AI Bias in Visual Data

a preliminary take

Symeon (Akis) Papadopoulos @sympap
Information Technologies Institute (ITI)
Centre for Research and Technology Hellas (CERTH)

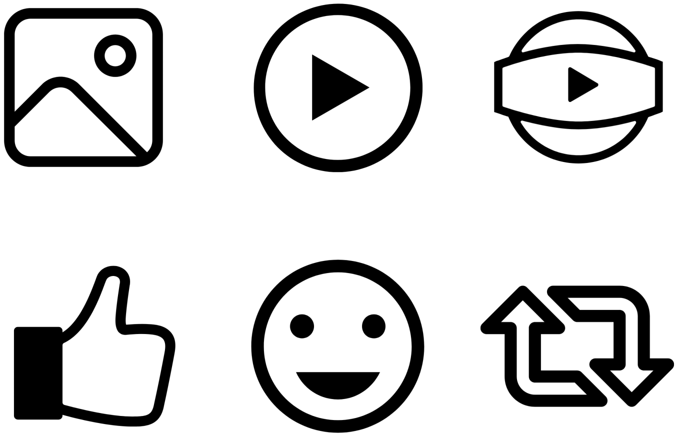
with contributions from Simone Fabbrizzi, Alaa Elobaid, Eirini Ntoutsi, Yiannis Kompatsiaris

Fairness in Artificial Intelligence — June 27, 2022 @ Bocconi University

Overview of Talk

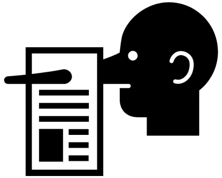
- introduction
- examples where things can go wrong with CV & AI
- background & motivation
- bias in visual datasets - a survey
- addressing visual bias
- parting thoughts

Volume of Online Visual Data



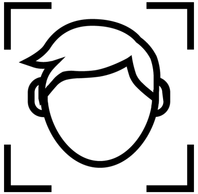
https://www.domo.com/blog/what-data-never-sleeps-9-0-proves-about-the-pandemic/

Media-related AI Applications

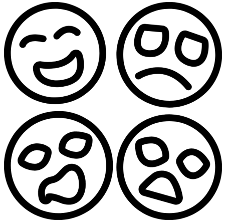


Fighting Disinformation

Content Moderation

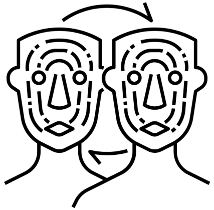


Face Recognition



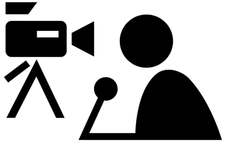
Emotional Profiling

Recommender Systems



Synthetic Media

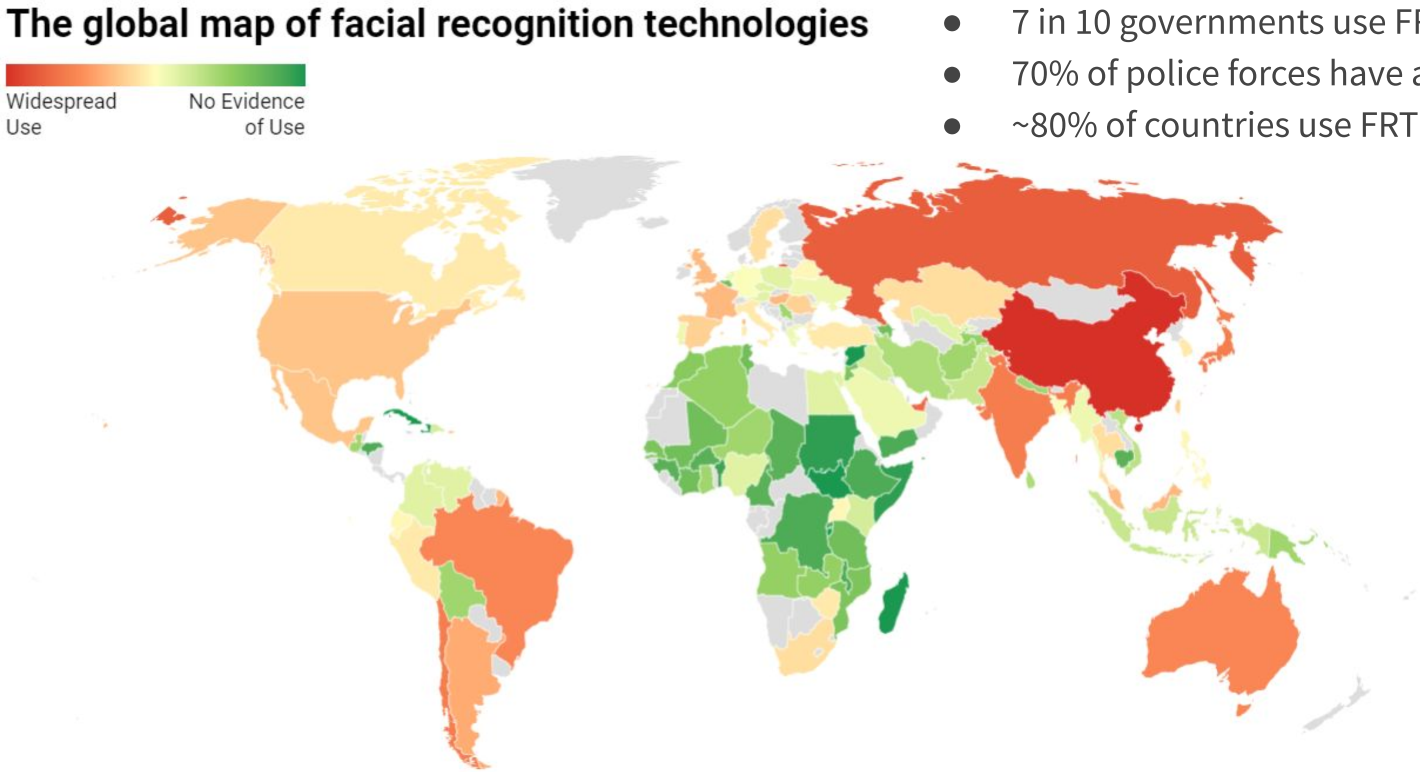
Media Search



Media Productions

Icons sourced from https://thenounproject.com/

Facial Recognition Technology in 100 Countries



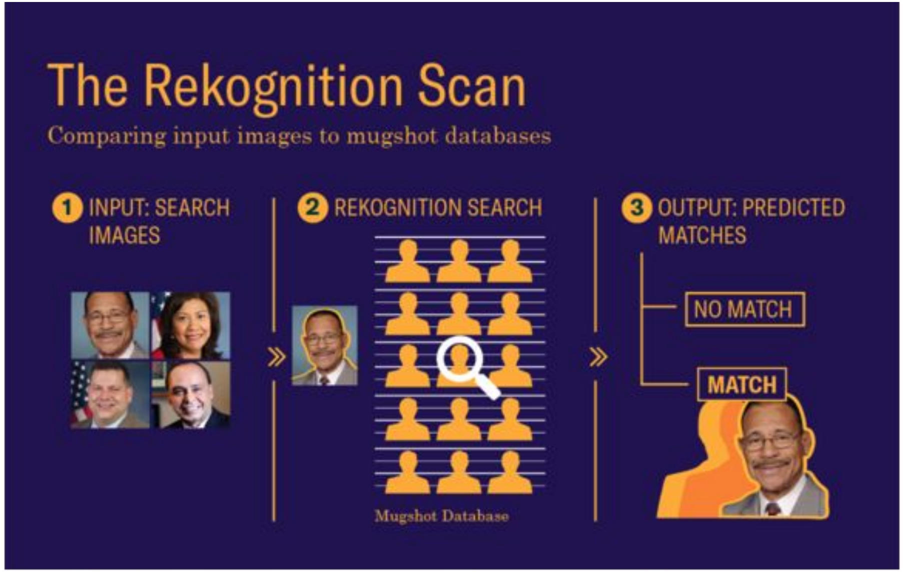
- 7 in 10 governments use FRT on a large-scale basis
- 70% of police forces have access to FRT
- ~80% of countries use FRT in banking

https://www.comparitech.com/blog/vpn-privacy/facial-recognition-statistics/

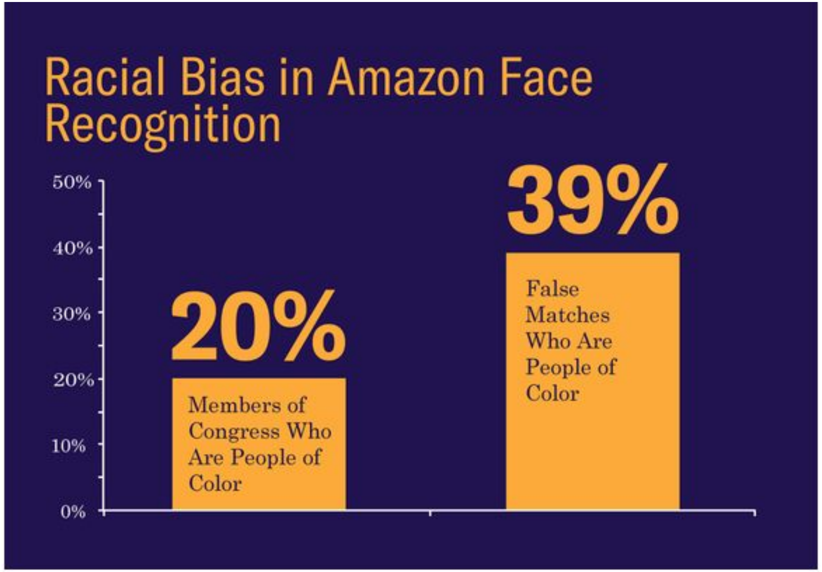
examples

where things can go wrong with CV & AI

Bias in Face Recognition



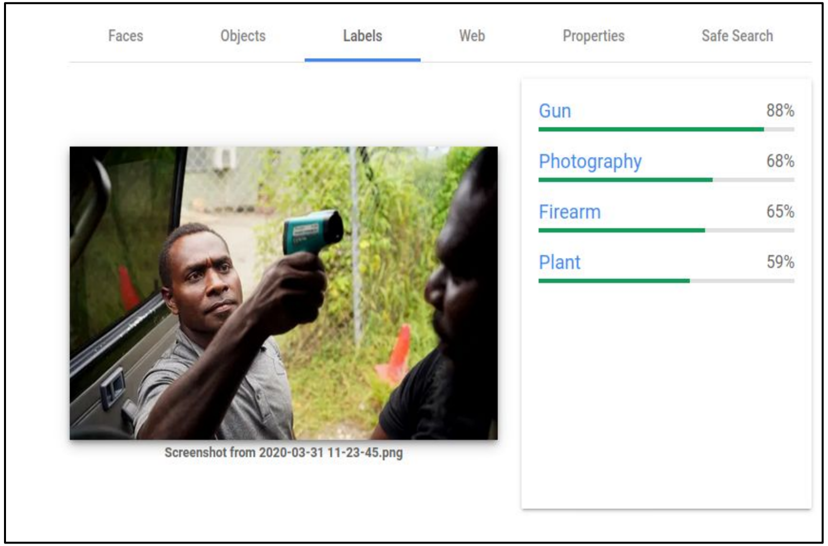
Rep. Sanford Bishop (D-Ga.) was falsely identified by Amazon Rekognition as someone who had been arrested for a crime.



People of color were disproportionately falsely matched in our test.

