

# Fairness in visual recognition

Olga Russakovsky

<http://visualai.princeton.edu>



PRINCETON  
UNIVERSITY



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”

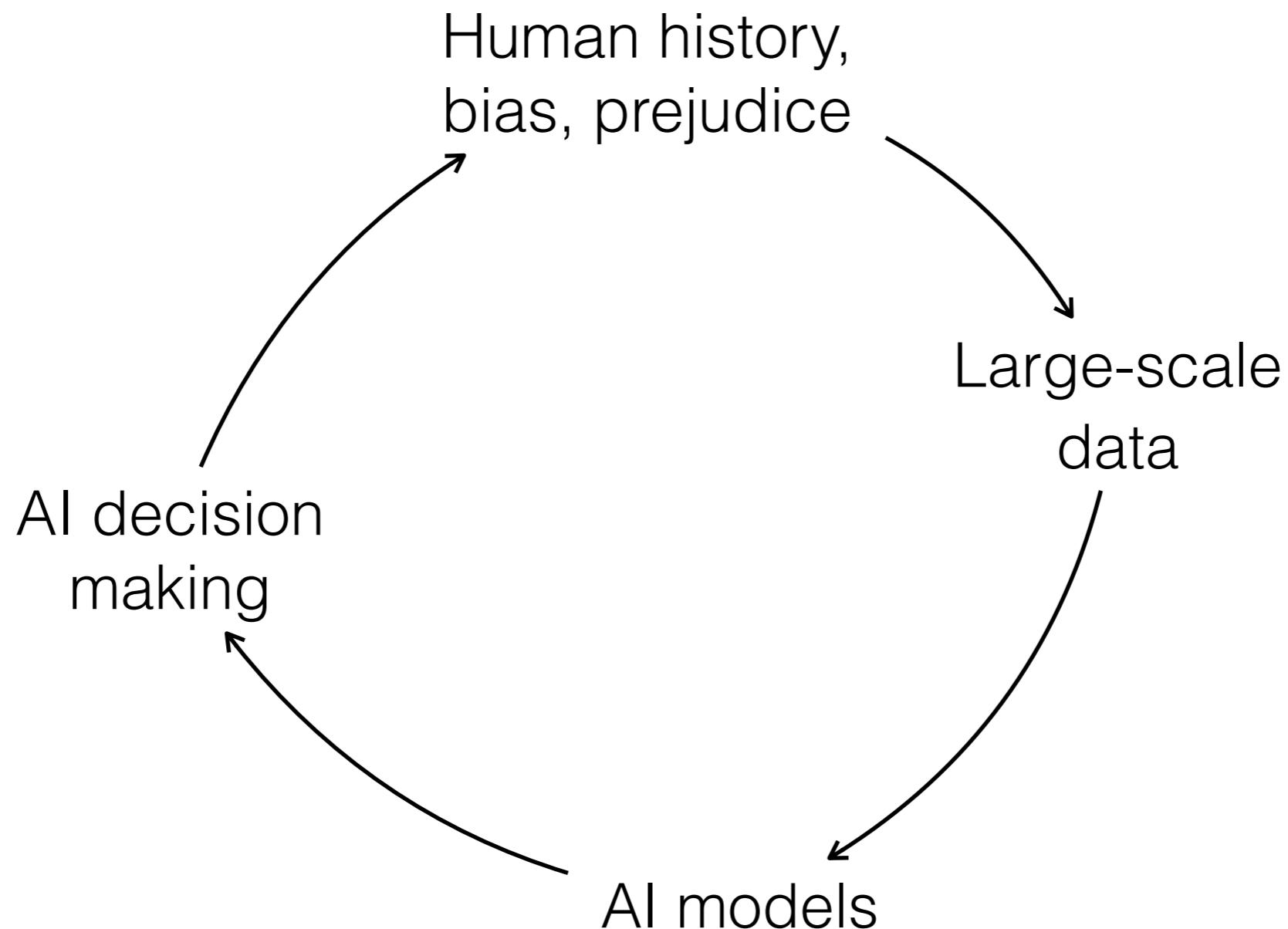
**The New York Times** Feb 9, 2018

Cropping algorithm prefers  
“lighter, slimmer, younger faces”

