Fairness in visual recognition

Olga Russakovsky

http://visualai.princeton.edu
Computer vision model learns to “increase attractiveness” by manipulating skin color

**THE VERGE**  April 25, 2017

“Machines taught by photos learn a sexist view of women”

**WIRED**  Aug 21, 2017

“Facial recognition is accurate, if you’re a white guy”

**The New York Times**  Feb 9, 2018

Cropping algorithm prefers “lighter, slimmer, younger faces”

**The Guardian**  Aug 10, 2021

See: Ruha Benjamin’s “Race after technology” — an excellent book
Can we adjust the AI design to mitigate these effects?

See: Ruha Benjamin’s “Race after technology” — an excellent book
Can we adjust the AI design to **mitigate** these effects?

![Diagram](https://i.imgur.com/3Q5Q5.png)

- Human history, bias, prejudice
- AI decision making
- AI models
- Large-scale data
Large scale ≠ fair representation

Demographic diversity

Image search diversity

Geographic diversity

(in ImageNet and OpenImages)

Data collection practices: concerns and interventions

[Timnit Gebru et al. CACM’21 “Datasheets for datasets.”]
[Eun Jo and Timnit Gebru. FAT’20 “Lessons from archives: strategies…”]
[Morgan Scheuerman et al. CSCW’21 “Do Datasets Have Politics?”]
[Amandalynne Paullada et al. Patterns’21 “Data and its (Dis)contents…”]
[Abeba Birhane et al. arxiv’21 “Multimodal Datasets: Misogyny, Pornography…”]
[Abeba Birhane et al. arxiv’21 “The Values Encoded in Machine Learning…”]
[Vinay Prabhu and Abeba Birhane. WACV’21 “Large datasets: a pyrrhic win…”]

[Joy Buolamwini and Timnit Gebru. FAT’18 “GenderShades: Intersectional accuracy…”]
[Shreya Shankar et al. NeurIPS Workshop’17 “No classification without representation…”]
[Terrance DeVries et al. CVPR Workshop’19 “Does object recognition work for everyone”]

[Matthew Kay et al. CHI’15 “Unequal representation and gender stereotypes…”]
[Safiya Noble. NYU Press’18 “Algorithms of Oppression: How search engines…”]

[Bernard Koch et al. NeurIPS D&B track’21 “Reduced, Reused and Recycled..”]
[Kenny Peng et al. NeurIPS D&B track’21 “Mitigating dataset harms requires…”]
[Margot Hanley et al. NeurIPS workshop’21 “An Ethical Highlighter…”]
[Milagros Miceli et al. FAccT’21 “Documenting Computer Vision Datasets…”]
[Ben Hutchinson et al. FAccT’21 “Towards Accountability for Machine…”]
[Kaiyu Yang et al. FAT’20 “Towards fairer datasets: filtering and…”]
[Kaiyu Yang et al. ICML’22 “A study of face obfuscation in ImageNet”]
REVISE: REvealing VIsual biaSEs tool

**Goal:** Develop a tool that inputs a visual dataset and reveals potential social biases

**Contributions:**

1. Aids dataset creators/users by revealing biases in *large-scale* visual data
2. Leverages the available annotations, pre-trained models, and census data to identify bias in the representation of *objects*, of *people* of different demographics, and of *geographic regions*
3. Suggests actionable insights to the user

**Example finding:**

Images: COCO dataset [Lin et al. ECCV’14]
Annotations: (1) binarized, socially-perceived inferred gender expression [Zhao et al. EMNLP’17], (2) predicted scenes with the Places model [Zhou et al. TPAMI’17]

Hot-off-the-press: Gender artifacts in visual datasets

Goal: Understand the extent to which gender artifacts are present in datasets

ROC AUC of a gender artifacts model (classifying if the image contains a person labeled “female” or “male”)

Images: COCO dataset [Lin et al. ECCV’14]
Gender labels: binarized, socially-perceived inferred gender expression [Zhao et al. EMNLP’17], [Zhao et al. ICCV’21]

Model trained on 3 features: (R,G,B) average image color

[Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, “Gender artifacts in visual datasets.”]
Can we adjust the AI design to **mitigate** these effects?

