F	M
angry	p^+	$1-p^+$
disgust	0.5	0.5
fear	0.5	0.5
happy	0.5	0.5
neutral	0.5	0.5
sad	p^-	$1-p^-$
surprise	0.5	0.5

(c) Multiclass stereotypical bias

Race

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p)/4$ ..				p
disgust	.. $(1-p)/4$ ..				p
fear	.. $(1-p)/4$ ..				p
happy	.. $(1-p)/4$ ..				p
neutral	.. $(1-p)/4$ ..				p
sad	.. $(1-p)/4$ ..				p
surprise	.. $(1-p)/4$ ..				p

(d) Representational bias

Class	Race				
	B	EA	LH	ME	W
angry	0.2	0.2	0.2	0.2	0.2
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	... $(1-p)/4$ ...				p
neutral	0.2	0.2	0.2	0.2	0.2
sad	0.2	0.2	0.2	0.2	0.2
surprise	0.2	0.2	0.2	0.2	0.2

(e) Single class stereotypical bias

Class	Race				
	B	EA	LH	ME	W
angry	.. $(1-p^+)/4$ ..				p^+
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	0.2	0.2	0.2	0.2	0.2
neutral	0.2	0.2	0.2	0.2	0.2
sad	.. $(1-p^-)/4$ ..				p^-
surprise	0.2	0.2	0.2	0.2	0.2

(f) Multiclass stereotypical bias

Gender

Class	Gender	
	F	M
angry	p	$1-p$
disgust	p	$1-p$
fear	p	$1-p$
happy	p	$1-p$
neutral	p	$1-p$
sad	p	$1-p$
surprise	p	$1-p$

(a) Representational bias

Class	Gender	
	F	M
angry	0.5	0.5
disgust	0.5	0.5
fear	0.5	0.5
happy	p	$1-p$
neutral	0.5	0.5
sad	0.5	0.5
surprise	0.5	0.5

(b) Single class stereotypical bias

Class	Gender	
	F	M
angry	p^+	$1-p^+$
disgust	0.5	0.5
fear	0.5	0.5
happy	0.5	0.5
neutral	0.5	0.5
sad	p^-	$1-p^-$
surprise	0.5	0.5

(c) Multiclass stereotypical bias

Race

Class	Race				
	B	EA	LH	ME	W
angry	..	$(1-p)/4$	..		p
disgust	..	$(1-p)/4$	..		p
fear	..	$(1-p)/4$	..		p
happy	..	$(1-p)/4$	..		p
neutral	..	$(1-p)/4$	..		p
sad	..	$(1-p)/4$	..		p
surprise	..	$(1-p)/4$	..		p

(d) Representational bias

Class	Race				
	B	EA	LH	ME	W
angry	0.2	0.2	0.2	0.2	0.2
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	...	$(1-p)/4$	...		p
neutral	0.2	0.2	0.2	0.2	0.2
sad	0.2	0.2	0.2	0.2	0.2
surprise	0.2	0.2	0.2	0.2	0.2

(e) Single class stereotypical bias

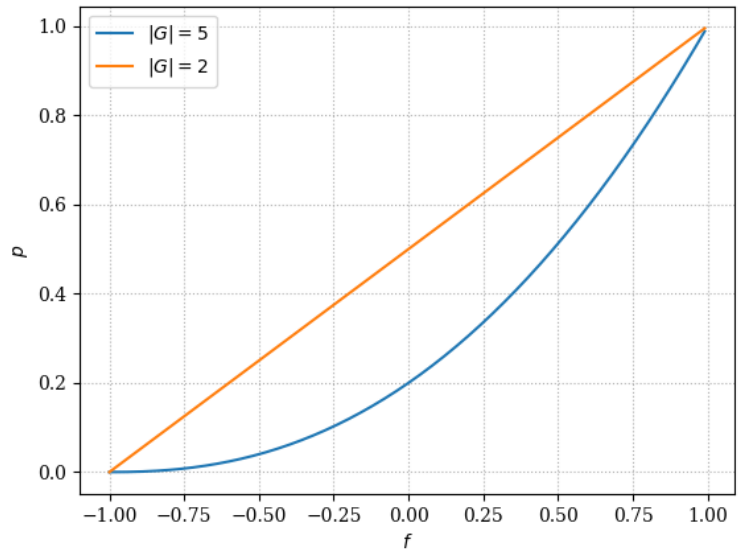
Class	Race				
	B	EA	LH	ME	W
angry	..	$(1-p^+)/4$	..		p^+
disgust	0.2	0.2	0.2	0.2	0.2
fear	0.2	0.2	0.2	0.2	0.2
happy	0.2	0.2	0.2	0.2	0.2
neutral	0.2	0.2	0.2	0.2	0.2
sad	..	$(1-p^-)/4$	..		p^-
surprise	0.2	0.2	0.2	0.2	0.2

(f) Multiclass stereotypical bias

The multigroup problem

We define a bias factor $f \in [-1, 1]$:

$$p = \left(\frac{f+1}{2}\right)^{\log_2 |G|}.$$



- **Original dataset:** Affectnet[44].
- **Network:** pretrained ResNet50[45].
- **Training:** 20 epochs, 1cycle policy, learning rate $1e^{-4}$.
- **Bias factor:** $[-1, -0.8, \ldots, 0.8, 1]$.
  - 3 repetitions per configuration.

[44]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.

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

- Adaptations from[46]:
  - Term-by-term Multiclass Equalized Odds (**TTEqOdds**)
  - Classwise Multiclass Equalized Odds (**CEqOdds**)
  - Multiclass Equality of Opportunity (**EqOpp**)
  - Multiclass Demographic Parity (**DemPar**)
- Previous metrics:
  - Overall disparity (**OD**) [47]
  - Combined Error Variance (**CVE**) [48]
  - Symmetric Distance Error (**SDE**) [48]

[46] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

[48] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proc. Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. IJCAI-ECAI-2022. –2022

[48] Cody Blakeney et al. "Measuring Bias and Fairness in Multiclass Classification". In: *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*. 2022, pp. 1–6

- Adaptations from[46]:
  - Term-by-term Multiclass Equalized Odds (**TTEqOdds**)
  - Classwise Multiclass Equalized Odds (**CEqOdds**)
  - Multiclass Equality of Opportunity (**EqOpp**)
  - Multiclass Demographic Parity (**DemPar**)
- Previous metrics:
  - Overall disparity (**OD**) [47]
  - Combined Error Variance (**CVE**) [48]
  - Symmetric Distance Error (**SDE**) [48]

[46] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

[48] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proc. Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. IJCAI-ECAI-2022. –2022

[48] Cody Blakeney et al. "Measuring Bias and Fairness in Multiclass Classification". In: *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*. 2022, pp. 1–6

- Adaptations from[46]:
  - Term-by-term Multiclass Equalized Odds (**TTEqOdds**)
  - Classwise Multiclass Equalized Odds (**CEqOdds**)
  - Multiclass Equality of Opportunity (**EqOpp**)
  - Multiclass Demographic Parity (**DemPar**)
- Previous metrics:
  - Overall disparity (**OD**) [47]
  - Combined Error Variance (**CVE**) [48]
  - Symmetric Distance Error (**SDE**) [48]

[46] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

[48] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proc. Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. IJCAI-ECAI-2022. –2022