**The Guardian** Aug 10, 2021

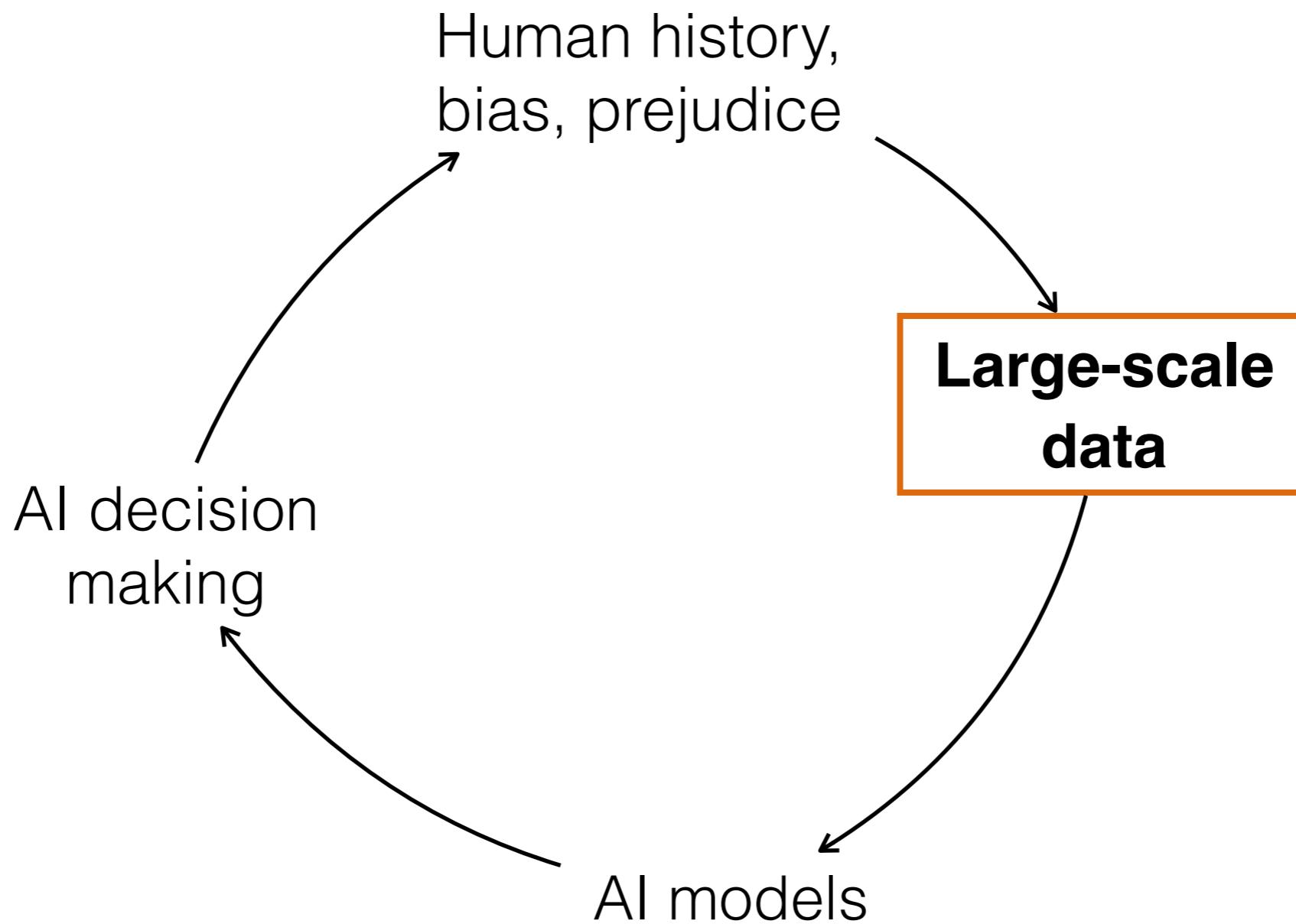
See: Ruha Benjamin’s “Race after technology” — an excellent book

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



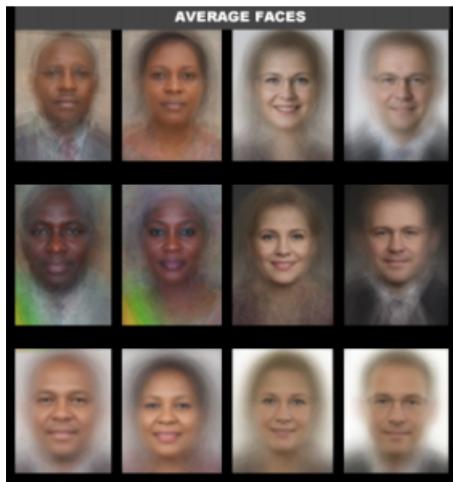
See: Ruha Benjamin's “Race after technology” — an excellent book

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



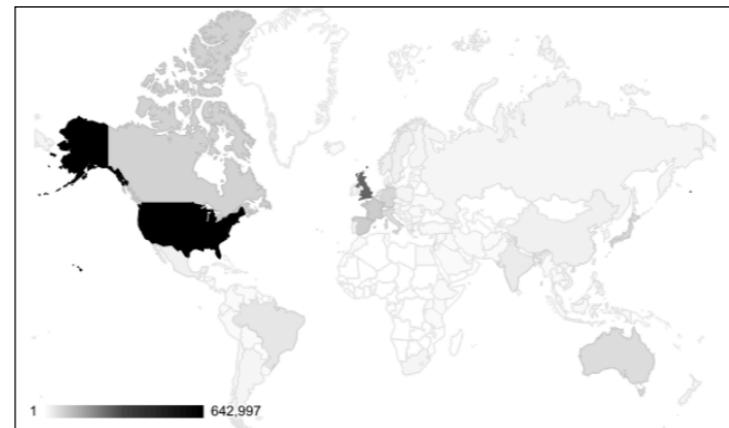
# Large scale $\neq$ fair representation

## Demographic diversity



[Joy Buolamwini and Timnit Gebru. FAT\*18  
“GenderShades: Intersectional accuracy...”]