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

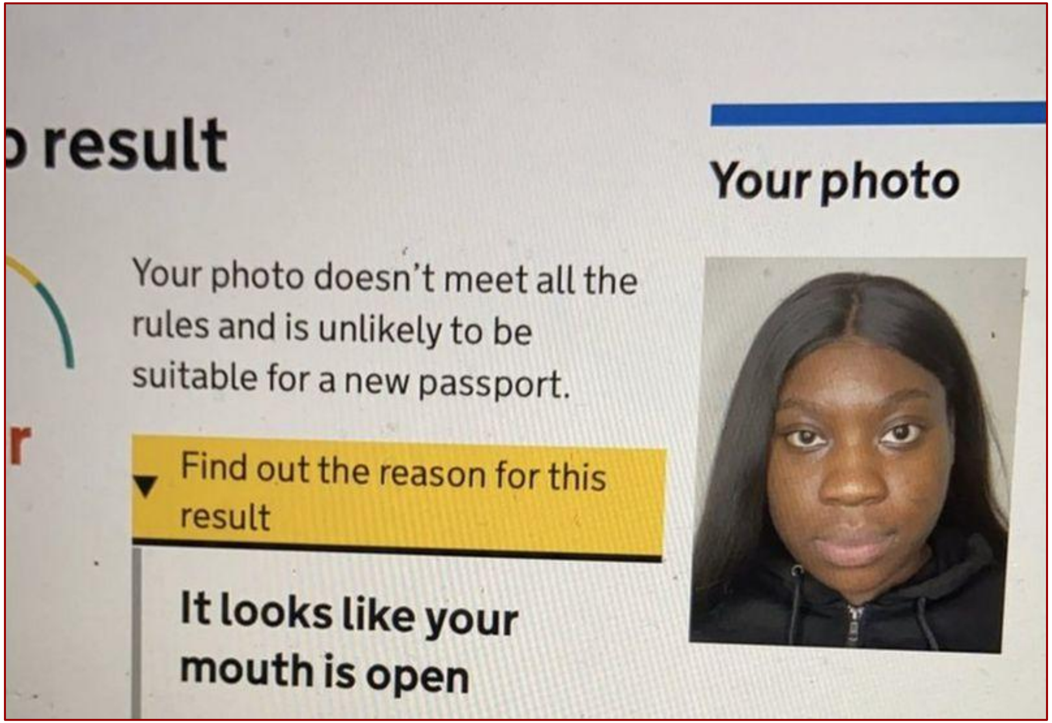
Bias in Object Recognition



Black person with hand-held thermometer → firearm
Asian person with hand-held thermometer → electronic device

https://twitter.com/nicolaskb/status/1244921742486917120

Bias in Passport Photo Checker



- Dark-skinned women are told their photos are poor quality **22%** of the time, while for light-skinned women this happens only **14%** of the time
- Dark-skinned men are told their photos are poor quality **15%** of the time, while the figure for light-skinned men is **9%**
- Photos of women with the darkest skin were **4x more likely to be graded poor quality**, than women with the lightest skin

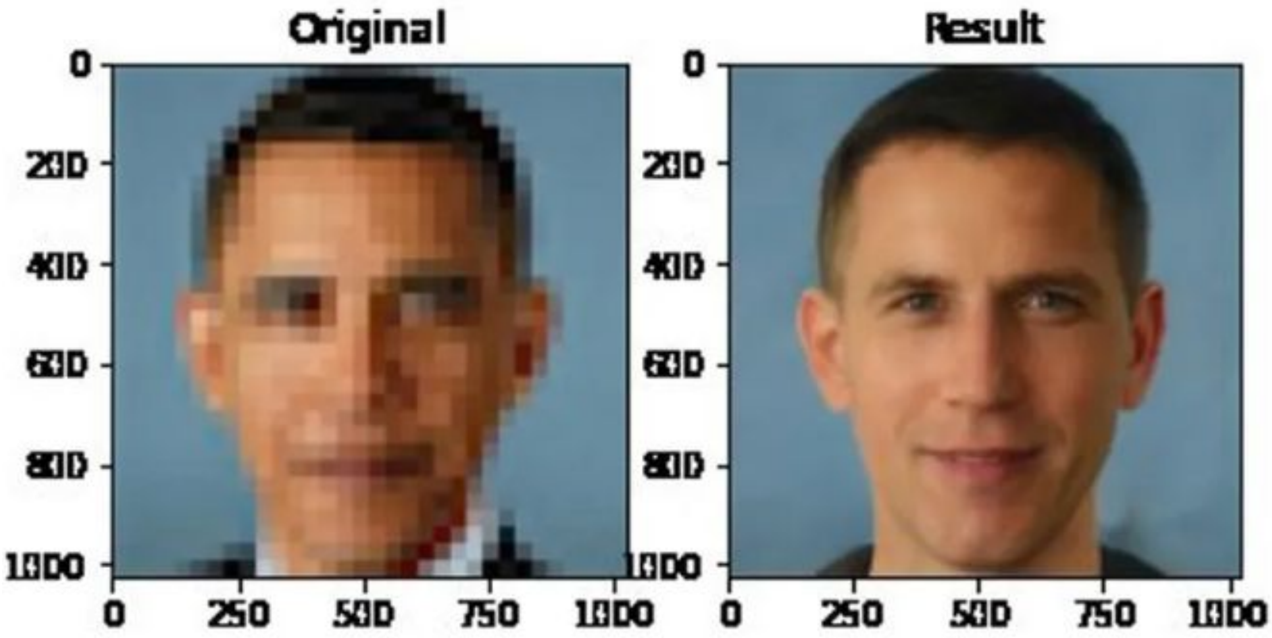
https://www.bbc.com/news/technology-54349538

Bias in Twitter Cropping Algorithm

https://petapixel.com/2020/09/21/twitter-photo-algorithm-draws-heat-for-possible-racial-bias/

Twitter's follow-up: https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm

Biased Super-Resolution

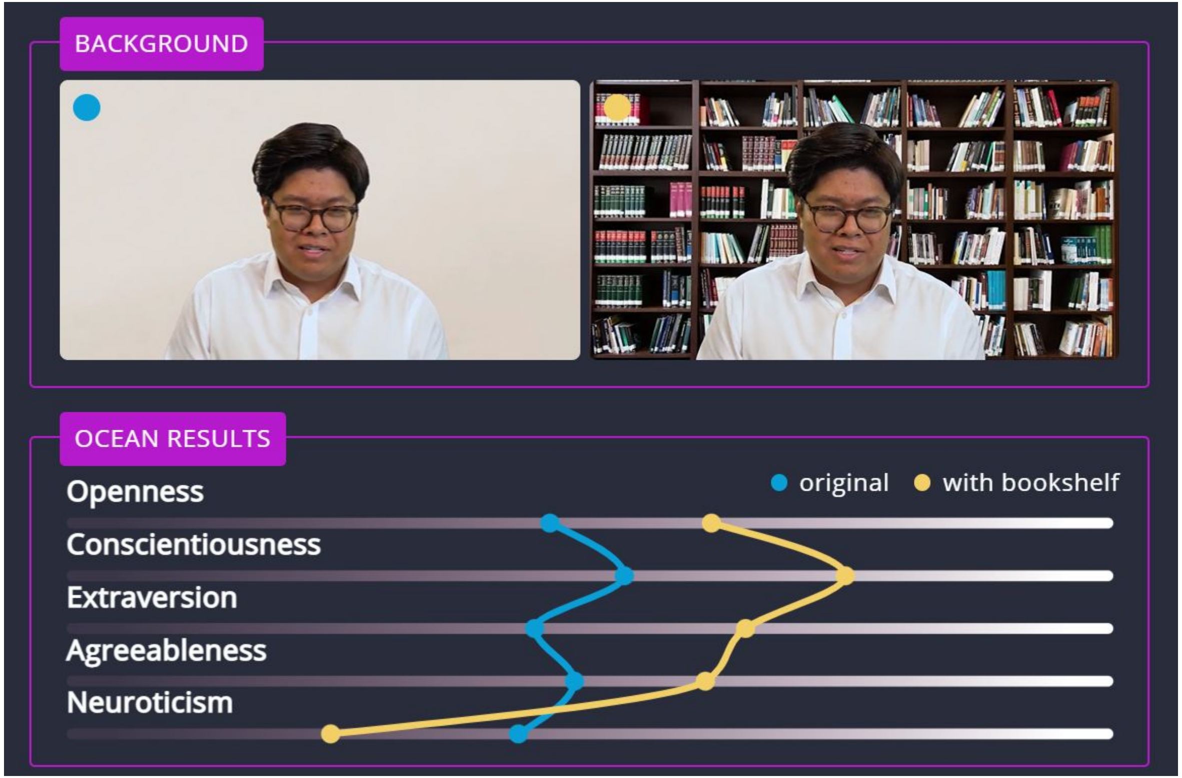


https://twitter.com/Chicken3gg/status/1274314622447820801

Blogpost: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

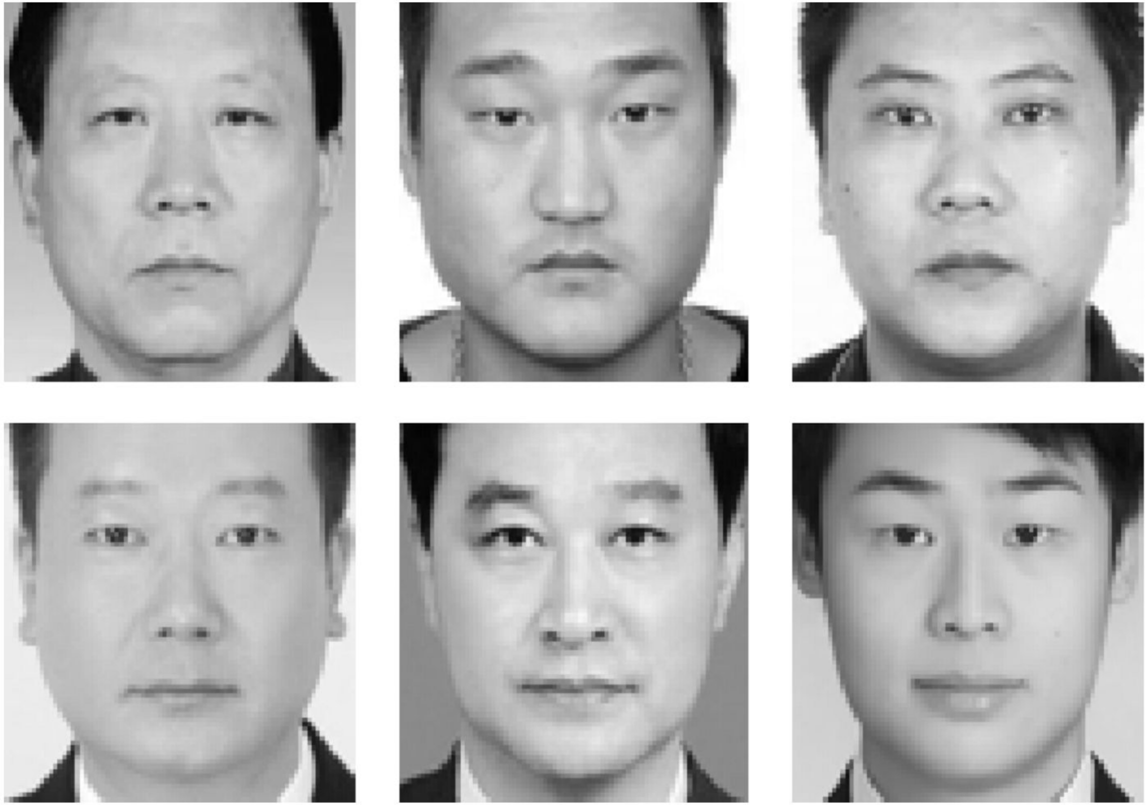
Objective or Biased?

On the questionable use of Artificial Intelligence for job applications



https://interaktiv.br.de/ki-bewerbung/en/index.html

Tell a Criminal Based on Their Face....



Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom). In the top images, the people are frowning. In the bottom, they are not. These types of superficial differences can be picked up by a deep learning system.

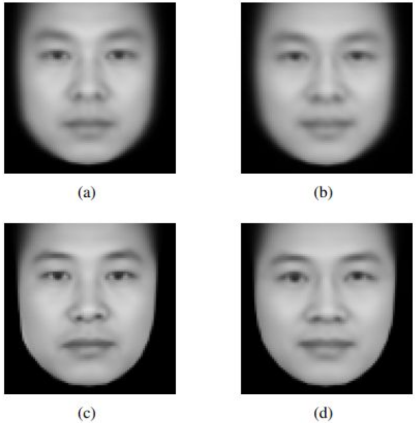


Figure 6. (a) and (b) are "average" faces for criminals and non-criminals generated by averaging of eigenface representations ; (c) and (d) are "average" faces for criminals and non-criminals generated by averaging of landmark points and image warping.

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. arXiv preprint arXiv:1611.04135, 4038-4052.