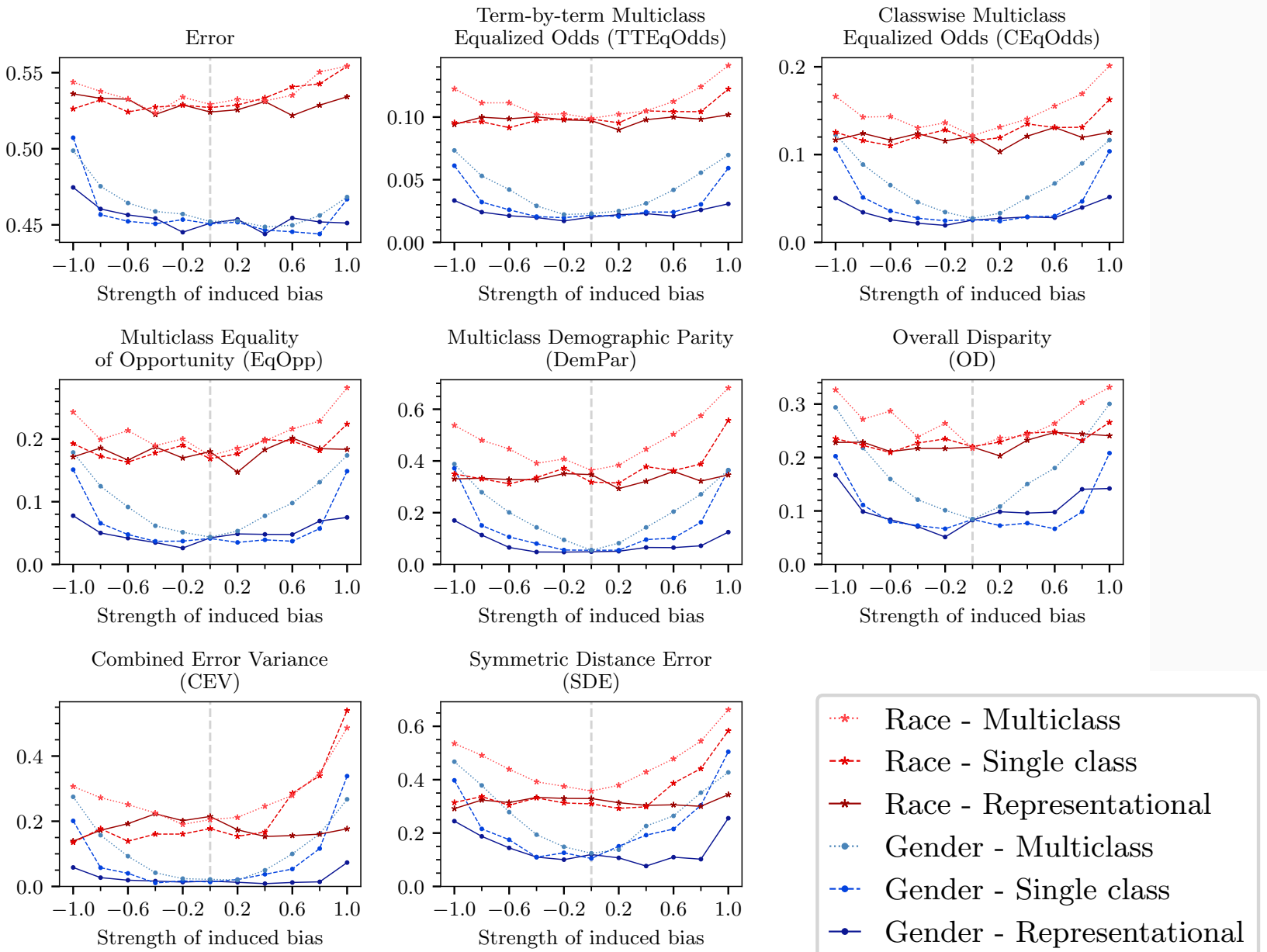
[48] Cody Blakeney et al. "Measuring Bias and Fairness in Multiclass Classification". In: *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*. 2022, pp. 1–6

Term-by-term Multiclass Equalized Odds (TTEqOdds): same confusion matrices across groups.

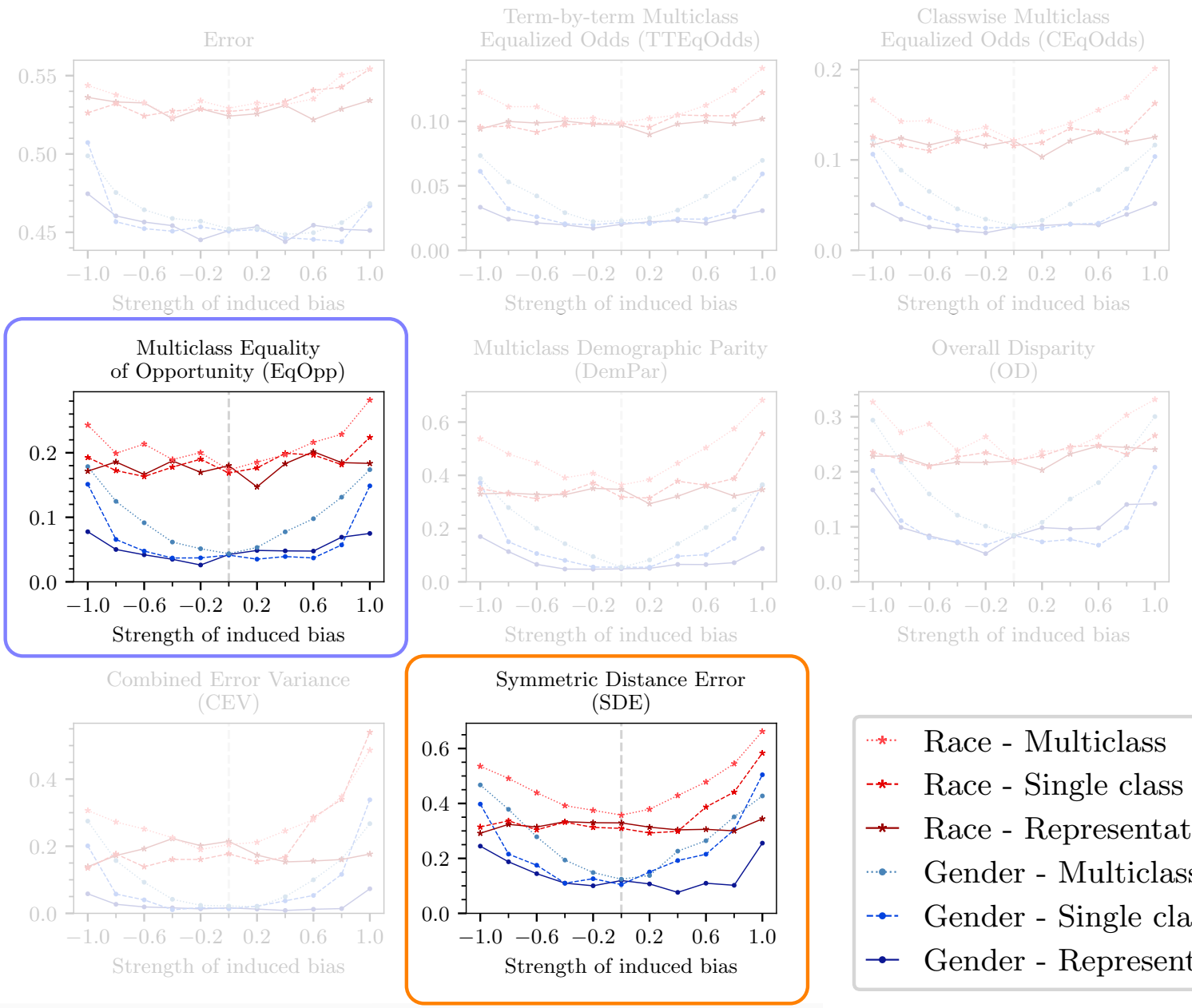
Bias metric: calculate the largest difference in the confusion matrices (W) across groups, then average across Y and $\hat{Y}$.

$$\text{TTEqOdds} = \frac{1}{|\text{Y}|^2} \sum_{i=1}^{|\text{Y}|} \sum_{j=1}^{|\text{Y}|} \max_{g_1, g_2 \in G} |\text{W}_{ij}^{g_1} - \text{W}_{ij}^{g_2}|. \tag{4}$$