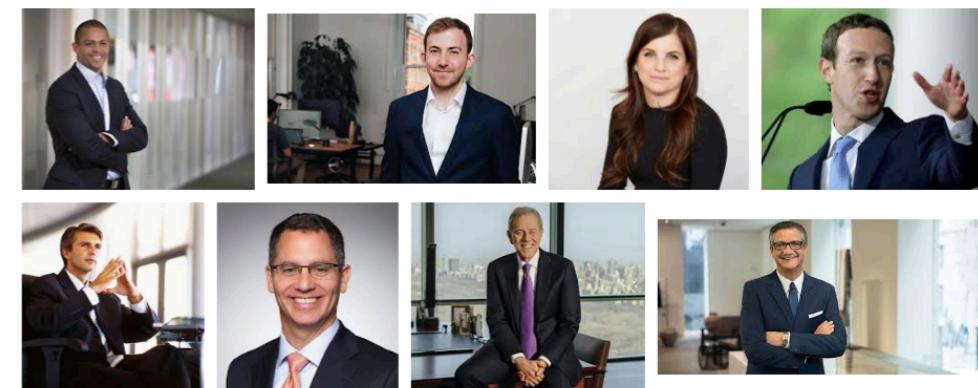
## Geographic diversity (in ImageNet and OpenImages)



[Shreya Shankar et al. NeurIPS Workshop’17  
“No classification without representation...”]  
[Terrance DeVries et al. CVPR Workshop’19  
“Does object recognition work for everyone”]

## Image search diversity

CEO



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

## 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...”]
- [Emily Denton et al. Big Data & Society’21 “On the Genealogy of Machine...”]
- [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**”]
- [Yuki Asano et al. NeurIPS D&B track’21 “PASS: An ImageNet replacement...”]

# REVISE: REvealing VIvisual biaSEs tool

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

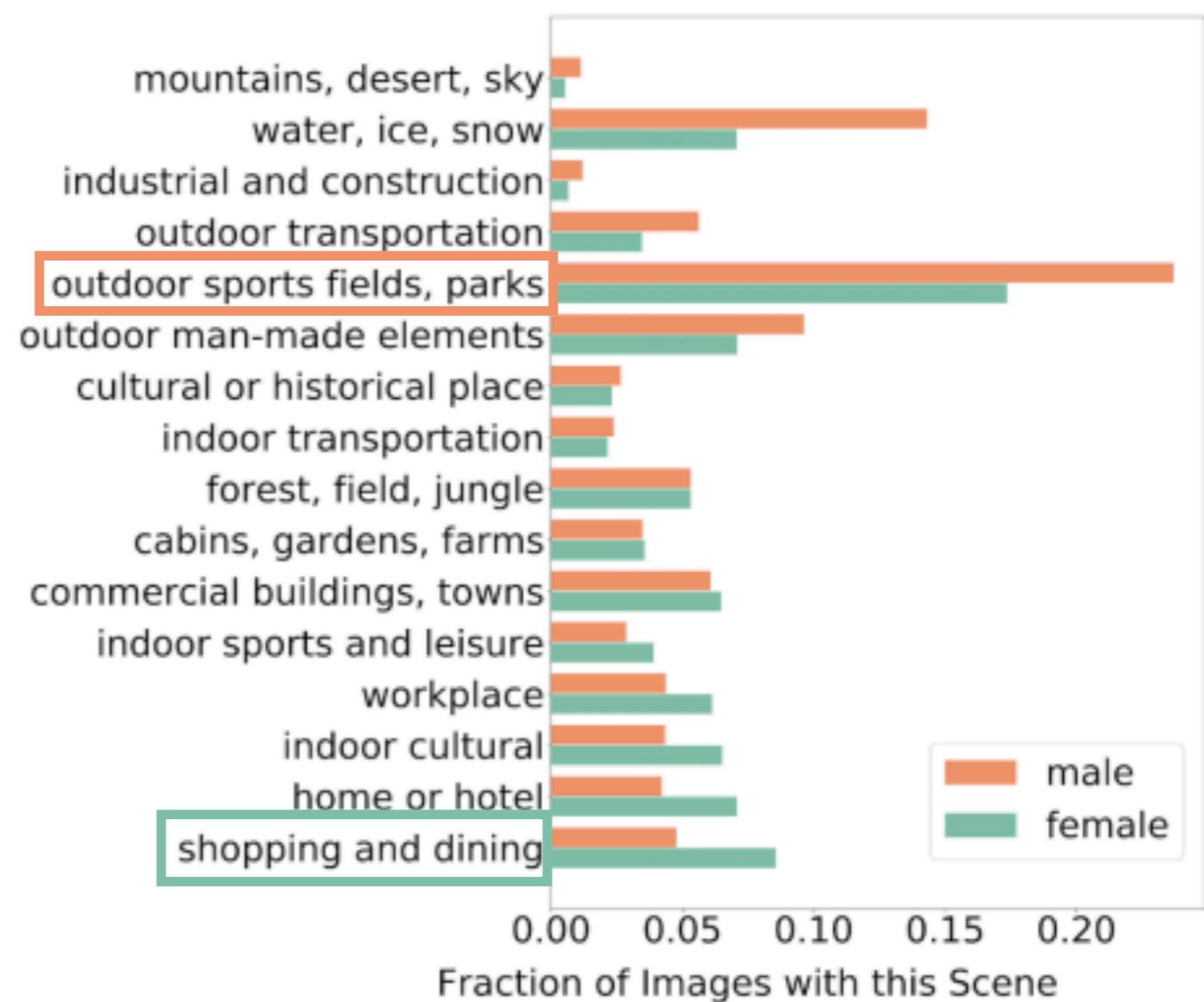
## Contributions:

- (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'7]



[Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, Olga Russakovsky. IJCV'22  
“REVISE: a tool for measuring and mitigating bias in visual datasets.” <https://github.com/princetonvisualai/revise-tool>]

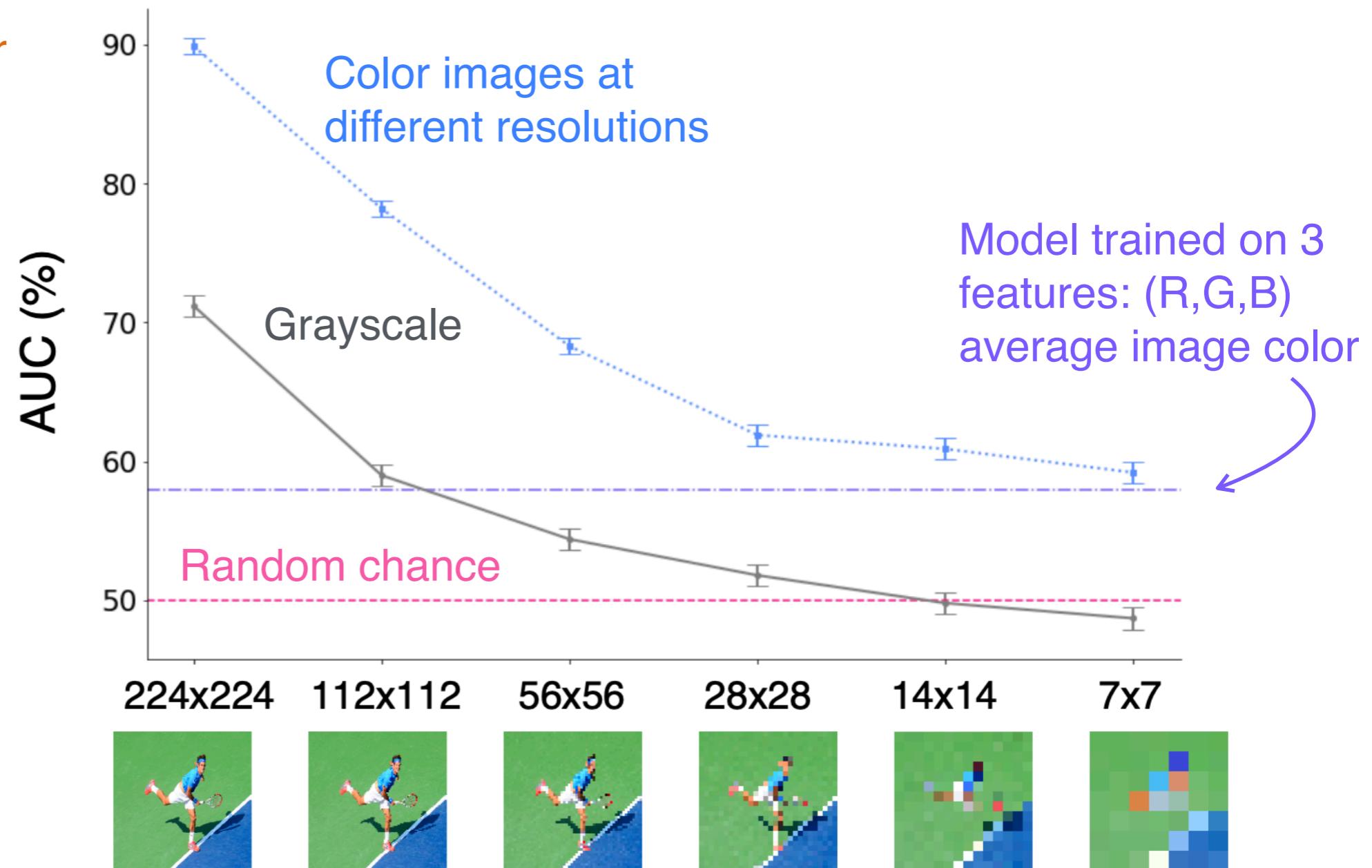
# Hot-off-the-press: Gender artifacts in visual datasets