Human history, bias, prejudice

Large-scale data

AI decision making

AI models
Can we adjust the AI design to **mitigate** these effects?

Human history, bias, prejudice

Large-scale data
Need targeted interventions to (1) increase representation, (2) examine and understand the data, (3) constructively engage with the raised ethical issues
Can we adjust the AI design to **mitigate** these effects?

**Human history, bias, prejudice**

**Large-scale data**
Need targeted interventions to
(1) increase representation, (2) examine and understand the data, (3) constructively engage with the raised ethical issues

**AI decision making**

**AI models**
Can we adjust the AI design to mitigate these effects?

- Human history, bias, prejudice
- Large-scale data
  Need targeted interventions to (1) increase representation, (2) examine and understand the data, (3) constructively engage with the raised ethical issues
- AI decision making
- AI models
Challenges in building fair models

Understanding bias propagation

Quantifying fairness

Interpreting the model

Mitigating model bias

Being right for the right reason

Engaging with the social context

[Bolei Zhou et al. TPAMI’18 “Interpreting deep...”]
[R. Selvaraju et al. ICCV’17 “GradCAM: Visual...”]
[Sunnie S. Y. Kim et al. arxiv’21 “HIVE: Evaluating...”]
[Vikram V. Ramaswamy et al. arxiv’22 “ELUDE...”]

[Jieyu Zhao et al. EMNLP’17 “Men also like...”]
[P. Stock, M. Cisse. ECCV’18 “ConvNets...”]
[Tianlu Wang et al. ICCV’19 “Balanced...”]
[Dora Zhao et al. ICCV’21 “Understanding...”]

[Bolei Zhou et al. TPAMI’18 “Interpreting deep...”]
[R. Selvaraju et al. ICCV’17 “GradCAM: Visual...”]
[Sunnie S. Y. Kim et al. arxiv’21 “HIVE: Evaluating...”]
[Vikram V. Ramaswamy et al. arxiv’22 “ELUDE...”]

Moritz Hardt et al. NeurIPS’16 “Equality of...”
[M. Chen, M.Wu. UAI’20 “Towards threshold...”]
[A. Jacobs, H. Wallach. FAccT’21 “Measurement...”]
[A. Wang, O. Russakovsky. ICML’21 “Directional...”]

[R. Zemel et al. ICML’13 “Learning fair representations”]
[B. Zhang et al. AIES’18 “Mitigating unwanted biases...”]
[Zeyu Wang et al. CVPR’20 “Towards fairness in...”]
[Vikram V. Ramaswamy et al. CVPR’21 “Fair attribute...”]

50 Years of Test (Un)fairness: Lessons for Machine Learning
[
B. Hutchinson and Margaret Mitchell

[Michael Bernstein et al. arxiv’21 “ESR: Ethics...”]
[Angelina Wang et al. FAccT’22 “Towards inters...”]

[Remi Cadene et al. NeurIPS’19 “RUBi:...”]
[Krishna Singh et al. CVPR’20 “Don’t judge...”]
[Sunnie S. Y. Kim et al. ReScience’21 “[Re] Don’t...”]

[B. Hutchinson, M. Mitchell FAT*19 “50 years...”]
[Solon Barocas et al. “Fairness and ML” book]

A. Wang, O. Russakovsky. ICML’21 “Directional...”

Moritz Hardt et al. NeurIPS’16 “Equality of...”
[M. Chen, M.Wu. UAI’20 “Towards threshold...”]
[A. Jacobs, H. Wallach. FAccT’21 “Measurement...”]

[R. Zemel et al. ICML’13 “Learning fair representations”]
[B. Zhang et al. AIES’18 “Mitigating unwanted biases...”]
[Zeyu Wang et al. CVPR’20 “Towards fairness in...”]
[Vikram V. Ramaswamy et al. CVPR’21 “Fair attribute...”]

B. Hutchinson, M. Mitchell FAT*19 “50 years...”
[Solon Barocas et al. “Fairness and ML” book]