Measured model bias



Measured model bias



Metric correlation I

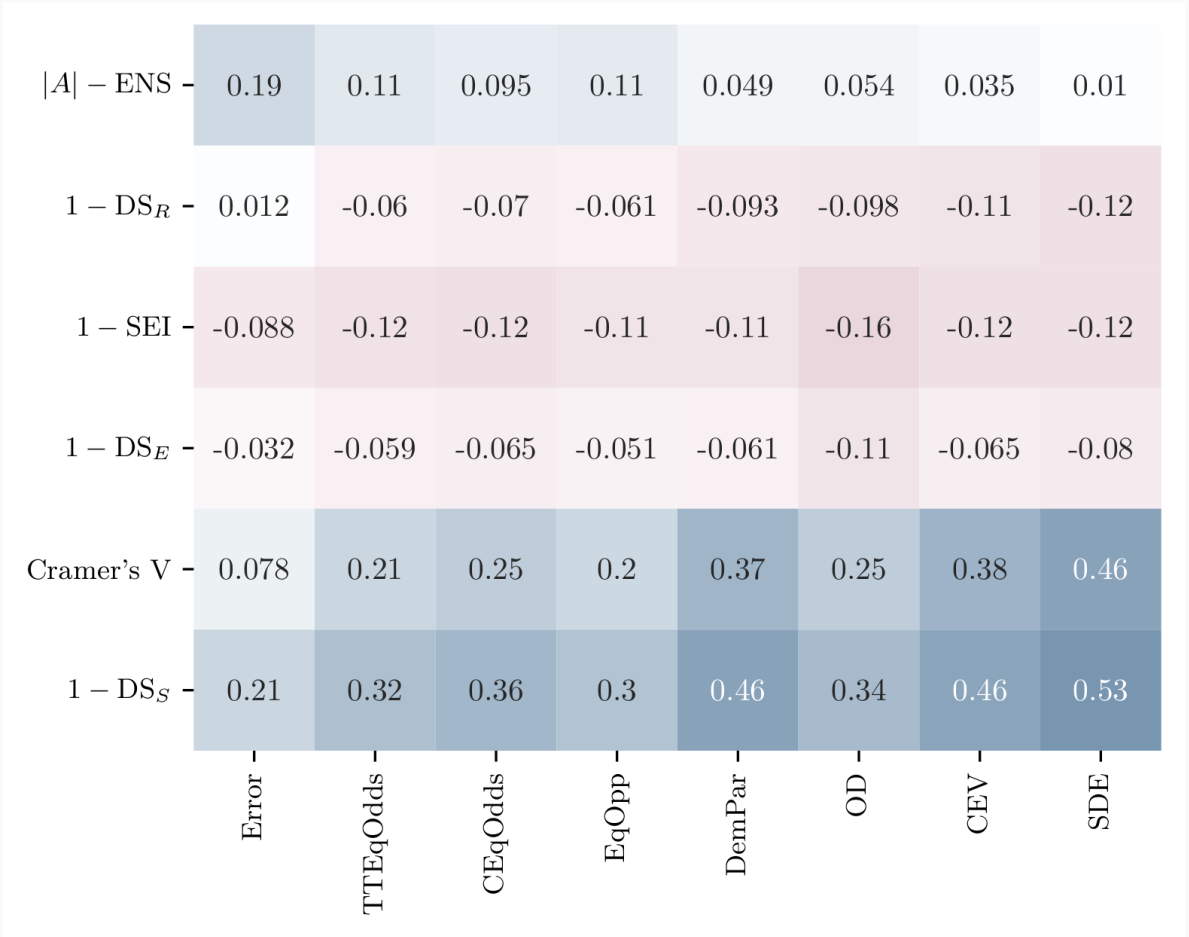


Figure 15: Spearman's ρ rank correlation between bias metrics.

Metric correlation I

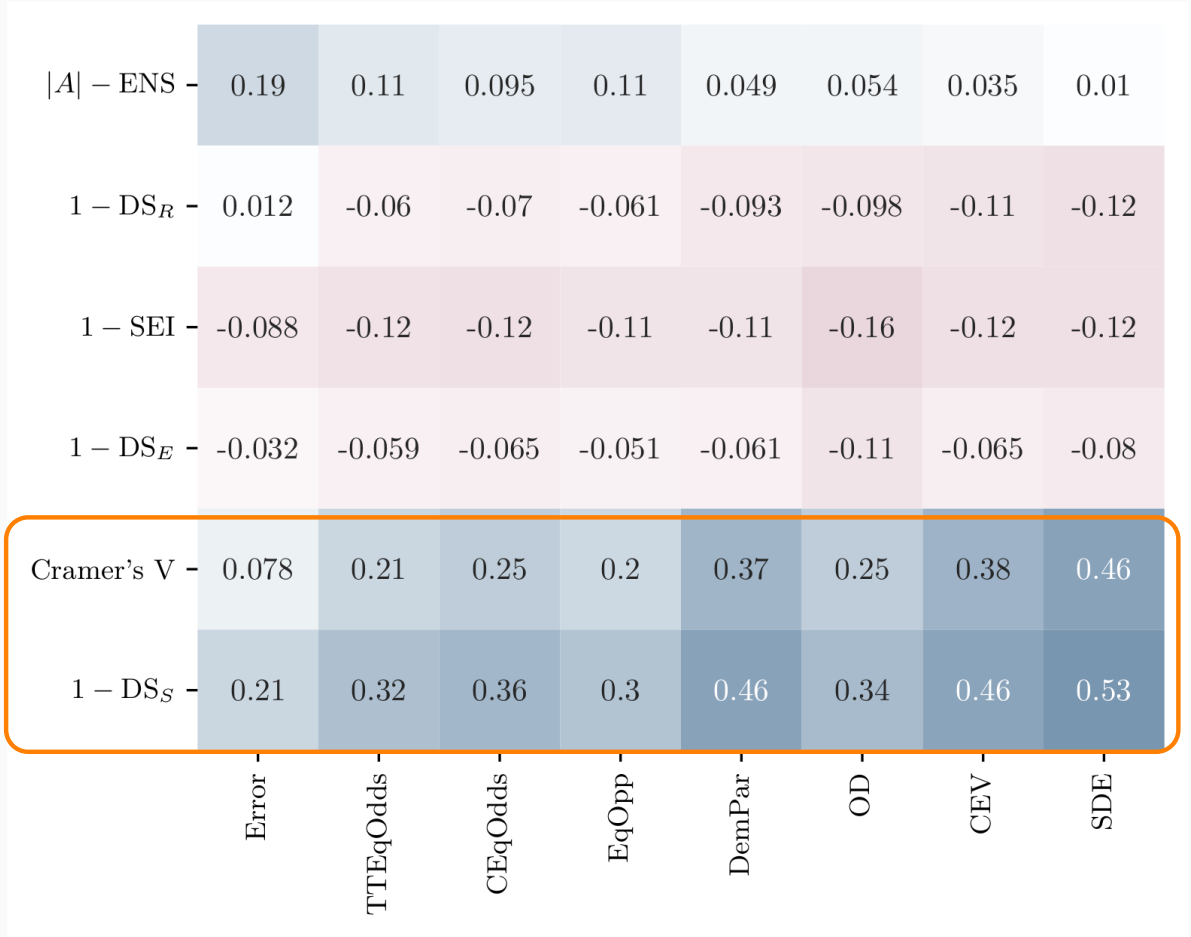


Figure 15: Spearman's ρ rank correlation between bias metrics.