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”)

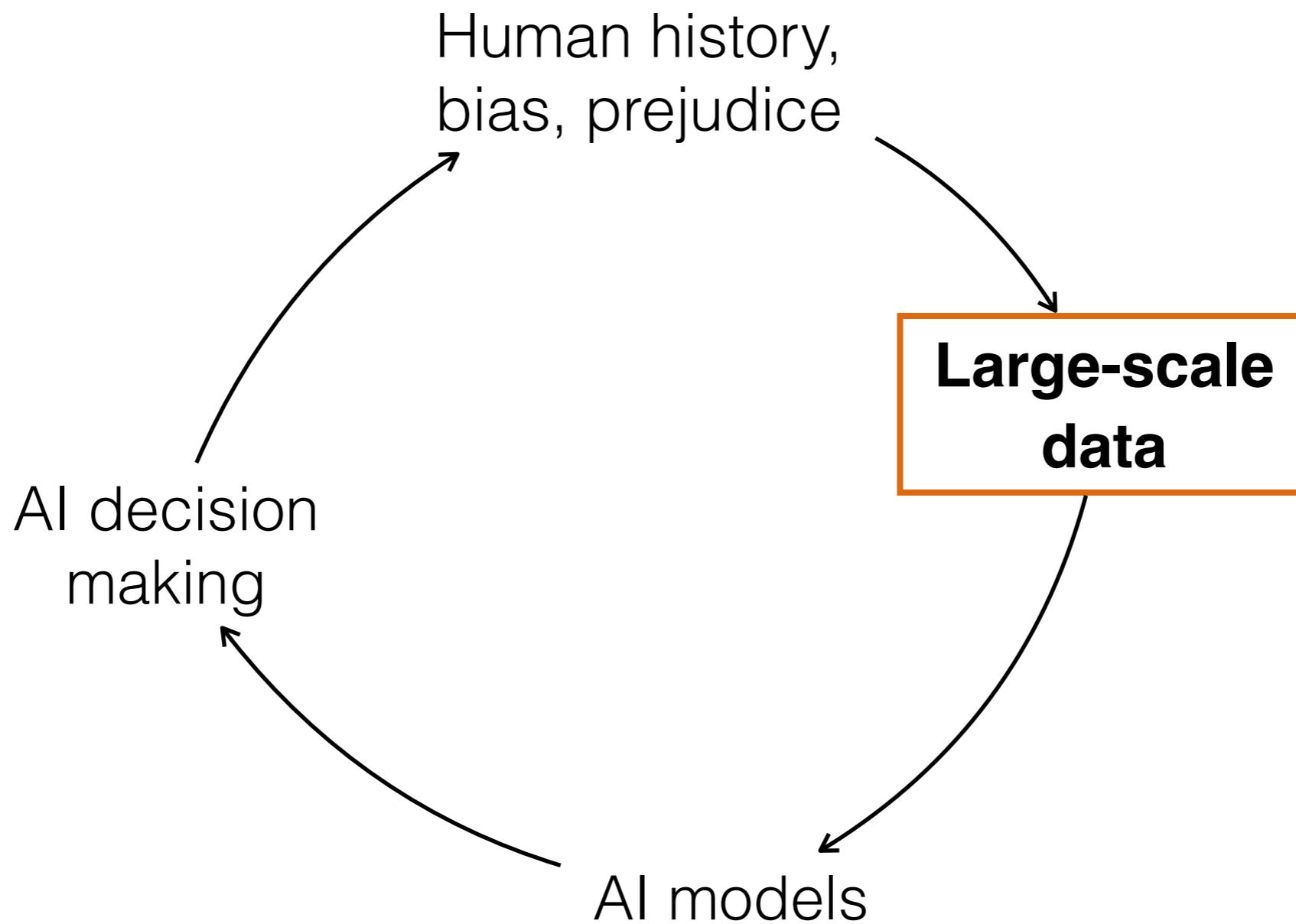
Images: COCO dataset [Lin et al. ECCV’14]

Gender labels: binarized, socially-perceived inferred gender expression [Zhao et al. EMNLP’17], [Zhao et al. ICCV’21]