AI Physiognomy

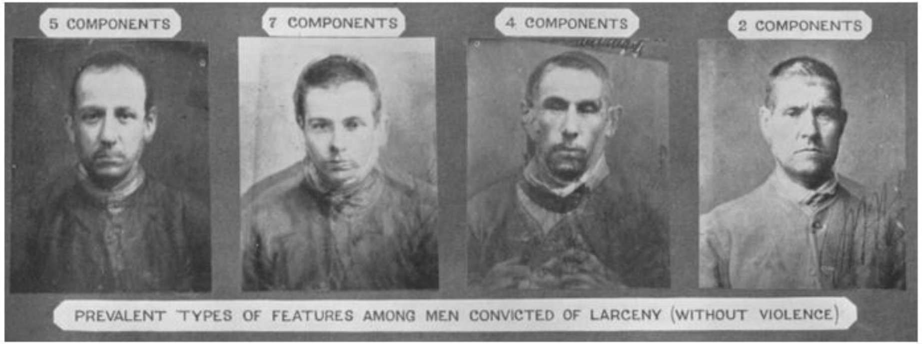
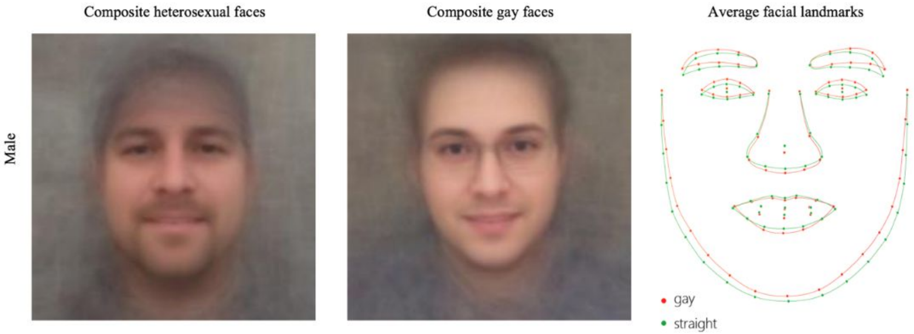


Figure 6. Francis Galton's attempt to reconstruct an "average criminal face".

https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a



Kosinski, M., and Wang, Y. (2018) Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images. *Journal of Personality and Social Psychology. February 2018, 114(2), 246–257.*

"personalities are affected by genes"

"Our face is a reflection of our DNA"

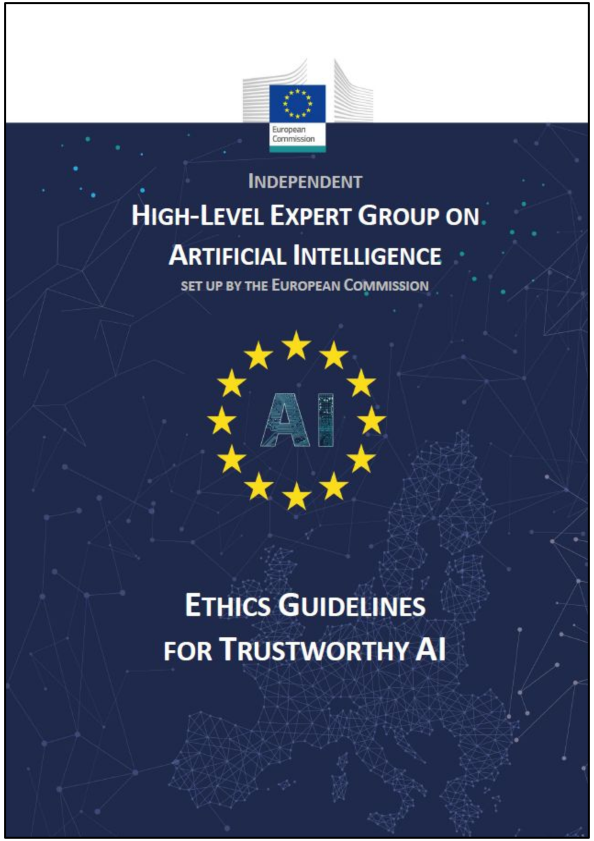
https://www.faception.com/

background concepts & motivation

AI bias basics, AI in media

Trustworthy (aka Responsible) AI

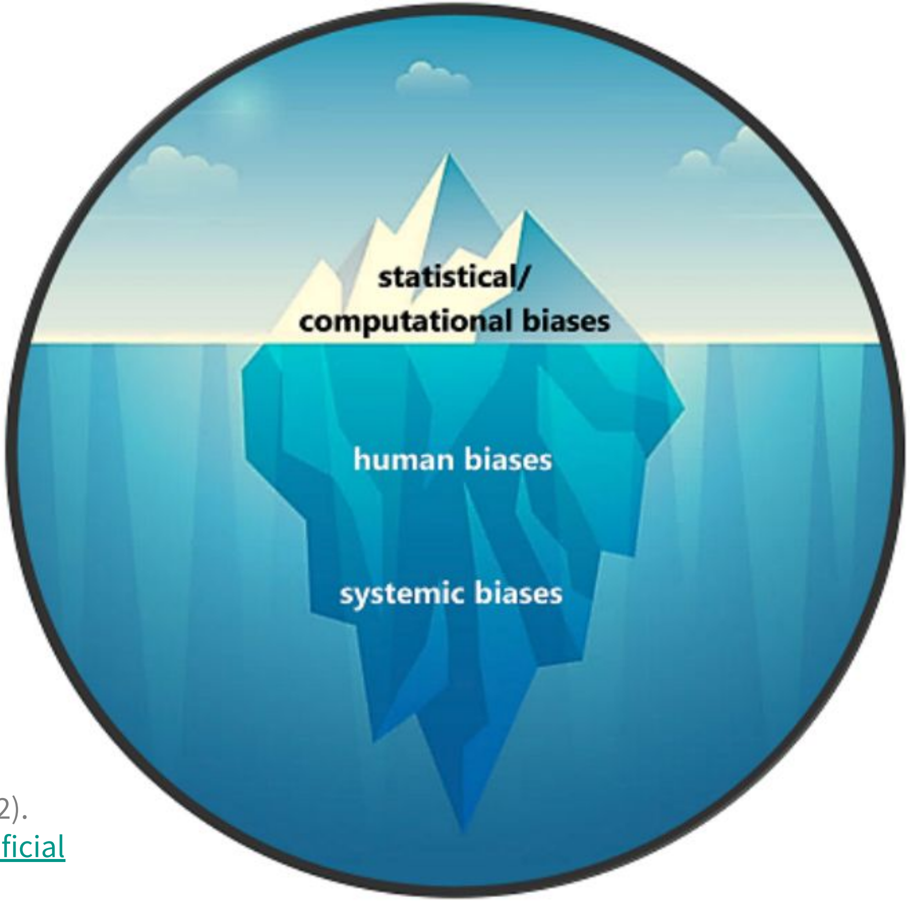
- 4 Ethical Principles
  - Respect for human autonomy
  - Prevention of harm
  - **Fairness**
  - Explicability
- 7 Key Requirements
  - Human agency and oversight
  - Technical robustness and safety
  - Privacy and data governance
  - Transparency
  - **Diversity, non-discrimination and fairness**
  - Societal and environmental wellbeing
  - Accountability



AI HLEG (2019). Ethics Guidelines for Trustworthy AI. European Commission

Bias and AI

- Bias is much more than the statistical and computation bias that we can “easily” measure
- What is needed is a broader socio-technical perspective linking AI practices with societal values



Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt., A. (2022). [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](). NIST Special Publication 1270

Contexts & Types of Bias

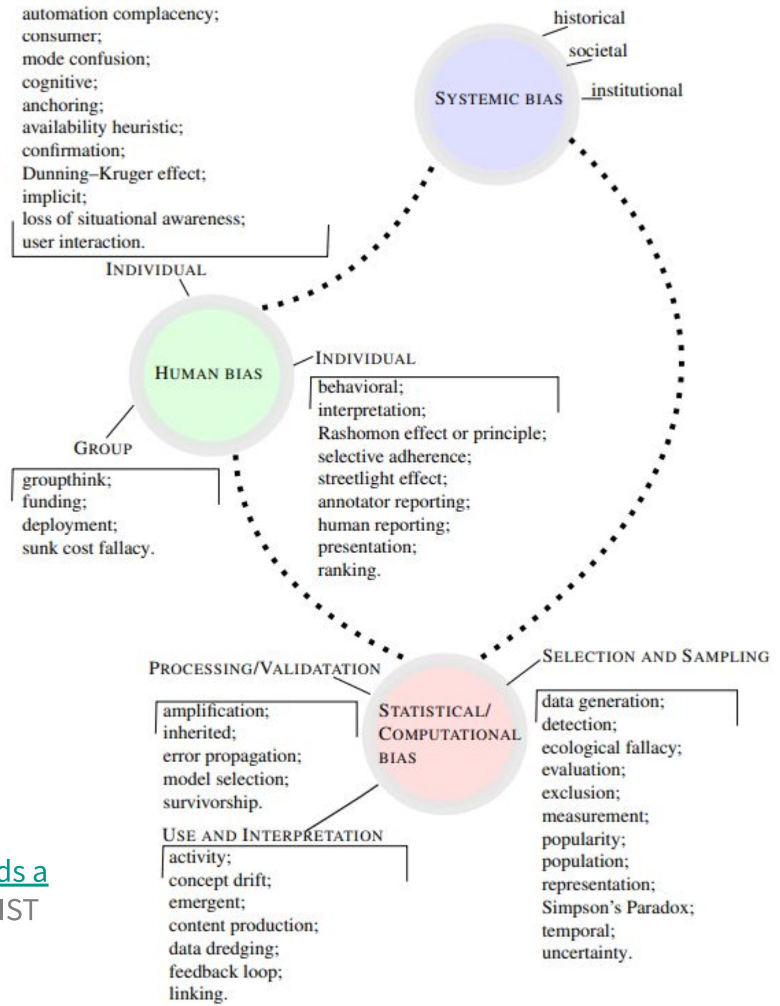
Contexts for addressing AI Bias

- Statistical
- Legal
- Cognitive and Societal

Types of AI Bias

- Systemic Bias
- Human Bias
- Statistical - Computational Bias

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt., A. (2022). [Towards a Standard for Identifying and Managing Bias in Artificial Intelligence](). NIST Special Publication 1270



Popular Fairness Definitions

- Equalized odds
- Equal opportunity
- Demographic (or statistical) parity
- Conditional statistical parity
- Treatment equality
- Test fairness

Group fairness

- Fairness through Awareness
- Fairness through Unawareness
- Counterfactual fairness

Individual fairness

- Diversity
- Fairness in relational domains
- Representational harms (e.g. bias ampl.)

other definitions

A. Narayanan (2018). “21 fairness definitions and their politics”. ACM FAT* 2018 tutorial

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Types of Harms as a Result of AI Bias

- Allocative Harms
    - When decision-making systems in criminal justice, health care, etc. are discriminatory, they create allocative harms, which are caused when a system withholds certain groups an opportunity or a resource.

banking, hiring, education, compensation

- Representational Harms
    - When systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc., they create stereotype perpetuation and cultural denigration.

news, social media, hate speech, disinformation, surveillance

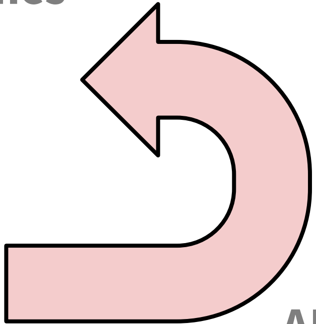
K. Crawford (2017). The Trouble with Bias, NIPS 2017 Keynote

Why AI Bias in (Social) Media Affects Us?