Metric correlation I

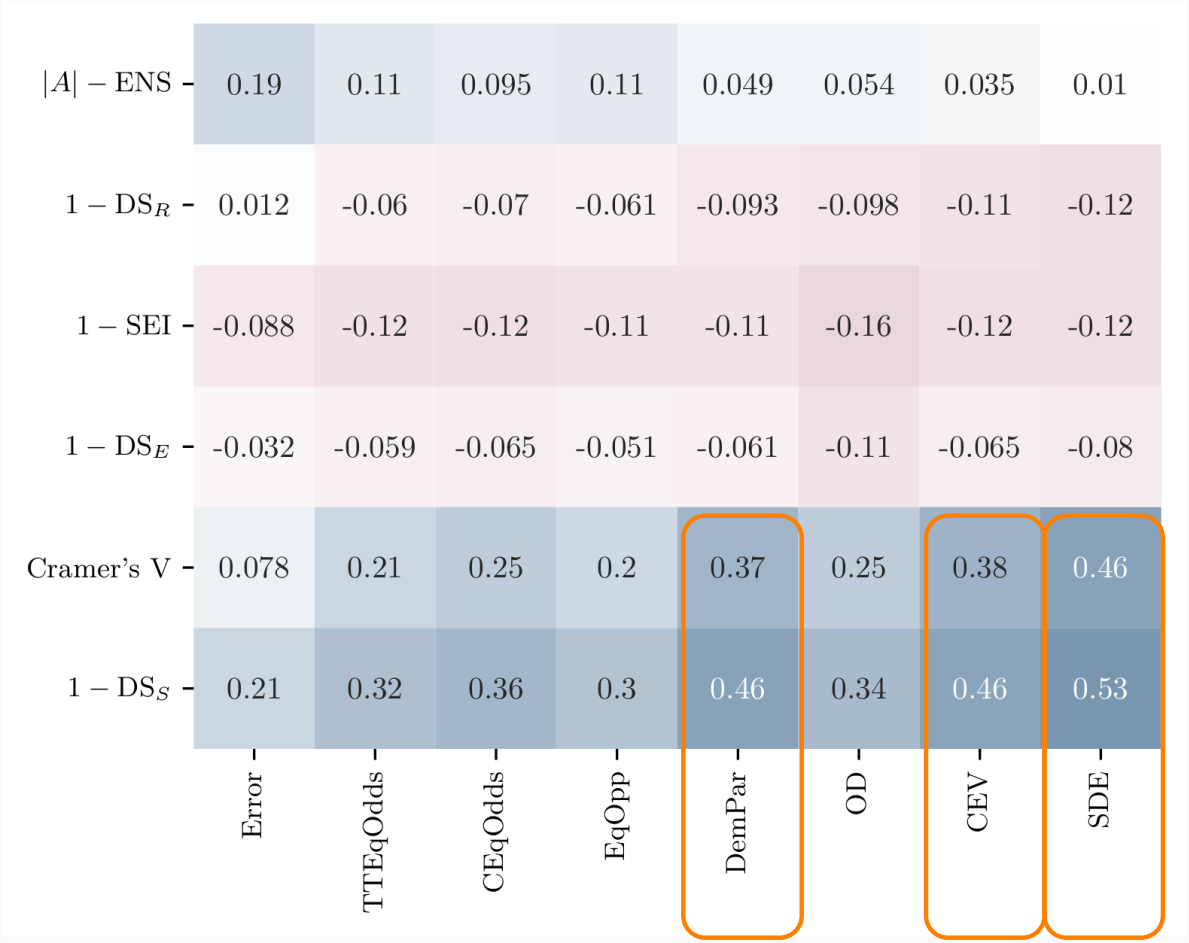


Figure 15: Spearman's ρ rank correlation between bias metrics.

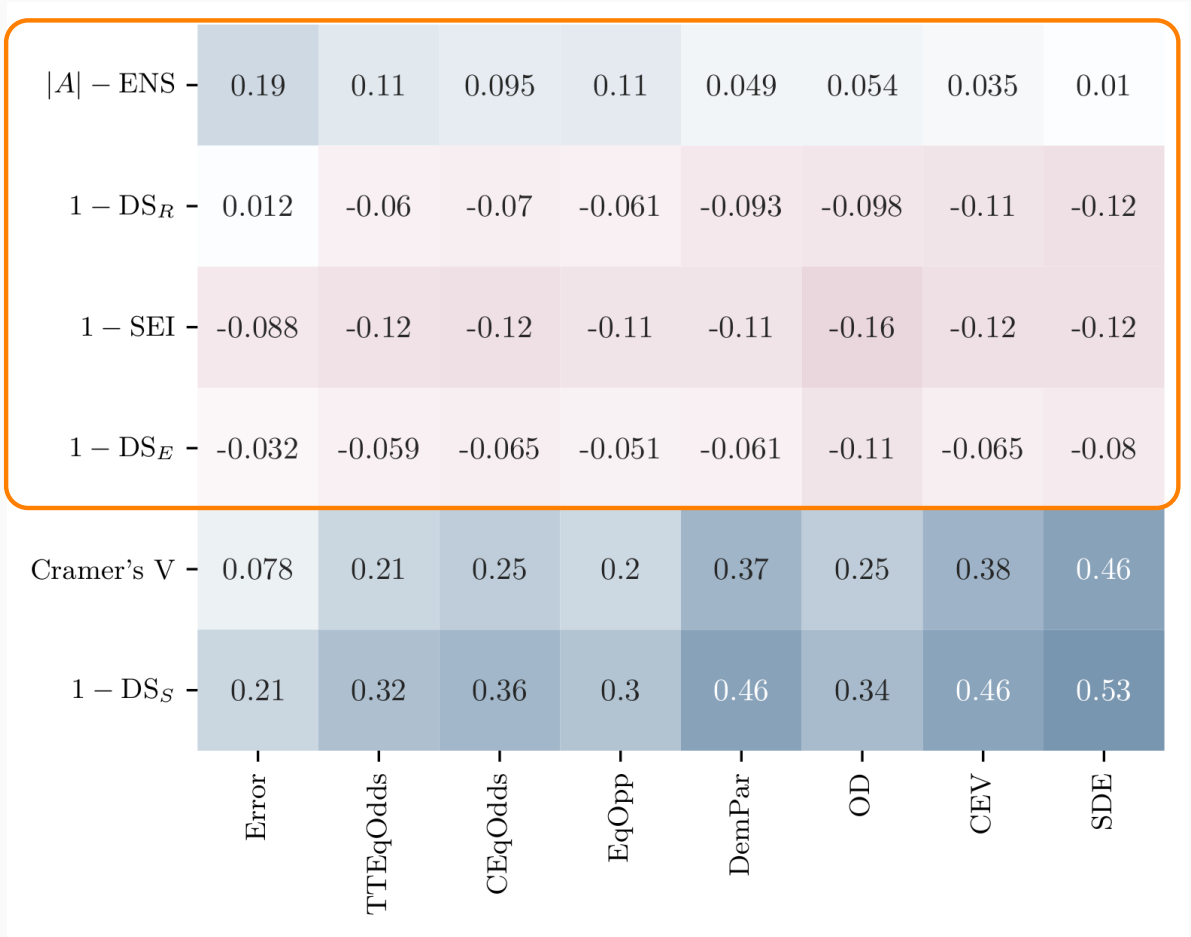


Figure 15: Spearman's ρ rank correlation between bias metrics.

Metric correlation II

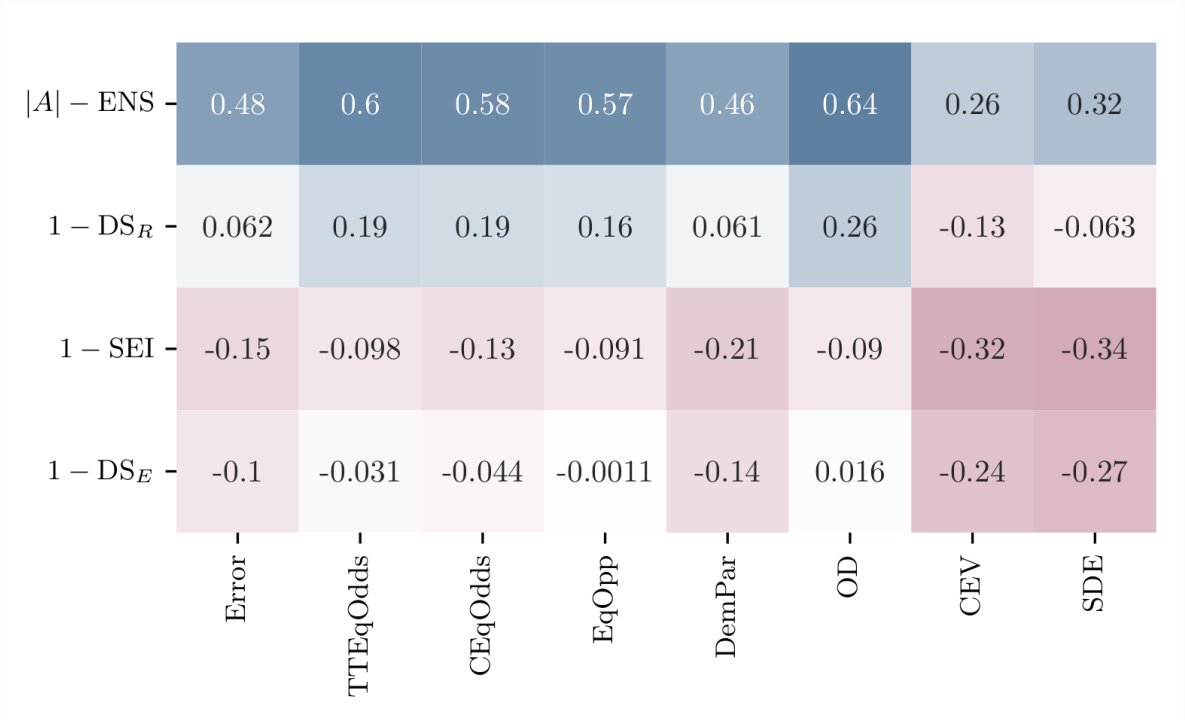


Figure 16: Spearman's ρ rank correlation between bias metrics, restricted to representational bias.