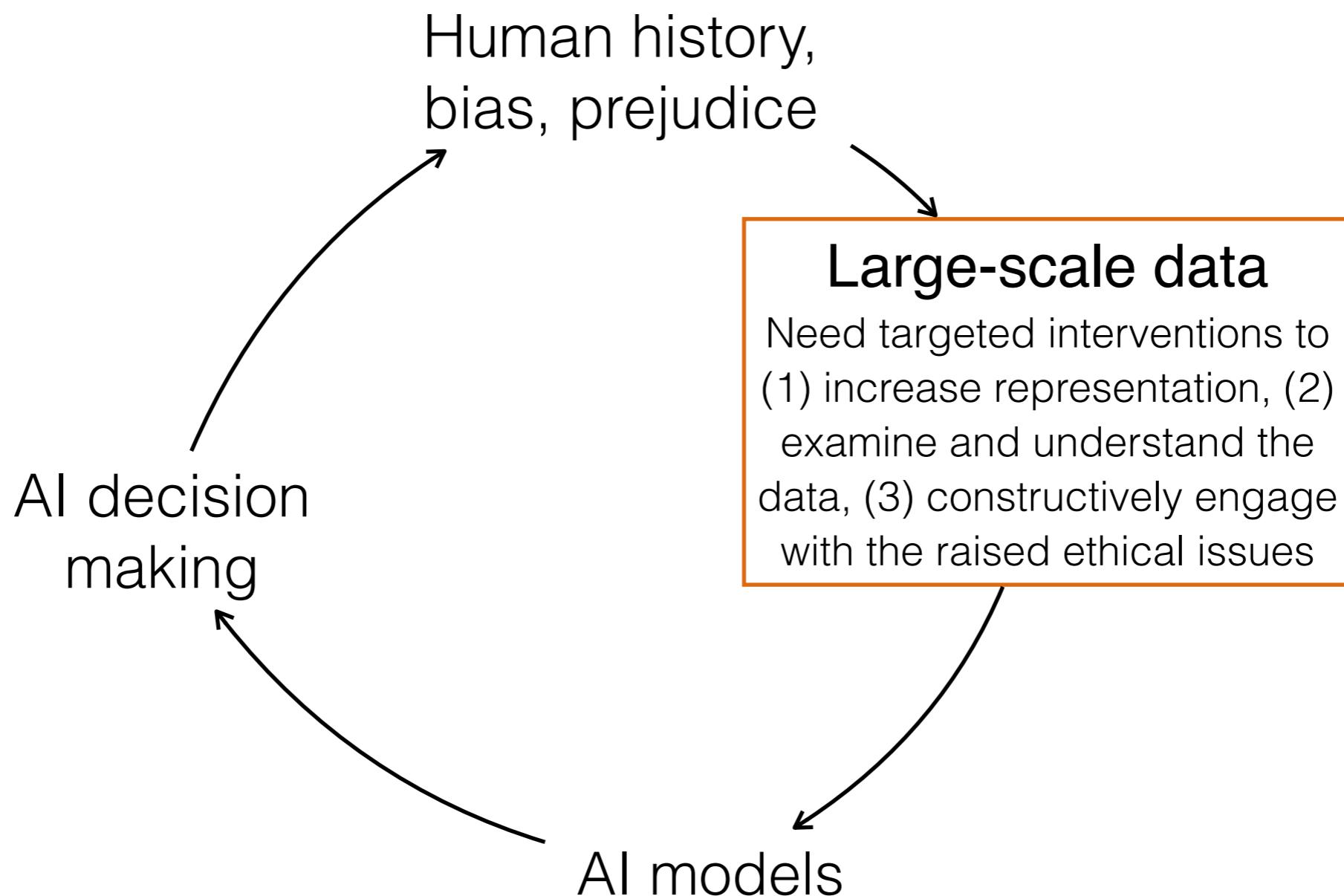


[Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky. “Gender artifacts in visual datasets.”]