- **“Active” engagement:** Continuous consumption and sharing → information/news/entertainment → opinion formation → decision making
    - Purchasing behaviour
    - Stance in topics of public interest
    - Voting
    - Health habits

Collective outcomes

- **“Latent” impact:** Continuous profiling of individuals
    - Online activities
    - Physical world activities (surveillance)
    - Beliefs
    - Intentions



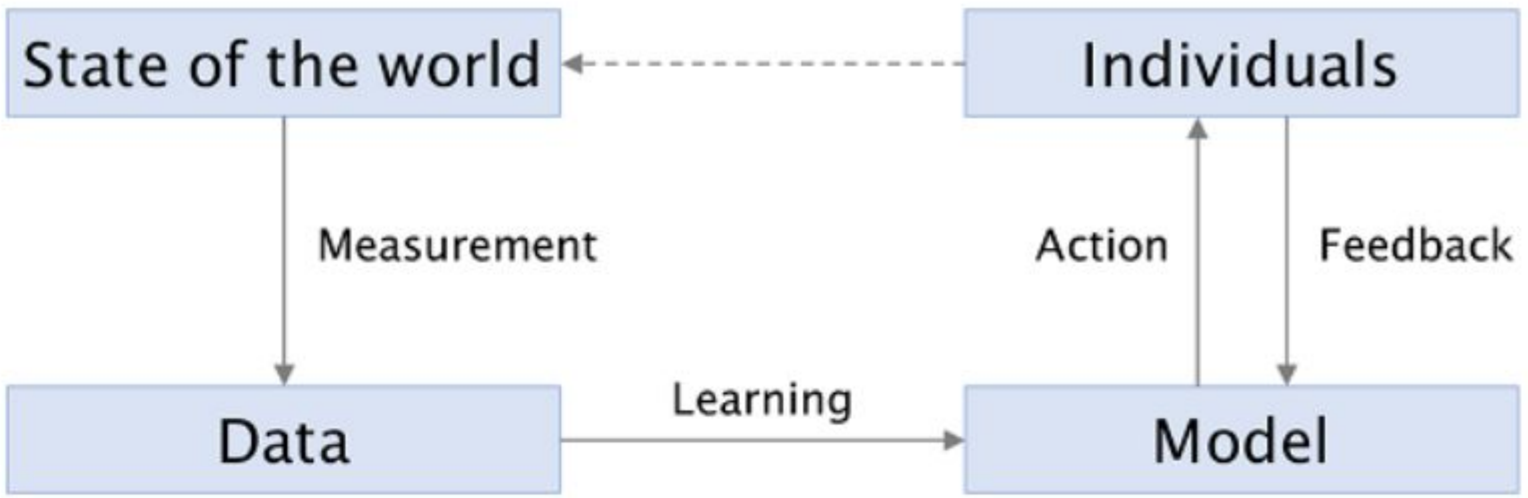
AI-mediated feedback loops

bias in visual datasets

a survey

The Machine Learning Loop

Internet users, crowd workers

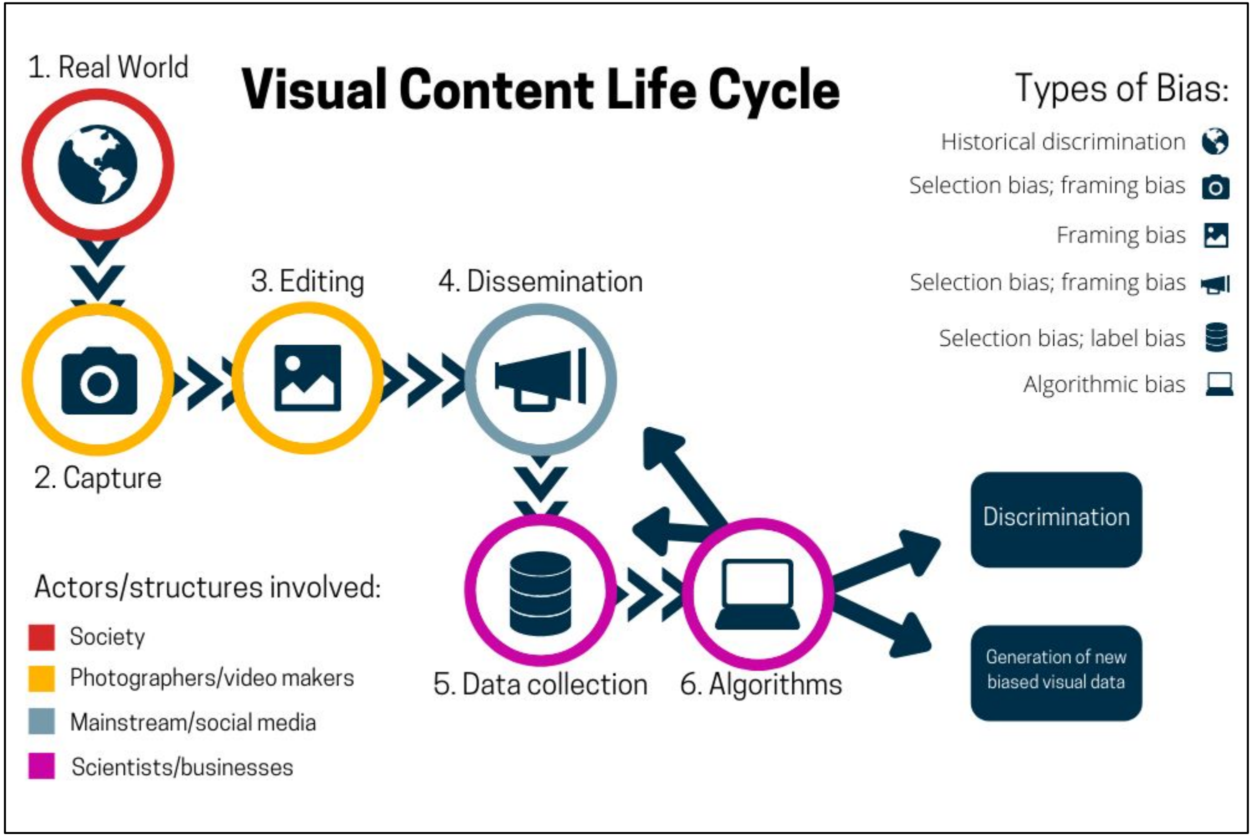


images, videos, text, labels

CNNs, ViTs, GANs

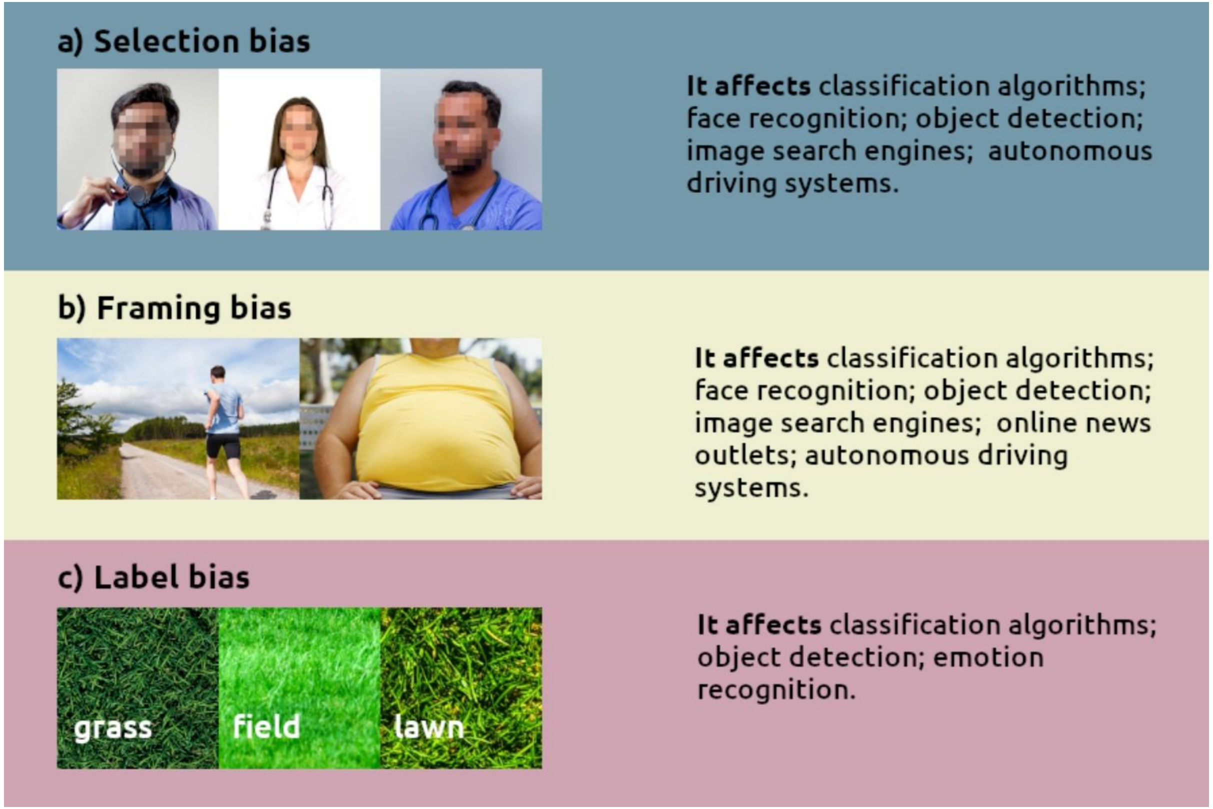
Barocas, S., Hardt, M., & Narayanan, A. (2021). Fairness and machine learning. Limitations and Opportunities.

The Media Bias Loop



Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2021). [A survey on bias in visual datasets](). *arXiv preprint arXiv:2107.07919*.

Visual Bias Taxonomy



any disparities or associations created as a result of the process by which subjects are included in a visual dataset

any associations or disparities that can be used to convey different messages and/or that can be traced back to the way in which the visual content has been composed.

any errors in the labelling of visual data, with respect to some ground truth, or the use of poorly defined or inappropriate semantic categories

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2021). [A survey on bias in visual datasets](). *arXiv preprint arXiv:2107.07919*.

Mapping Specific Types of Bias to the three overarching Visual Bias categories

Name	Description	Selection	Framing	Label
Sampling bias[*]	Bias that arises from the sampling of the visual data. It includes class imbalance.	•		
Negative set bias (Torralba and Efros, 2011)	When a negative class (say `non-white` in a `white/non-white` categorisation) is not representative enough.	•		•
Availability bias[†]	Distortion arising from the use of the most readily available data (e.g., using search engines).	•		
Platform bias	Bias that arises as a result of a data collection being carried out on a specific digital platform (e.g., Twitter, Instagram, etc.).	•		
Volunteer bias[†]	When data is collected in a controlled setting instead of being collected in-the-wild, volunteers that participate in the data collection procedure may differ from the general population.	•		
Crawling bias	Bias that arises as a result of the crawling algorithm/system used to collect images from the Web or with the use of an API (e.g., the keywords used to query an API, the seed websites used in a crawler).	•	•	
Spurious correlation	Presence of spurious correlations in the dataset that falsely associate a certain group of subjects with any other features.	•	•	
Exclusion bias[*]	Bias that arise when the data collection excludes partly or completely a certain group of people.	•	•	
Chronological bias[†]	Distortion due to temporal changes in the visual world the data is supposed to represent.	•	•	•
Geographical bias (Shankar et al., 2017)	Bias due to the geographic provenance of the visual content or of the photographer/video maker (e.g., brides and grooms depicted only in western clothes).	•	•	
Capture bias (Torralba and Efros, 2011)	Bias that arise from the way a picture or video is captured (e.g., objects always in the centre ,exposure, etc.).		•	
Apprehension bias[†]	Different behaviour of the subjects when they are aware of being photographed/filmed (e.g., smiling).		•	
Contextual bias (Singh et al., 2020)	Association between a group of subjects and a specific visual context (e.g., women and men respectively in household and working contexts)		•	
Stereotyping[§]	When a group is depicted according to stereotypes (e.g., female nurses vs. male surgeons).		•	
Measurement bias (Jacobs and Wallach, 2021)	Every distortion generated by the operationalisation of an unobservable theoretical construct (e.g., race operationalised as a measure of skin colour).			•
Observer bias[†]	Bias due to the way a annotator records the information.			•
Perception bias[†]	When data is labelled according to the possibly flawed perception of a annotator (e.g., perceived gender or race) or when the annotation protocol is not specific enough or is misinterpreted.			•
Automation bias[§]	Bias that arises when the labelling/data selection process relies excessively on (biased) automated systems.	•		•

Visual Bias Quantification Approaches

While the dataset bias literature is vast for other data types, for visual data it appears to be more limited. We review the relevant literature and found out four major categories of bias detection methods for visual data:

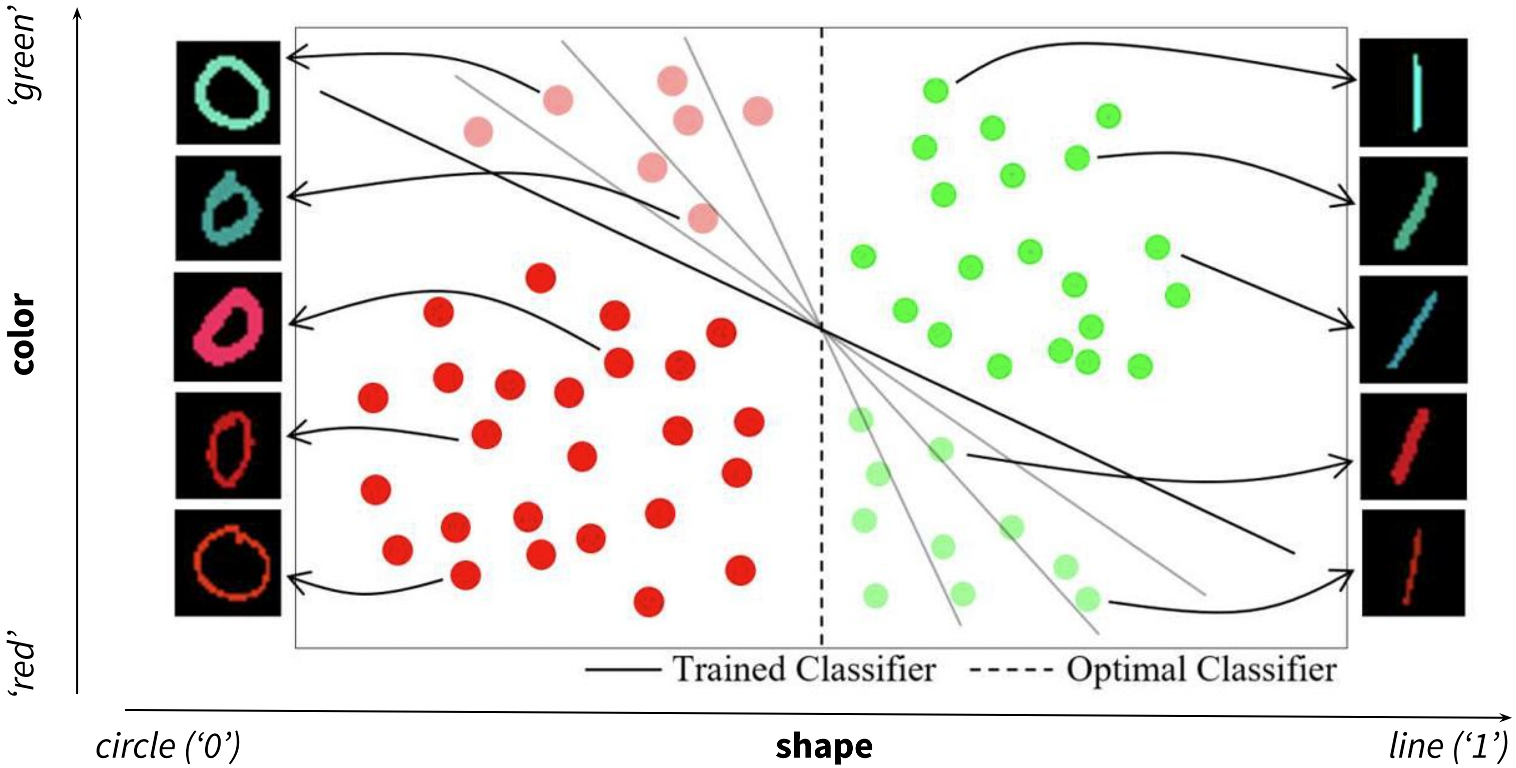
- Reduction to tabular data
  - Parity-based
  - Information theoretic
- Biased image representation
- Cross-dataset bias detection
- Other

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2021). A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919*.

Bias Discovery & Quantification Methods

	No.	Paper	Year	Selection	Framing	Label	Type of measures/methods
Reduction to tabular data	1	Dulhanty and Wong (2019)	2019	•	•		Count; Demographic parity
	2	Yang et al. (2020)	2020	•	•	•	Count; Demographic parity
	3	Zhao et al. (2017)	2017	•	•		Demographic parity
	4	Shankar et al. (2017)	2017	•			Count
	5	Buolamwini and Gebru (2018)	2018	•			Count
	6	Merler et al. (2019)	2019	•			Entropy-based; Information theoretical
	7	Panda et al. (2018)	2018	•	•		Entropy-based
	8	Kim et al. (2019)	2019	•	•		Information theoretical
	9	Wang et al. (2019)	2019	•	•		Dataset leakage
	10	Wachinger et al. (2021)	2021	•			Causality
	11	Jang et al. (2019)	2019		•		4 different measures
	12	Wang et al. (2020)	2020	•	•	•	13 different measures
Biased image representation	13	Kärkkäinen and Joo (2021)	2021	•			Distance-based
	14	Steed and Caliskan (2021)	2021		•		Distance-based
	15	Balakrishnan et al. (2020)	2020		•		Interventions
Cross-dataset bias detection	16	Torralba and Efros (2011)	2011	•	•		Cross-dataset generalisation
	17	Tommasi et al. (2015)	2015	•	•		Cross-dataset generalisation
	18	Khosla et al. (2012)	2012	•	•		Modelling bias
	19	López-López et al. (2019)	2019	•			Nearest neighbour in a latent space
Other	20	Model and Shamir (2015)	2015	•	•		Model-based
	21	Thomas and Kovashka (2019)	2019		•		Model-based
	22	Clark et al. (2020)	2020	•	•		Modelling bias
	23	Lopez-Paz et al. (2017)	2017	•			Causality
	24	Hu et al. (2020)	2020	•	•		Crowd-sourcing

Unknowns in the Visual Feature Space → Bias



Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). [Learning not to learn: Training deep neural networks with biased data](). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9012-9020).

Reduction of Visual to Tabular Data

parity-based

A standard technique for quantifying bias is to reduce the problem to tabular data.

- For example Zhao et al. (2017) measured the correlation between the occurrences of certain objects/activities with a protected attribute in a scene

$$b(o,g) = \frac{c(o,g)}{\sum_{g' \in G} c(o,g')} \qquad \frac{c(verb, \mathrm{man})}{c(verb, \mathrm{man}) + c(verb, \mathrm{woman})}.$$

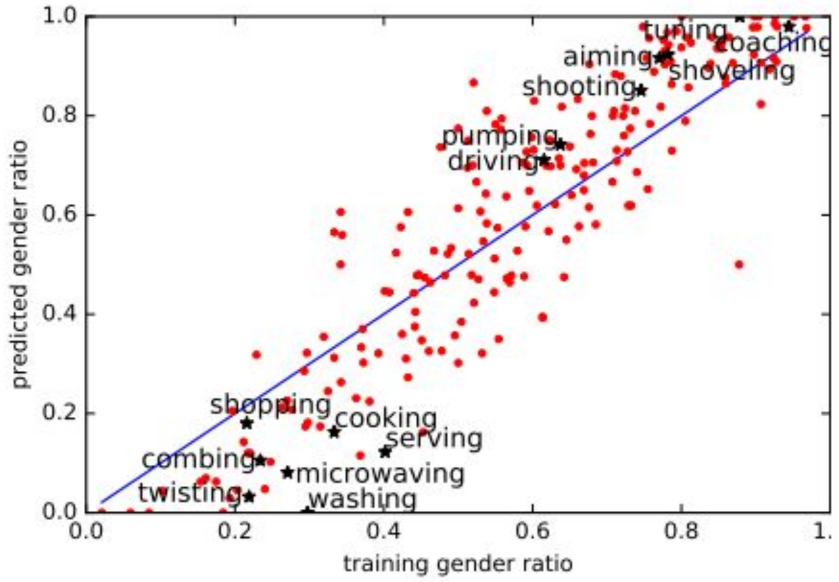
where *c(o,g)* is the number of co-occurrences between an object/activity *o* and the protected attribute value *g* (e.g. man/woman)

In a popular dataset such as MS-COCO, men are more likely associated with sports-related objects while women are more likely associated with kitchen objects.

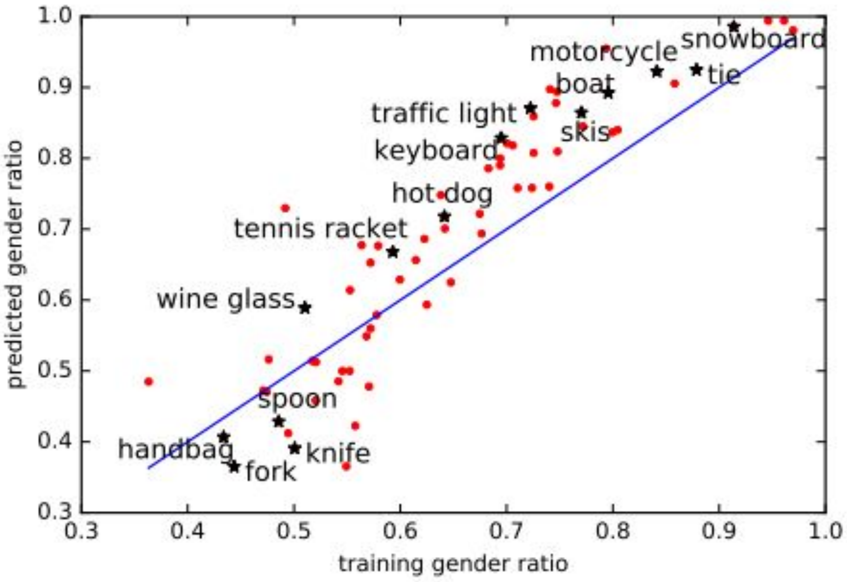
Zhao, J., et al., (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Reduction of Visual to Tabular Data

parity-based



(a) Bias analysis on imSitu vSRL



(b) Bias analysis on MS-COCO MLC

Zhao, J., et al., (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

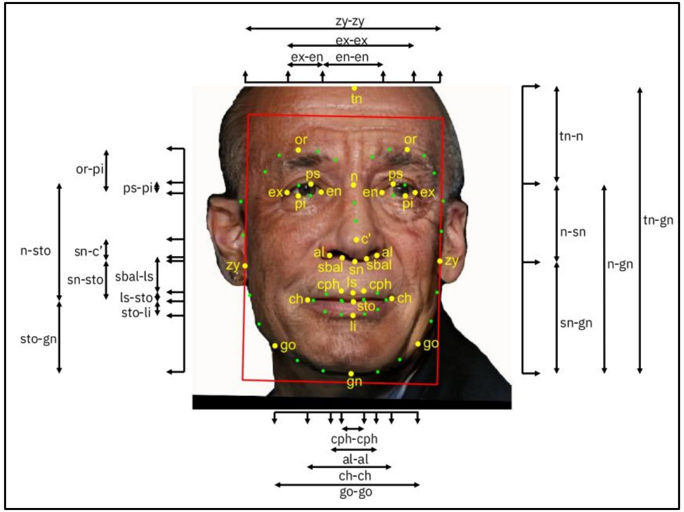
Reduction of Visual to Tabular Data

information theoretic

Bias in CV tasks such as face recognition might be due to limited coverage/representativeness of the training set. To increase the variety and "coverage" of the training set, one would like to achieve high **diversity**. If attributes are available in tabular form, information theoretic techniques can be used to measure diversity.

- Merler et al. (2019) applied information-theoretic measures (e.g., Shannon entropy) to facial attributes (e.g., skin colour, craniofacial distances, gender, etc.) to ensure diversity in the data they collected. (S is the number of attribute values and p_i is the probability of an image to have the attribute i)

Diversity	**Evenness**
Shannon $H = -\sum_{i=1}^{S} p_i * ln(p_i)$	Shannon $E = \frac{H}{ln(S)}$
Simpson $D = \frac{1}{\sum_{i=1}^{S}(p_i * p_i)}$	Simpson $E = \frac{D}{S}$

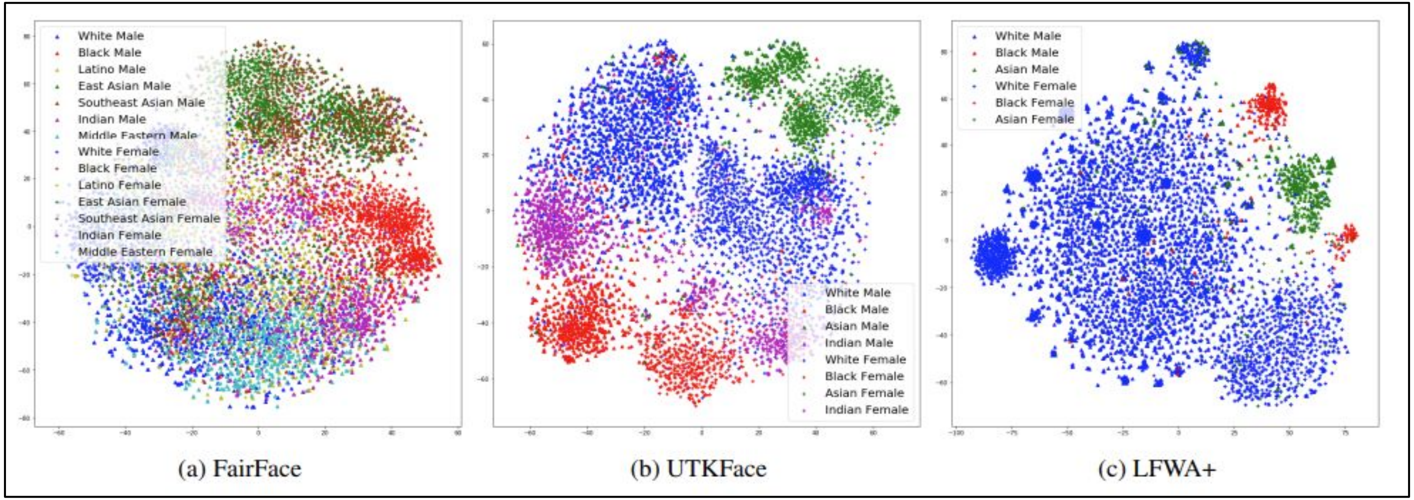


Merler, M., et al., (2019). Diversity in Faces. *arXiv pre-print*, arXiv:1901.10436.