Metric correlation II

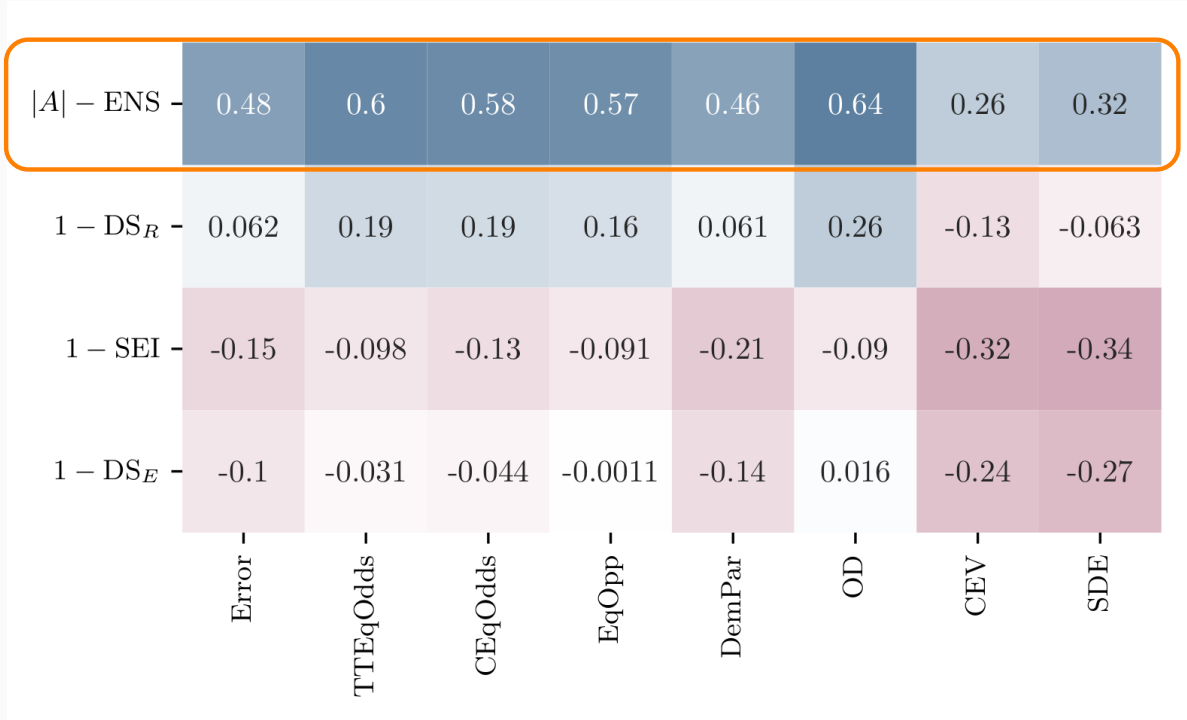


Figure 16: Spearman's ρ rank correlation between bias metrics, restricted to representational bias.

Metric correlation II

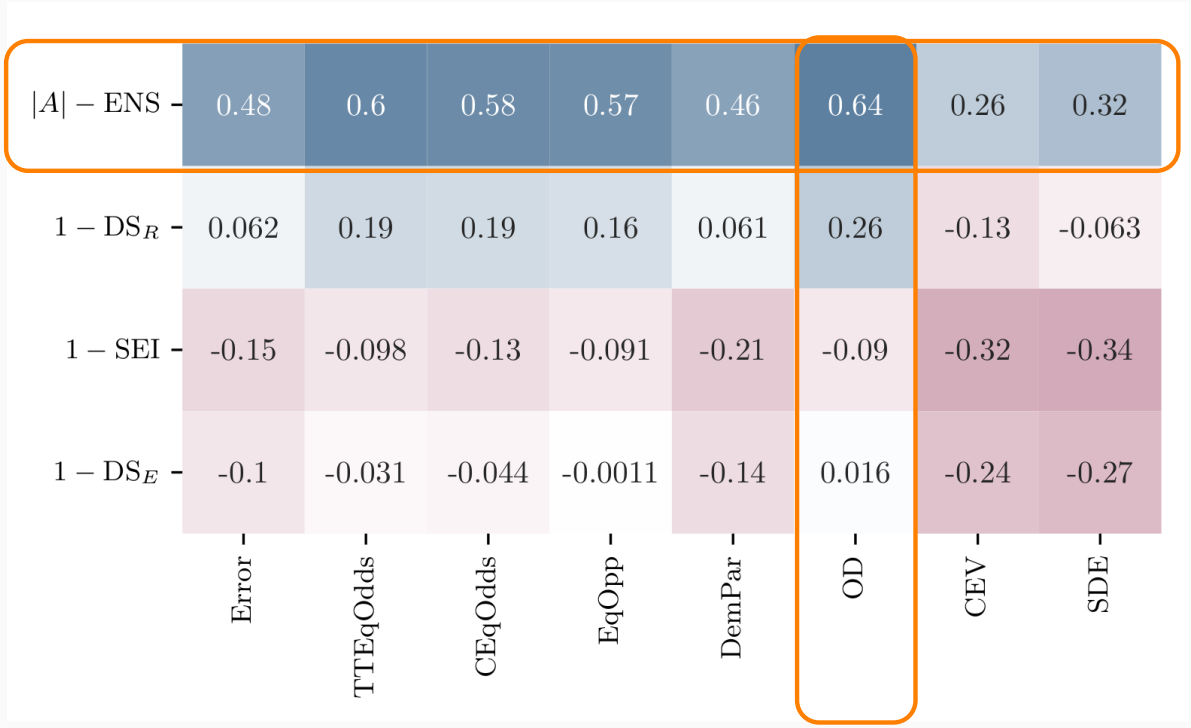


Figure 16: Spearman's ρ rank correlation between bias metrics, restricted to representational bias.

Conclusion

- Representational and stereotypical biases **propagate differently**, affecting fairness metrics in distinct ways.
- Current metrics are more sensitive to stereotypical bias.
- Different types and strengths of dataset bias can threaten different fairness notions in the model:
  - Stereotypical bias (measured by DS_S) strongly linked to CVE and SDE model bias metrics.
  - Representational bias (measured by ENS) correlates better with the OD metric.
- Dataset bias not only propagates to model bias, threatening fairness, but also strongly **correlates with lower accuracy**.

Publication 4

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

In colaboration with the Ghent University.

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

UNIVERSITEIT GENT

To be submitted.

Conference papers

- IJCAI-ECAI 2022:
  - Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proc. Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. IJCAI-ECAI-2022. –2022
- ECML-PKDD 2022:
  - 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. 2023, pp. 9–22

About the measurement of **dataset bias**:

- We have conducted a comprehensive **taxonomy of dataset bias** types, distinguishing metrics for each type.
- We have developed the **DSAP methodology** to compare datasets.
  - DSAP serves as the base for more interpretable and versatile **dataset bias metrics**.
- More recent In-The-Wild (ITW) datasets exhibit a shift from representational to stereotypical bias.

About the **transference of dataset bias** into the model:

- **Stereotypical bias has a stronger impact** on models than representational bias, at least in FER.
- We have developed **new multiclass and multigroup model bias metrics**.
- We identify strong correlations between **stereotypical dataset bias** and most **model bias** metrics.
- Findings highlight the need to **proactively identify and mitigate biases** at all stages of the machine learning pipeline, especially for ITW datasets.