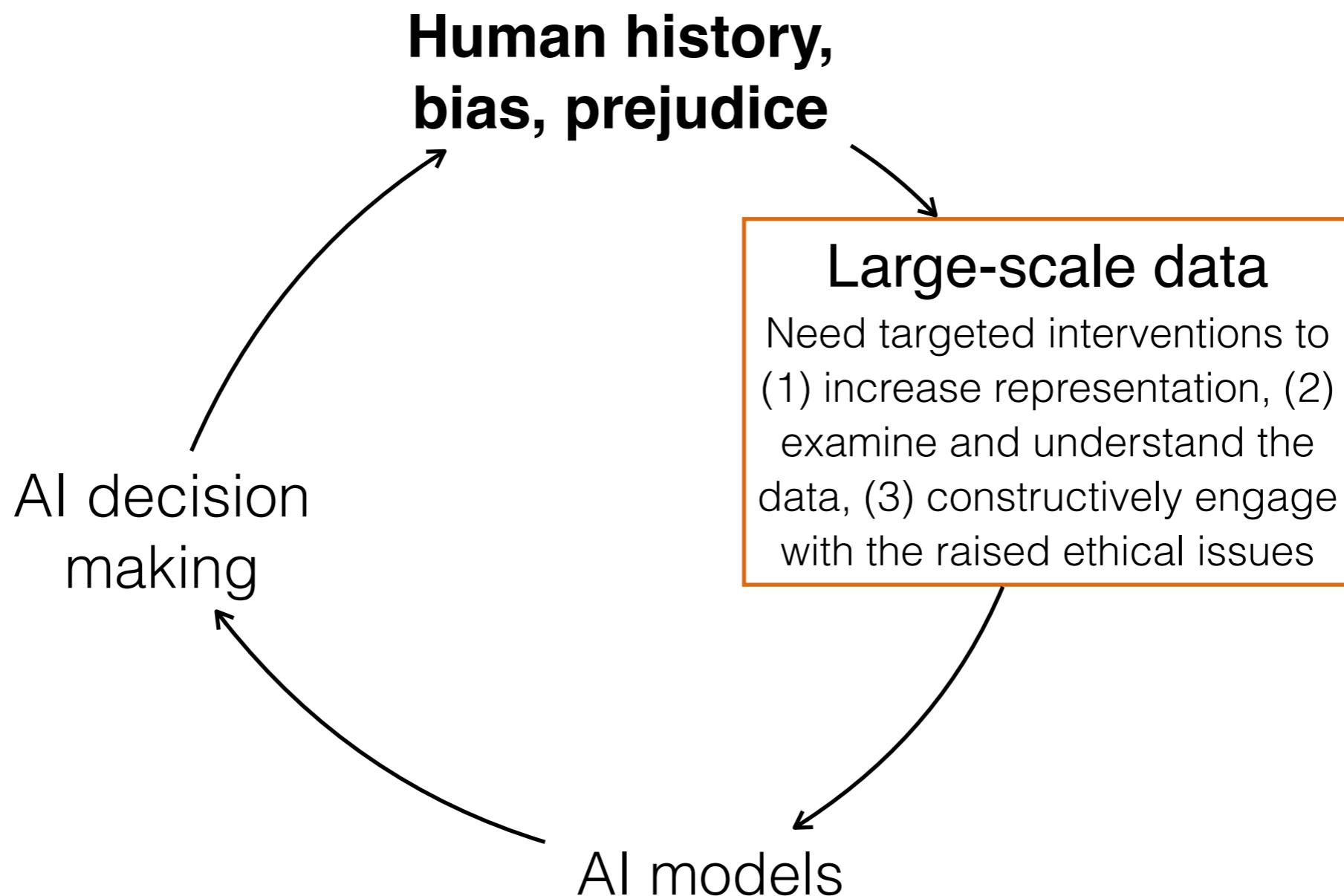
Can we adjust the AI design to **mitigate** these effects?



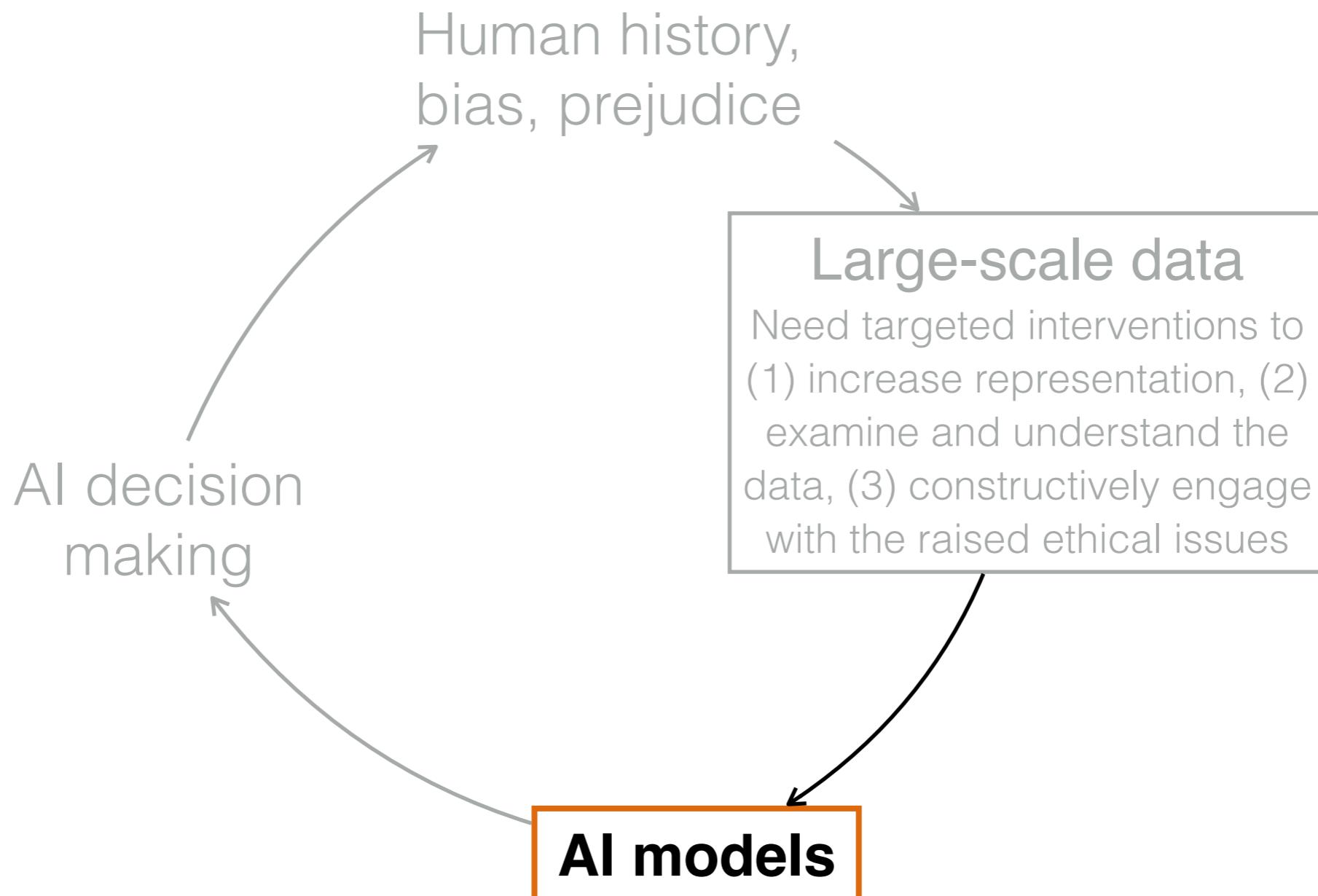
# Can we adjust the AI design to **mitigate** these effects?