Low-dimensional Visual Representations

Another strategy is to measure bias in a lower dimensional representation space and measure separability and coverage of the space.

These approaches rely on the assumption that the projection onto the representation space is reasonably unbiased.



t-SNE visualizations of ResNet-34 face embeddings

Kärkkäinen, K., and Joo, J., (2021). [Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation](). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558.

Low-dimensional Visual Representations

Some works are inspired from similar work in NLP

- Steed and Caliskan (2021) devised a version of an Image Association Test to be applied to image representations. The association were measured in terms of the cosine similarity of the representation vectors.

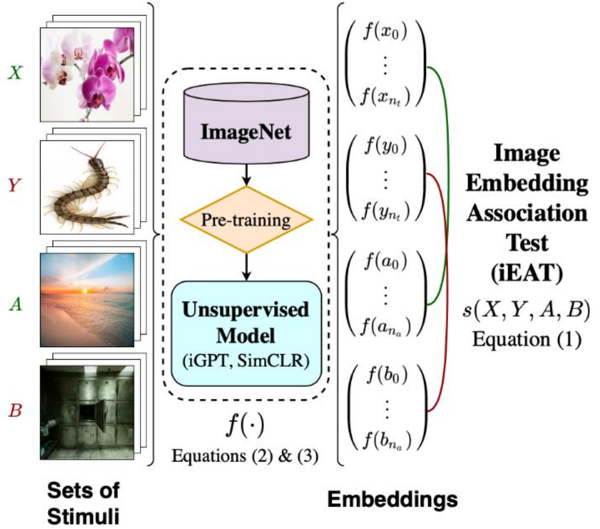


Figure 2: Example iEAT replication of the Insect-Flower IAT [31], which measures the differential association between flowers vs. insects and pleasantness vs. unpleasantness.

Steed, R., and Caliskan, A., (2021). [Image representations learned with unsupervised pre-training contain human-like biases](). In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 701–713.

Cross-dataset Bias Detection

The first attempts to discovering biases in image datasets were done by comparing different datasets.

- Torralba and Efros (2011) found out that it is easy for an algorithm to classify images according to their appearance in different benchmarks.
- They also looked at how badly a classification algorithm trained on a given dataset generalises to other benchmarks.
  The worse the generalisation, the greater the bias (but does not necessarily imply higher discrimination).

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528.

Human-in-the-Loop for Bias Assessment

- Step 1: crowd workers inspect images and try to identify similarities between them and attributes that are responsible for these similarities in the form of questions
- Step 2: crowd workers are asked to answer some questions from step 1 for a different sample of images
- Step 3: crowd workers are asked whether statements coming from step 2 correspond to the real world

Figure 2: Inputs and outputs of the workflow in our two evaluation studies. Top panel: sample images of the image datasets used in Study 1 (the airplane dataset) and Study 2 (the car dataset); bottom panel: Top 10 "biases" with distinct meanings that are detected by the crowd using our workflow for each dataset. Each bias is coded into one of the 4 categories: Known bias (KB), additional bias (AB), unbiased similarity (US) or unrelated (U). KB and AB are considered correct detection of sampling biases (highlight in green), while US and U are considered incorrect detection (highlight in red).

Hu, X., et al. (2020). [Crowdsourcing detection of sampling biases in image datasets](). In *Proceedings of The Web Conference 2020, WWW '20*, page 2955–2961.

Visual Bias Quantification Approaches

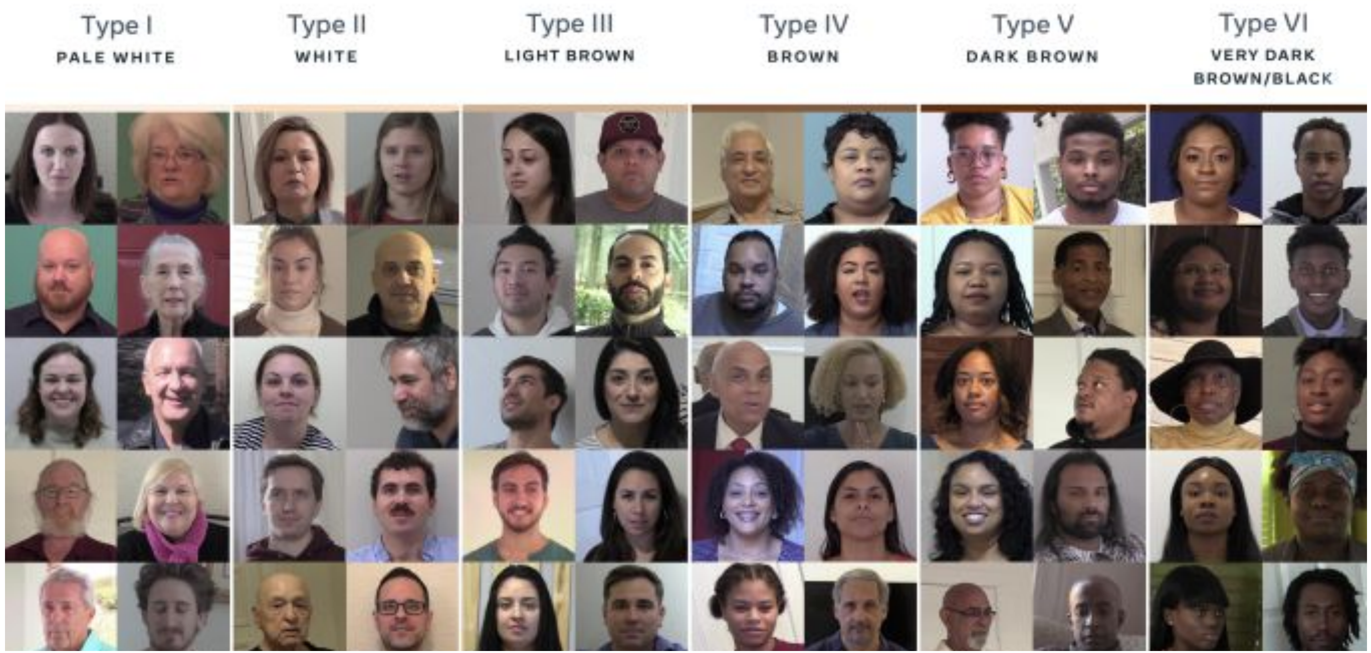
Pros and Cons

- Reduction to tabular data
  - + Tabular data are much easier to work with and the wealth of fairness toolkits can be leveraged
  - - The reduction to tabular data might introduce bias or over-simplify
- Biased image representation
  - + In theory, they should preserve more of the complexity/nuance of visual content
  - - Depend a lot on embedding/projection and similarity function
- Cross-dataset bias detection
  - - Only applicable when multiple datasets are available
  - - Give little insight with respect to the type of bias
- Other
  - - Depend a lot on the domain/task under consideration.
  - - Human-in-the-loop approaches are expensive and require very careful design.

Bias-aware Visual Datasets

- Pilot Parliaments Benchmark (PPB) dataset (used in Gender Shades paper) balanced in terms of gender and skin color
- FairFace (Kärkkäinen & Joo, 2021) contains 108,500 images containing faces of people from 7 races
- Diversity in Faces (Merler et al., 2021) contains almost one million face images from YFCC100m and annotating them in terms of cranio-facial features, age, gender, skin
- KANFace (Georgopoulos et al., 2020) consists of 40K still images and 44K videos (14.5M frames in total) from 1,045 subjects captured in real-world conditions
- Casual Conversations (Hazirbas et al., 2021) is composed of over 45,000 videos (3,011 participants) and intended to be used for assessing the performance of already trained models in computer vision and audio applications
- ObjectNet (Barbu et al., 2019) a large real-world test set for object recognition with control where object backgrounds, rotations, and imaging viewpoints are random

Casual Conversations Dataset



Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A. and Ferrer, C.C., 2021. [Towards measuring fairness in AI: the Casual Conversations dataset](). *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

The Trouble with CV Datasets

- Numerous ethical issues and controversial practices in the collection, curation and labelling of web-scale image-text datasets
- Many types of harms:
  - harmful stereotypes
  - inappropriate/NSFW content
  - privacy intrusion

Birhane, A., & Prabhu, V. U. (2021, January). [Large image datasets: A pyrrhic win for computer vision?](). In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1536-1546). IEEE.

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). [Multimodal datasets: misogyny, pornography, and malignant stereotypes](). *arXiv preprint arXiv:2110.01963.*

https://excavating.ai/

LARGE DATASETS: A PYRRHIC WIN FOR COMPUTER VISION?

arXiv:2110.01963v1 [cs.CY] 5 Oct 2021

Multimodal datasets: misogyny, pornography, and malignant stereotypes

Abeba Birhane* University College Dublin & Lero, Dublin, Ireland, abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu* Independent Researcher, vinaypra@alumni.cmu.edu

Emmanuel Kahembwe University of Edinburgh, Edinburgh, UK, e.kahembwe@ed.ac.uk

Abstract

We have now entered the era of trillion parameter machine learning models trained on billion-sized datasets scraped from the internet. The rise of these gargantuan datasets has given rise to formidable bodies of critical work that has called for caution while generating these large datasets. These address concerns surrounding the dubious curation practices used to generate these datasets, the sordid quality of alt-text data available on the world wide web, the problematic content of the CommonCrawl dataset often used as a source for training large language models, and the entrenched biases in large-scale visio-linguistic models (such as OpenAI's CLIP model) trained on opaque datasets (WebImageText). In the backdrop of these specific calls of caution, we examine the recently released LAION-400M dataset, which is a CLIP-filtered dataset of Image-Alt-text pairs parsed from the Common-Crawl dataset. We found that the dataset contains, troublesome and explicit images and text pairs of rape, pornography, malign stereotypes, racist and ethnic slurs, and other extremely problematic content. We outline numerous implications, concerns and downstream harms regarding the current state of large scale datasets while raising open questions for various stakeholders including the AI community, regulators, policy makers and data subjects.

Warning: This paper contains NSFW content that some readers may find disturbing, distressing, and/or offensive.