Future work

- Apply proposed tools to **new problems** and demographic variables.
- Study other **complex bias scenarios**.
  - Multiple biases in different classes and combinations of representational and stereotypical bias.
- Adapt metrics to **different types of variable encodings**.
- Create **new model bias metrics** tailored to the effects of each type of bias.
- **Improve auxiliary demographic models**.
- **Bias mitigation** strategies.

Thank you for the attention.

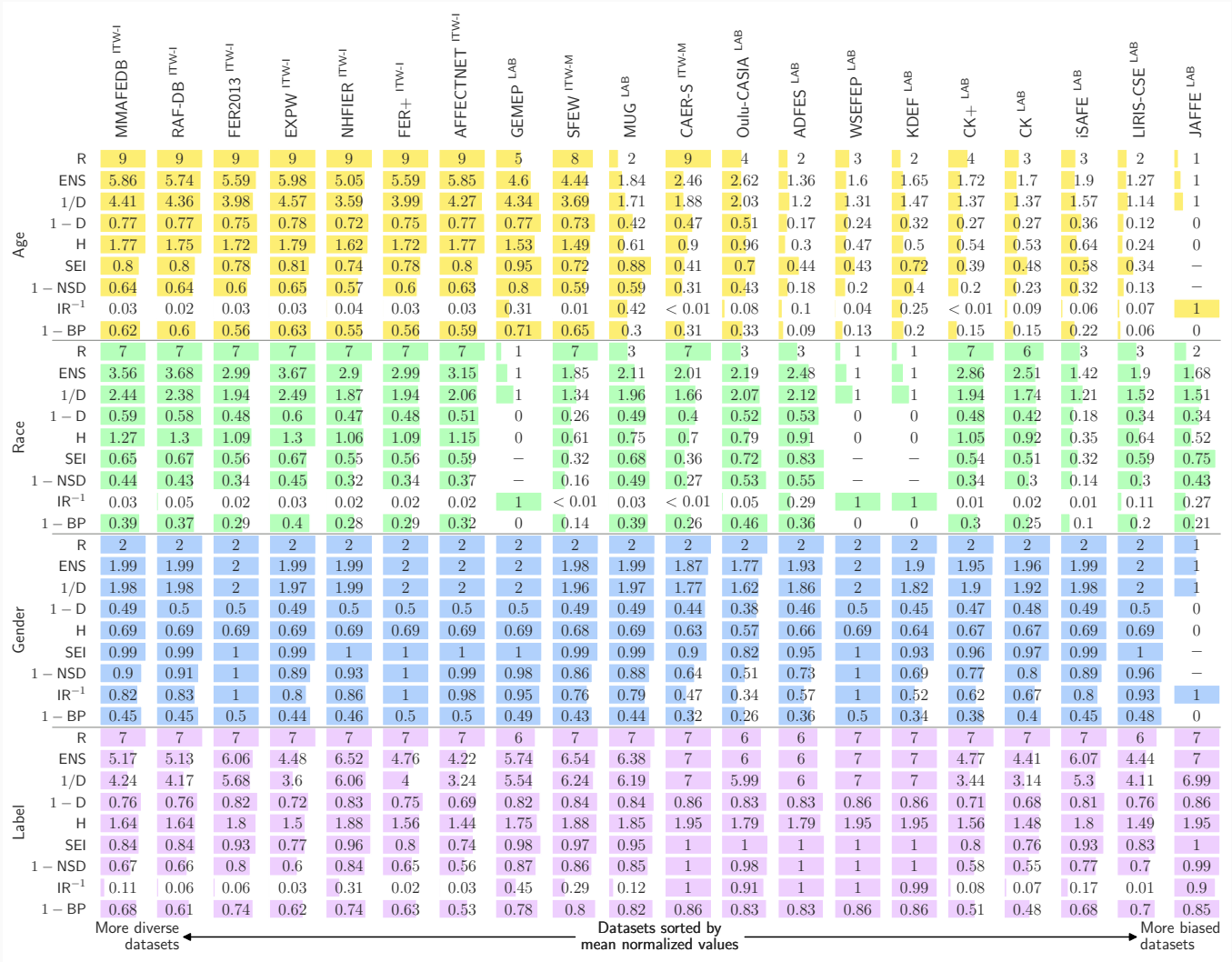
¿Questions?

✉ iris.dominguez@unavarra.es

https://irisai.neocities.org

Results

Measured bias: Representational Stereotypical



		MMAFEDB ITW-I	RAF-DB ITW-I	FER2013 ITW-I	EXPW ITW-I	NHFIER ITW-I	FER+ ITW-I	AFFECTNET ITW-I	GEMEP LAB	SFEW ITW-M	MUG LAB	CAER-S ITW-M	Oulu-CASIA LAB	ADFES LAB	WSEFEP LAB	KDEF LAB	CK+ LAB	CK LAB	iSAFE LAB	LIRIS-CSE LAB	JAFFE LAB
Age	R	9	9	9	9	9	9	9	5	8	2	9	4	2	3	2	4	3	3	2	1
	ENS	5.86	5.74	5.59	5.98	5.05	5.59	5.85	4.6	4.44	1.84	2.46	2.62	1.36	1.6	1.65	1.72	1.7	1.9	1.27	1
	1/D	4.41	4.36	3.98	4.57	3.59	3.99	4.27	4.34	3.69	1.71	1.88	2.03	1.2	1.31	1.47	1.37	1.37	1.57	1.14	1
	1 − D	0.77	0.77	0.75	0.78	0.72	0.75	0.77	0.77	0.73	0.42	0.47	0.51	0.17	0.24	0.32	0.27	0.27	0.36	0.12	0
	H	1.77	1.75	1.72	1.79	1.62	1.72	1.77	1.53	1.49	0.61	0.9	0.96	0.3	0.47	0.5	0.54	0.53	0.64	0.24	0
	SEI	0.8	0.8	0.78	0.81	0.74	0.78	0.8	0.95	0.72	0.88	0.41	0.7	0.44	0.43	0.72	0.39	0.48	0.58	0.34	–
	1 − NSD	0.64	0.64	0.6	0.65	0.57	0.6	0.63	0.8	0.59	0.59	0.31	0.43	0.18	0.2	0.4	0.2	0.23	0.32	0.13	–
	IR^{-1}	0.03	0.02	0.03	0.03	0.04	0.03	0.03	0.31	0.01	0.42	< 0.01	0.08	0.1	0.04	0.25	< 0.01	0.09	0.06	0.07	1
	1 − BP	0.62	0.6	0.56	0.63	0.55	0.56	0.59	0.71	0.65	0.3	0.31	0.33	0.09	0.13	0.2	0.15	0.15	0.22	0.06	0
Race	R	7	7	7	7	7	7	7	1	7	3	7	3	3	1	1	7	6	3	3	2
	ENS	3.56	3.68	2.99	3.67	2.9	2.99	3.15	1	1.85	2.11	2.01	2.19	2.48	1	1	2.86	2.51	1.42	1.9	1.68
	1/D	2.44	2.38	1.94	2.49	1.87	1.94	2.06	1	1.34	1.96	1.66	2.07	2.12	1	1	1.94	1.74	1.21	1.52	1.51
	1 − D	0.59	0.58	0.48	0.6	0.47	0.48	0.51	0	0.26	0.49	0.4	0.52	0.53	0	0	0.48	0.42	0.18	0.34	0.34
	H	1.27	1.3	1.09	1.3	1.06	1.09	1.15	0	0.61	0.75	0.7	0.79	0.91	0	0	1.05	0.92	0.35	0.64	0.52
	SEI	0.65	0.67	0.56	0.67	0.55	0.56	0.59	–	0.32	0.68	0.36	0.72	0.83	–	–	0.54	0.51	0.32	0.59	0.75
	1 − NSD	0.44	0.43	0.34	0.45	0.32	0.34	0.37	–	0.16	0.49	0.27	0.53	0.55	–	–	0.34	0.3	0.14	0.3	0.43
	IR^{-1}	0.03	0.05	0.02	0.03	0.02	0.02	0.02	1	< 0.01	0.03	< 0.01	0.05	0.29	1	1	0.01	0.02	0.01	0.11	0.27
	1 − BP	0.39	0.37	0.29	0.4	0.28	0.29	0.32	0	0.14	0.39	0.26	0.46	0.36	0	0	0.3	0.25	0.1	0.2	0.21
Gender	R	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1
	ENS	1.99	1.99	2	1.99	1.99	2	2	2	1.98	1.99	1.87	1.77	1.93	2	1.9	1.95	1.96	1.99	2	1
	1/D	1.98	1.98	2	1.97	1.99	2	2	2	1.96	1.97	1.77	1.62	1.86	2	1.82	1.9	1.92	1.98	2	1
	1 − D	0.49	0.5	0.5	0.49	0.5	0.5	0.5	0.5	0.49	0.49	0.44	0.38	0.46	0.5	0.45	0.47	0.48	0.49	0.5	0
	H	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.68	0.69	0.63	0.57	0.66	0.69	0.64	0.67	0.67	0.69	0.69	0
	SEI	0.99	0.99	1	0.99	1	1	1	1	0.99	0.99	0.9	0.82	0.95	1	0.93	0.96	0.97	0.99	1	–
	1 − NSD	0.9	0.91	1	0.89	0.93	1	0.99	0.98	0.86	0.88	0.64	0.51	0.73	1	0.69	0.77	0.8	0.89	0.96	–
	IR^{-1}	0.82	0.83	1	0.8	0.86	1	0.98	0.95	0.76	0.79	0.47	0.34	0.57	1	0.52	0.62	0.67	0.8	0.93	1
	1 − BP	0.45	0.45	0.5	0.44	0.46	0.5	0.5	0.49	0.43	0.44	0.32	0.26	0.36	0.5	0.34	0.38	0.4	0.45	0.48	0
Label	R	7	7	7	7	7	7	7	6	7	7	7	6	6	7	7	7	7	7	6	7
	ENS	5.17	5.13	6.06	4.48	6.52	4.76	4.22	5.74	6.54	6.38	7	6	6	7	7	4.77	4.41	6.07	4.44	7
	1/D	4.24	4.17	5.68	3.6	6.06	4	3.24	5.54	6.24	6.19	7	5.99	6	7	7	3.44	3.14	5.3	4.11	6.99
	1 − D	0.76	0.76	0.82	0.72	0.83	0.75	0.69	0.82	0.84	0.84	0.86	0.83	0.83	0.86	0.86	0.71	0.68	0.81	0.76	0.86
	H	1.64	1.64	1.8	1.5	1.88	1.56	1.44	1.75	1.88	1.85	1.95	1.79	1.79	1.95	1.95	1.56	1.48	1.8	1.49	1.95
	SEI	0.84	0.84	0.93	0.77	0.96	0.8	0.74	0.98	0.97	0.95	1	1	1	1	1	0.8	0.76	0.93	0.83	1
	1 − NSD	0.67	0.66	0.8	0.6	0.84	0.65	0.56	0.87	0.86	0.85	1	0.98	1	1	1	0.58	0.55	0.77	0.7	0.99
	IR^{-1}	0.11	0.06	0.06	0.03	0.31	0.02	0.03	0.45	0.29	0.12	1	0.91	1	1	0.99	0.08	0.07	0.17	0.01	0.9
	1 − BP	0.68	0.61	0.74	0.62	0.74	0.63	0.53	0.78	0.8	0.82	0.86	0.83	0.83	0.86	0.86	0.51	0.48	0.68	0.7	0.85

More diverse datasets ← Datasets sorted by mean normalized values → More biased datasets

Results

Measured bias: Representational Stereotypical

		WSEFEP LAB	ADFES LAB	KDEF LAB	JAFFE LAB	Oulu-CASIA LAB	MUG LAB	CAER-S ITW-M	EXPW ITW-I	CK+ LAB	MMAFEDB ITW-I	CK LAB	AFFECTNET ITW-I	FER+ ITW-I	FER2013 ITW-I	iSAFE LAB	RAF-DB ITW-I	GEMEP LAB	NHFIER ITW-I	SFEW ITW-M	LIRIS-CSE LAB
Label - Age	ϕ_C	0	0	.002	–	.021	.053	.052°	.072°	.125°	.07°	.137°	.095°	.105°	.104°	.117°	.144△	.165△	.138△	.115°	.134°
	T	0	0	< .001	–	.019	.034	.048	.067	.105	.066	.104	.088	.097	.097	.089	.134	.156	.128	.111	.089
	C	0	0	.002	0	.037	.053	.126	.174	.212	.17	.19	.226	.248	.247	.163	.334	.314	.32	.271	.132
	U	< .001	< .001	< .001	0	< .001	< .001	.004	.009	.011	.009	.009	.017	.021	.018	.008	.036	.037	.031	.019	.007
	U^R	0	< .001	< .001	–	< .001	.002	.009	.008	.032	.008	.026	.014	.019	.019	.023	.034	.042	.035	.024	.042
	NMI	< .001	< .001	< .001	0	< .001	< .001	.003	.004	.008	.004	.007	.008	.01	.009	.006	.018	.02	.017	.011	.006
Label - Race	ϕ_C	–	0	–	.039	.038	.035	.04	.041	.073°	.058°	.096°	.041	.068°	.086°	.197°	.058°	–	.087°	.076°	.256△
	T	–	0	–	.025	.03	.026	.04	.041	.073	.058	.092	.041	.068	.086	.149	.058	–	.087	.076	.204
	C	0	0	0	.039	.054	.049	.096	.1	.175	.141	.21	.099	.164	.206	.268	.14	0	.208	.184	.34
	U	0	< .001	0	< .001	< .001	< .001	.003	.003	.013	.006	.019	.004	.009	.012	.021	.006	0	.011	.01	.051
	U^R	–	0	–	.002	.002	.002	.007	.004	.019	.008	.031	.004	.013	.02	.107	.008	–	.02	.03	.118
	NMI	0	< .001	0	< .001	< .001	< .001	.002	.002	.008	.004	.012	.002	.006	.007	.018	.004	0	.007	.008	.037
Label - Gender	ϕ_C	0	0	.002	–	.026	.054	.138°	.157°	.023	.167°	.053	.195°	.171°	.178°	.108°	.131°	.093	.172°	.261°	.313△
	T	0	0	.001	–	.018	.035	.088	.1	.014	.107	.034	.125	.109	.114	.069	.084	.062	.11	.167	.21
	C	0	0	.002	0	.026	.054	.136	.155	.023	.164	.053	.192	.169	.176	.107	.13	.093	.17	.253	.299
	U	< .001	< .001	< .001	0	< .001	< .001	.005	.008	< .001	.009	< .001	.013	.009	.009	.003	.005	.002	.008	.018	.034
	U^R	< .001	0	< .001	–	< .001	.002	.015	.018	< .001	.02	.002	.028	.021	.023	.009	.012	.006	.022	.05	.074
	NMI	0	< .001	< .001	0	< .001	< .001	.004	.006	< .001	.006	< .001	.009	.007	.006	.002	.004	.002	.006	.014	.024

Less biased datasets ← Datasets sorted by mean normalized values → More biased datasets

Representational bias metrics selection

- General representational: **Effective Number of Species** (ENS)[49]

$$\mathrm{ENS}(X) = \exp\left(-\sum_{g \in G} p_g \ln p_g\right) . \tag{5}$$

- Evenness between represented groups: **Shannon Evenness Index** (SEI)[50]

$$\mathrm{SEI}(X) = \frac{\mathrm{H}(X)}{\ln(\mathrm{R}(X))} , \tag{6}$$

- Dominance: **Berger-Parker Index** (BP)[51]

$$\mathrm{BP}(X) = \frac{\max_{g \in G} n_g}{n} . \tag{7}$$

---

[49] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[50] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144.

[51] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347.

Stereotypical bias metrics selection

Cramer's V (ϕ_C)[52]:

$$\chi^2(X) = \sum_{g \in G} \sum_{y \in Y} \frac{(n_{g \wedge y} - \frac{n_g n_y}{n})^2}{\frac{n_g n_y}{n}} , \tag{8}$$

$$\phi_C(X) = \sqrt{\frac{\chi^2(X)/n}{\min(|G| - 1, |Y| - 1)}} , \tag{9}$$

Preferred for interpretability reasons.