1 Introduction

The emergence of deep learning aided computer vision as a notable field of Artificial Intelligence (AI) ushered the so-termed *AI spring* [1] and has been characterized by its voracious need for vast volumes of data. The recent multi-modality drive within AI seeks to break away from the template of training siloed task-specific models for image classification, segmentation, or detection and entails curating cross-domain datasets and training cross-domain models that will jointly model the modalities of *vision*, *text*, and *speech* data. In the specific context of the *vision-text dyad*, the endeavor begins with curating large-scale datasets of tuples of the form: $D = \{(x_i, t_i, \mu_i)\}_{i=1}^{N}$ where x_i is the i^{th} image, t_i is the textual description associated with the i^{th} image, and μ_i is the i^{th} image's meta-data. As has been the case with much of state-of-the-art (SotA) AI endeavors [2, 3], the dataset is expected to be

*Equal contribution

Preprint.

addressing visual bias

aka fairness-aware learning in visual content

Bias in Data-driven AI Systems

NOBIAS

UNDERSTANDING BIAS

Socio-technical causes of bias
- Data generation
- Data collection
- Institutional bias

Bias manifestation in data
- Sensitive features & causal inferences
- Data representativeness
- Data modalities

Fairness definition
- Similarity-based
- Causal reasoning
- Predicted outcome
- Predicted & actual outcome
- Predicted probabilities & actual outcome

MITIGATING BIAS

Pre-processing
- Class label modification
- Instance selection
- Instance weighting

In-processing
- Classification model adaptation
- Regularization / Loss function s.t. constraints
- Latent fair classes

Post-processing
- Confidence/probability score corrections
- Promoting/demoting boundary decisions
- Wrapping a fair classifier on top of black-box base classifier

ACCOUNTING FOR BIAS

Bias-aware data collection
- Bias elicitation: individual assessors, mathematical pooling, group elicitation, consensus building
- Crowdsourcing

Describing and modelling bias
- Description and causal logics
- Ontological formalisms and reasoning

Explaining AI decisions
- Model explanation by approximation
- Inherently interpretable models
- Local behaviour explanation

LEGAL ISSUES

Regulations provisions
- Data accuracy (GDPR)
- Equality, prohibition of discrimination (CFR-EU)

Are data modifications legal?
- Intellectual Property issues
- Legal basis for data/model modification

Application of existing rules
- Applicability to algorithmic decision-making
- Limited scope of anti-discrimination law. Indirect discrimination

Ntoutsi, E., et al (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

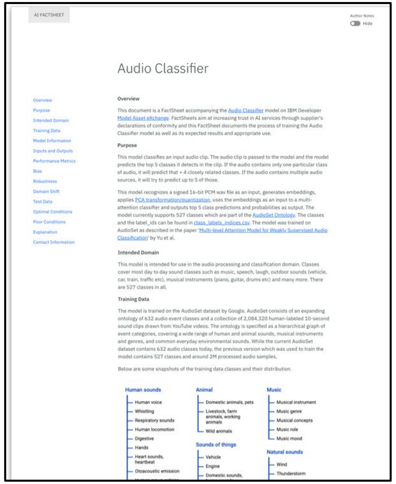
How to Address Bias in Visual Data

- Transparency
  - document and bring forward
- Proactive approaches / Check Lists
  - avoid at creation time
- Algorithmic bias mitigation
- Fairness Toolkits

Transparency

Information Sheets and Model Cards

- Datasheets for Datasets (Microsoft): seminal work on dataset transparency
- Model cards (Google): Based on seminal work by (Mitchell et al., 2019)
- AI FactSheets 360 (IBM): offers a variety of example templates



AI FACTSHEET

Author Notes

Hide

Audio Classifier

Overview

This document is a FactSheet accompanying the Audio Classifier model on IBM Developer Model Asset eXchange. FactSheets aim at increasing trust in AI services through supplier's declarations of conformity and this FactSheet documents the process of training the Audio Classifier model as well as its expected results and appropriate use.

Purpose

This model classifies an input audio clip. The audio clip is passed to the model and the model predicts the top 5 classes it detects in the clip. If the audio contains only one particular class of audio, it will predict that + 4 closely related classes. If the audio contains multiple audio sources, it will try to predict up to 5 of those.

This model recognizes a signed 16-bit PCM wav file as an input, generates embeddings, applies PCA transformation/quantization, uses the embeddings as an input to a multi-attention classifier and outputs top 5 class predictions and probabilities as output. The model currently supports 527 classes which are part of the AudioSet Ontology. The classes and the label_ids can be found in class_labels_indices.csv. The model was trained on AudioSet as described in the paper 'Multi-level Attention Model for Weakly Supervised Audio Classification' by Yu et al.

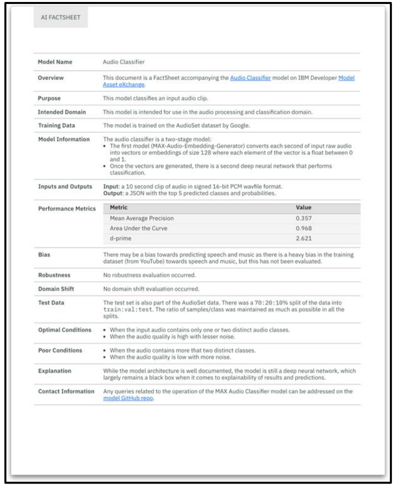
Intended Domain

This model is intended for use in the audio processing and classification domain. Classes cover most day to day sound classes such as music, speech, laugh, outdoor sounds (vehicle, car, train, traffic etc), musical instruments (piano, guitar, drums etc) and many more. There are 527 classes in all.

Training Data

The model is trained on the AudioSet dataset by Google. AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds. While the current AudioSet dataset contains 632 audio classes today, the previous version which was used to train the model contains 527 classes and around 2M processed audio samples.

Below are some snapshots of the training data classes and their distribution.



AI FACTSHEET

Model Name	Audio Classifier
Overview	This document is a FactSheet accompanying the Audio Classifier model on IBM Developer Model Asset eXchange.
Purpose	This model classifies an input audio clip.
Intended Domain	This model is intended for use in the audio processing and classification domain.
Training Data	The model is trained on the AudioSet dataset by Google.
Model Information	The audio classifier is a two-stage model: • The first model (MAX-Audio-Embedding-Generator) converts each second of input raw audio into vectors or embeddings of size 128 where each element of the vector is a float between 0 and 1. • Once the vectors are generated, there is a second deep neural network that performs classification.
Inputs and Outputs	**Input**: a 10 second clip of audio in signed 16-bit PCM wavfile format. **Output**: a JSON with the top 5 predicted classes and probabilities.
Performance Metrics	**Metric** / **Value** Mean Average Precision: 0.357 Area Under the Curve: 0.968 d-prime: 2.621
Bias	There may be a bias towards predicting speech and music as there is a heavy bias in the training dataset (from YouTube) towards speech and music, but this has not been evaluated.
Robustness	No robustness evaluation occurred.
Domain Shift	No domain shift evaluation occurred.
Test Data	The test set is also part of the AudioSet data. There was a `70:20:10%` split of the data into `train:val:test`. The ratio of samples/class was maintained as much as possible in all the splits.
Optimal Conditions	• When the input audio contains only one or two distinct audio classes. • When the audio quality is high with lesser noise.
Poor Conditions	• When the audio contains more that two distinct classes. • When the audio quality is low with more noise.
Explanation	While the model architecture is well documented, the model is still a deep neural network, which largely remains a black box when it comes to explainability of results and predictions.
Contact Information	Any queries related to the operation of the MAX Audio Classifier model can be addressed on the model GitHub repo.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Hind, M., Houde, S., Martino, J., Mojsilovic, A., Piorkowski, D., Richards, J., & Varshney, K. R. (2020). Experiences with improving the transparency of AI models and services. In *Extended Abstracts of 2020 CHI Conf. on Human Factors in Computing Systems* (pp. 1-8).

Check Lists

- Deon: A command-line tool for adding ethics checklists to data science projects (includes fairness and bias aspects as part of the default list)
- AI Fairness Checklist (Microsoft): a checklist co-designed with practitioners, incl. how organizational/team processes shape how AI teams address fairness harms
- Legal and Ethical Checklist for AI Systems: this checklist is sectioned by legal priorities, incl. human agency & oversight, security & safety, privacy & data governance, transparency, accessibility, etc.

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proc. of 2020 CHI Conf. on Human Factors in Computing Systems* (pp. 1-14).

Lifshitz, L. R., & McMaster, C. (2020). Legal and Ethics Checklist for AI Systems. *SciTech Lawyer*, *17*(1), 28-34.

Visual Dataset Bias CheckList

We proposed a checklist to help scientist and practitioners to spot possible biases in the visual data they collect. The CheckList is organized in four main parts:

- General
- Selection bias
- Framing bias
- Label bias

Our questions are partly inspired by works on reflective data practices (Gebru et al., 2021; Jacobs & Wallach, 2021)

Gebru, T., et al. (2021). Datasheets for datasets. *Communications of the ACM, December 2021,* Vol. 64 No. 12, Pages 86-92.

Jacobs, A. Z. and Wallach., H. Measurement and fairness. I*n Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 375–385

Visual Dataset Bias CheckList

General	What are the purposes the data is collected for?
	Are there uses of the data that should be discouraged because of possible biases?
	What kind of bias can be inserted by the way the collection process is designed?
Selection	Do we need balanced data or statistically representative data?
	Does the selection of the subjects create any spurious associations?
	Is the dataset representative enough? Are the negative sets representative enough?
	Is there any group of subjects that is systematically excluded from the data?
	Do the data come from or depict a specific geographical area?
	Will the data remain representative for a long time?
Framing	Are there any spurious correlation that can contribute to framing different subjects in different ways?
	Are there any biases due to the way images/videos are captured?
	Did the capture induce some behaviour in the subjects (e.g. smiling when photographed)?
	Are there any images that can possibly convey different messages depending on the viewer?
	Are subjects of a certain group depicted in a particular context more often than others?
	Do the data agree with harmful stereotypes?
Label	If the labelling process relies on machines: have their biases been taken into account?
	If the labelling process relies on human annotators: is there an adequate and diverse pool of annotators? Have their possible biases been taken into account?
	If the labelling process relies on crowdsourcing: are there any biases due to the workers' access to crowd platforms?
	Do we use fuzzy labels? (e.g, race or gender)
	Do we operationalise any unobservable theoretical constructs/use proxy variables? (Jacobs & Wallach, 2021)

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2021). A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919*.

Bias Mitigation

- Pre-processing
  - Instance selection and/or weighting (Stone et al., 2022)
  - Instance label modification/massaging
  - Synthetic instance generation (incl. augmentation, GANs, etc.)
- In-processing
  - Regularization, Multi-task learning (Das et al., 2018)
  - Constraints
  - Training on latent variables
  - Adversarial debiasing (Kim et al., 2019, Wang et al., 2019)
- Post-processing
  - Confidence score correction
  - Class label correction
  - Decision boundary change

Epistemic uncertainty-weighted loss function for sample weighting.

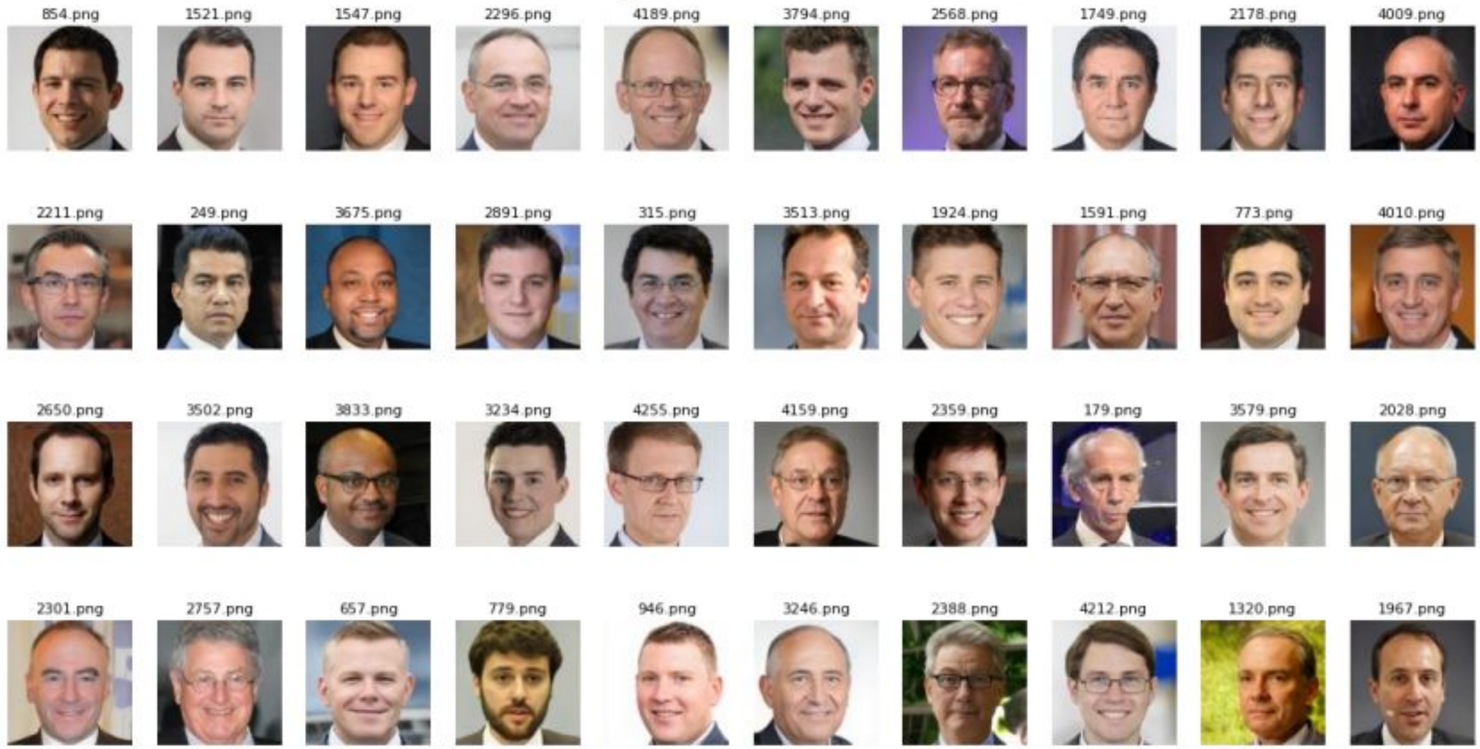
MTCNN with dynamic loss weight adjustment for three tasks

Minimize mutual information between feature embedding and target bias by adversially unlearning.