[52] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. 1991, p. 282.

Local stereotypical bias metric selection

Ducher's Z (Z)[53]:

$$Z(X, g, y) = \begin{cases} \frac{p_{g\wedge y} - p_g p_y}{\min[p_g, p_y] - p_g p_y} & \text{if } p_{g\wedge y} - p_g p_y > 0 \\ \frac{p_{g\wedge y} - p_g p_y}{p_g p_y - \max[0, p_g + p_y - 1]} & \text{if } p_{g\wedge y} - p_g p_y < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

[53] 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 (1994), pp. 649–655.

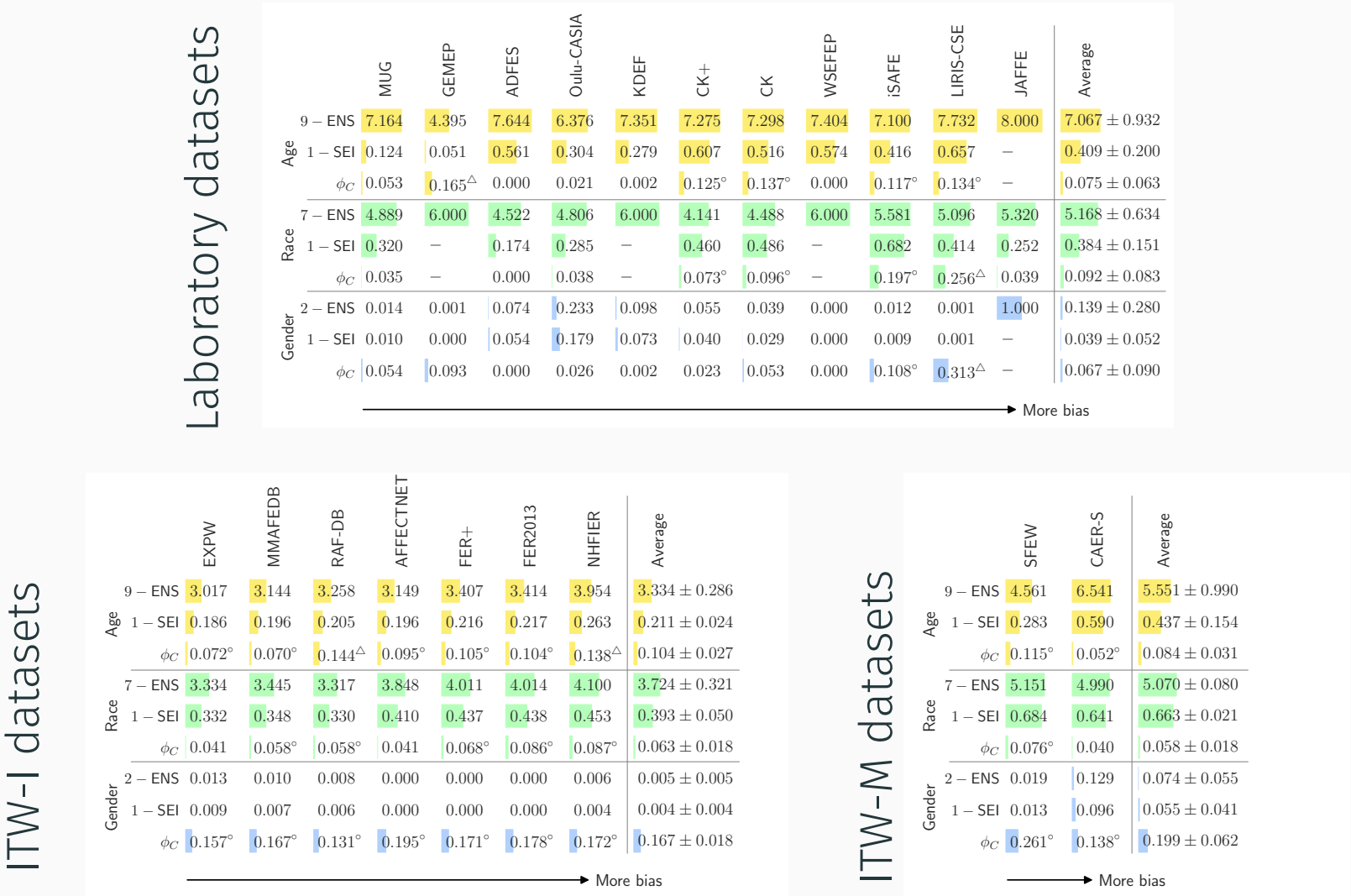


Figure 17: Measured representational bias (ENS), evenness (SEI) and stereotypical bias (ϕ_C).

Results: Bias measurement II

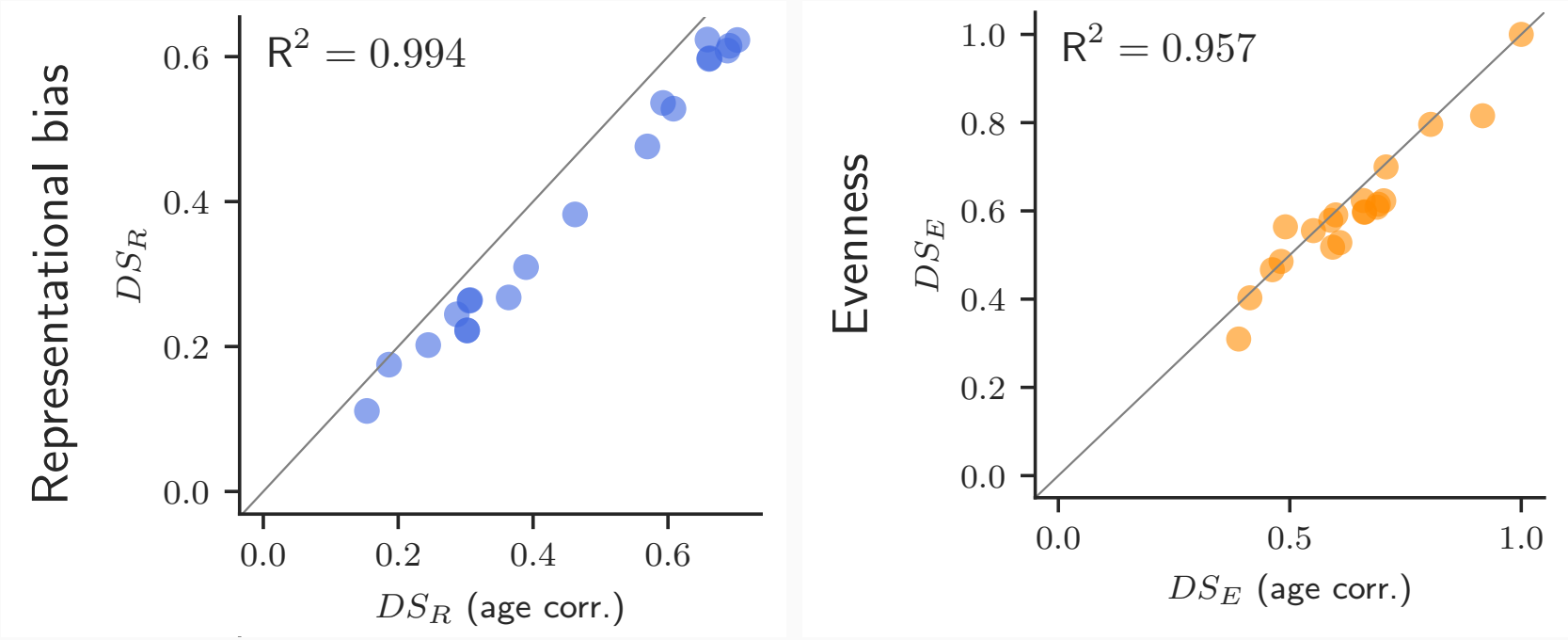


Figure 18: Effect of age-correction on DS_E and DS_R, based on 2021 world age distribution.

Measured model bias