Adversarially train critic model on gender-related loss vs a task specific model

Bias in StyleGAN2

Top 40 generated images in terms of GIQA



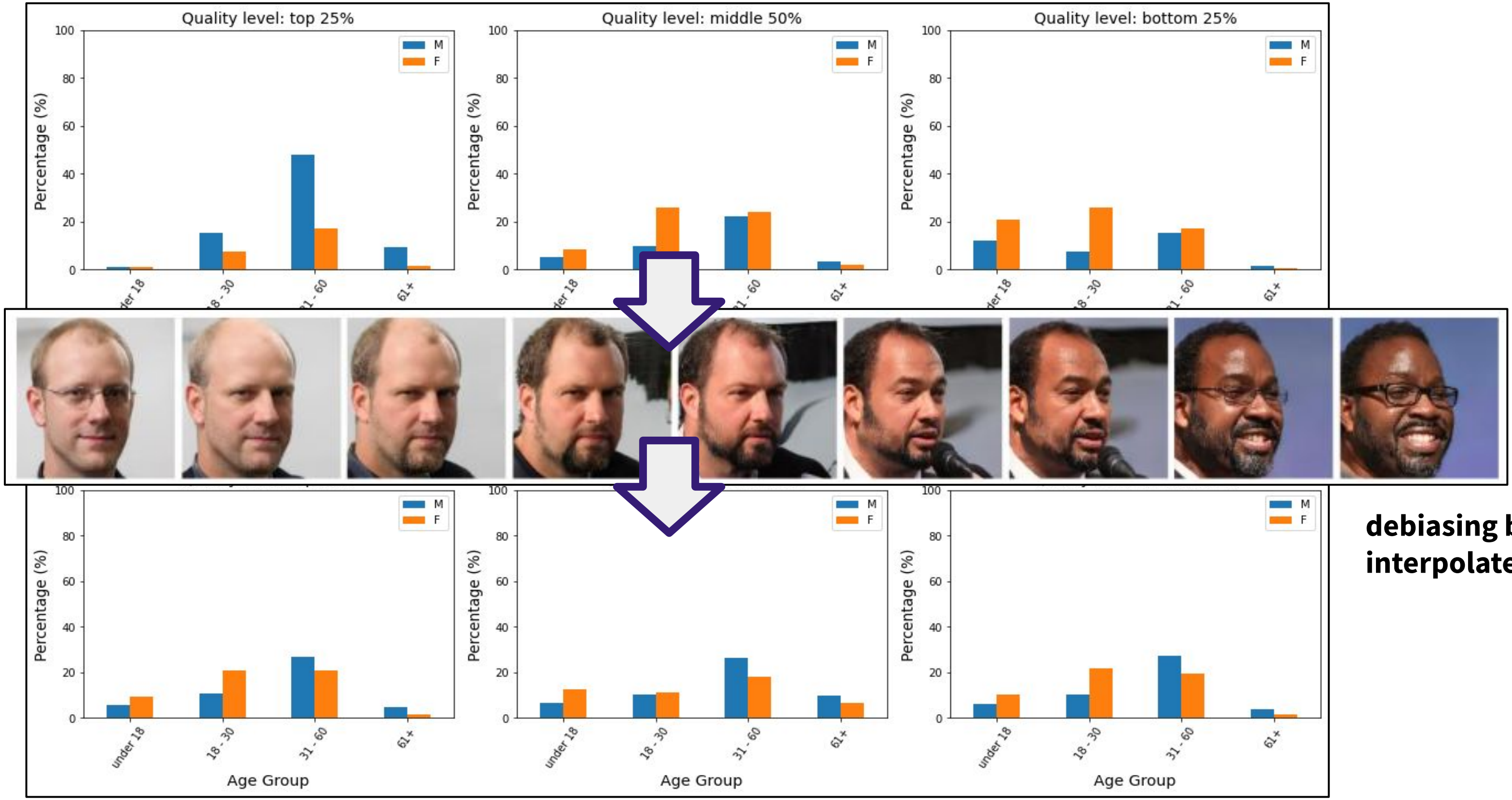
Maragkoudakis, E. (2022). “Study of bias in face synthesis methods” Bachelor Thesis in Harokopio University of Athens

Bias in StyleGAN2

Bottom 40 generated images in terms of GIQA

Maragkoudakis, E. (2022). “Study of bias in face synthesis methods” Bachelor Thesis in Harokopio University of Athens

Distribution of Quality vs Protected Attributes



debiasing by sampling on the interpolated z-space

Maragkoudakis, E. (2022). “Study of bias in face synthesis methods” Bachelor Thesis in Harokopio University of Athens

Fairness Software Toolkits

Reducing visual data as tabular

- AI Fairness 360 (IBM): arguably the most popular fairness toolkit
- FairLearn (originally Microsoft): comparable to AI Fairness 360
- TensorFlow Fairness Indicators (Google): emphasis on large scale applications
- TensorFlow What-If Tool (Google): emphasis on interpretation/exploration
- Aequitas (U Chicago): includes a web audit tool
- LiFT (LinkedIn): emphasis on large-scale machine learning workflows
- audit-AI (Pymetrics): regulatory compliance and checks for practical/statistical bias
- algofairness (Haverford C.): contains fairness-comparison & BlackBoxAuditing
- ML-fairness-gym (Google): enables the study of ML impact via social simulations

Richardson, B., & Gilbert, J. E. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *arXiv preprint arXiv:2112.05700*.

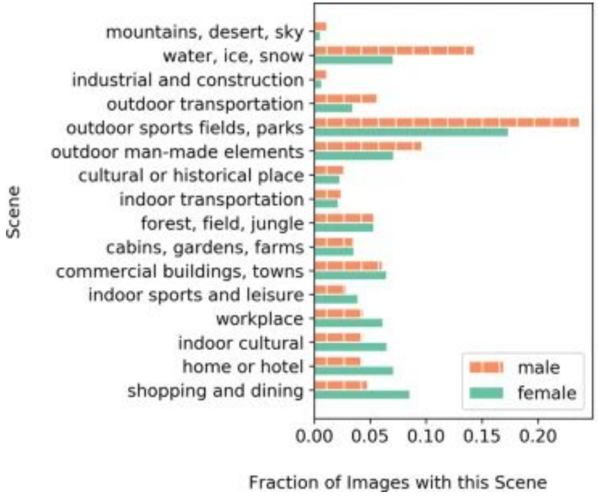
https://www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/

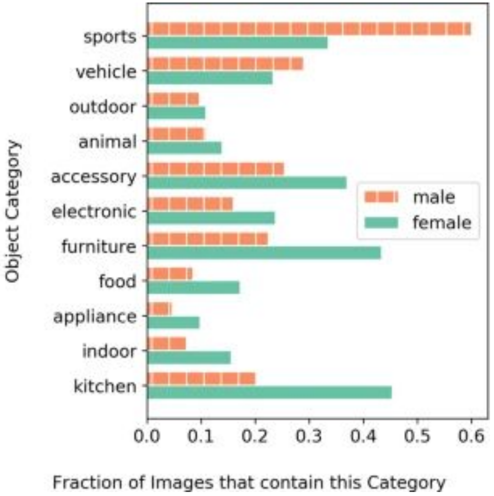
REVISE

a tool for measuring and mitigating bias in visual datasets

Input: image dataset → Output: metrics along person, object, geography

- Object
  - Object counts
  - Duplicate annotations
  - Object scale
  - Object co-occurrence
  - Scene diversity
  - Appearance diversity
- Person
  - Person prominence
  - Contextual representation
  - Instance counts and differences
  - Appearance differences
- Geography
  - Geographic distribution
  - Geography by object/people/language/income/weather





Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., ... & Russakovsky, O. (2022). [REVISE: A tool for measuring and mitigating bias in visual datasets]. *International Journal of Computer Vision*, 1-21. https://github.com/princetonvisualai/revise-tool

Conclusions

- AI Bias in Visual Data - while a specific area of AI Bias - raises many new challenges, incl. how to define bias considering the whole lifecycle of media data and their impact on individuals and society
  - Big multimodal datasets in the spotlight
- Different types of quantifying visual AI bias, with reduction to tabular and low-dimensional representations being the most common
- Approaches and toolsets for addressing bias in tabular data are useful but not sufficient → new methods emerge and new tools needed

Open Questions / Future Work

- Good ways of quantifying **visual framing** bias: important for assessing and auditing media and social media outlets
- Bias in **generative models**: recent big models like DALL-E 2 and Imagen consider it, but still no comprehensive or standardized assessment out there
- **Label bias** is much less studied: definition of labels, comprehensiveness of human and machine annotations, free-text captions, etc.
- Conceptual and formalization work: what is a **good overall definition** for visual bias? What are good **operational** measures? What are good ways to describe visual bias beyond numerical indicators?

Acknowledgements

Simone Fabbrizzi
(NoBIAS Fellow, CERTH)

Alaa Elobaid
(NoBIAS Fellow, CERTH)

Eirini Ntoutsi
(prof. FUB)

Yiannis Kompatsiaris
(research director CERTH)

AI4media

NOBIAS

Manios Krasanakis
(researcher CERTH)

Christos Diou
(assist. prof. HUA)

Thank you!

Symeon Papadopoulos

@sympap / papadop@iti.gr

mever.iti.gr

Visual Bias Taxonomy

Name	Selection bias	Framing Bias	Label Bias
Sampling bias	✅		
Platform bias	✅		
Chronological bias	✅	✅	✅
Spurious correlation	✅	✅	
Stereotyping		✅	
Measurement bias			✅
Automation bias	✅		✅

Visual Dataset Bias CheckList

Selection Bias	Do we need balanced data or statistically representative data?
	Does the selection of the subjects create any spurious associations?
	Is the dataset representative enough? Are the negative sets representative enough?
	Is there any group of subjects that is systematically excluded from the data?
	Do the data come from or depict a specific geographical area?
	Will the data remain representative for a long time?

Visual Dataset Bias CheckList

Framing Bias	Are there any spurious correlation that can contribute to framing different subjects in different ways?
	Is there any biases due to the way images/videos are captured?
	Did the capture induce some behaviour in the subjects (e.g. smiling when photographed)?
	Are there any images that can possibly convey different messages depending on the viewer?
	Are subjects of a certain group depicted in a particular context more often than others?
	Do the data agree with harmful stereotypes?

Visual Dataset Bias CheckList

Label Bias	If the labelling process relies on machines: have their biases been taken into account?
	If the labelling process relies on human annotators: is there an adequate and diverse pool of annotators? Have their possible biases been taken into account?
	If the labelling process relies on crowd sourcing: are there any biases due to the workers' access to crowd sourcing platforms?
	Do we use fuzzy labels? (e.g, race or gender)
	Do we operationalise any unobservable theoretical constructs/use proxy variables? (Jacobs & Wallach, 2021)

Jacobs, A. Z. and Wallach., H. Measurement and fairness. I*n Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 375–385

Popular Fairness Definitions (2/2)

- **Treatment equality:** treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories
- **Test fairness:** for any predicted probability score S, people in both protected and unprotected groups must have equal probability of correctly belonging to the positive class
- **Counterfactual fairness:** a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group
- **Fairness in relational domains:** capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Fairness Metrics

- Statistical bias
- Group fairness (demographic parity, equal pos./neg. pred. Value, equal FPR/FNR, accuracy equity)
- Blindness
- Individual fairness (equal thresholds, similarity metric)
- Process fairness (feature rating)
- Diversity
- Representational harms (stereotype mirroring/exaggeration, cross-dataset generalization, bias in representation learning, bias amplification)

A. Narayanan (2018). "21 fairness definitions and their politics". ACM FAT* 2018 tutorial

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.