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

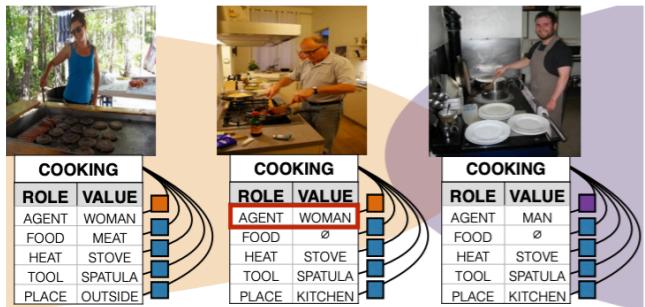


# Can we adjust the AI design to mitigate these effects?



# Challenges in building fair models

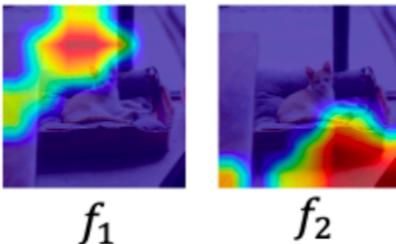
## Understanding bias propagation



- [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..."]

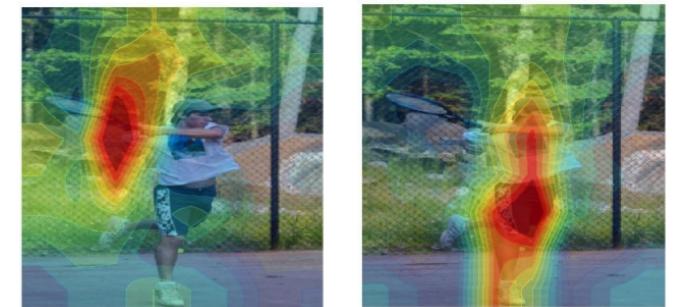
## Interpreting the model

$$\begin{aligned} \text{fur: } & +1.2 \times (\textcolor{green}{1}) \\ \text{paw: } & +0.7 \times (\textcolor{green}{1}) \\ \text{tree: } & -0.6 \times (\textcolor{red}{0}) \\ & = 1.9 \end{aligned}$$



- [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..."]

## Being right for the right reason



- [Kaylee Burns et al. ECCV'18 "Women also..."]  
 [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..."]

## Quantifying fairness

**Definition 2.1** (Equalized odds). We say that a predictor  $\hat{Y}$  satisfies *equalized odds* with respect to protected attribute  $A$  and outcome  $Y$ , if  $\hat{Y}$  and  $A$  are independent conditional on  $Y$ .

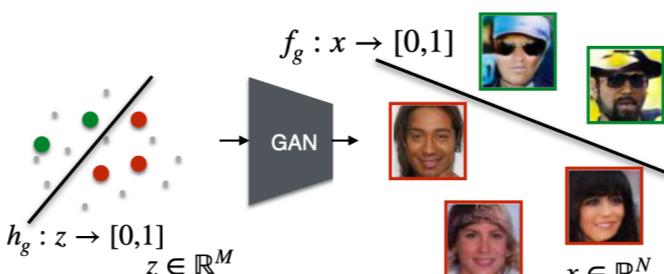
Unlike demographic parity, equalized odds allows  $\hat{Y}$  to depend on  $A$  but only through the target variable  $Y$ . This encourages the use of features that relate to  $Y$  directly, not through  $A$ .

As stated, equalized odds applies to targets and protected attributes taking values in any space, including discrete and continuous spaces. But in much of our presentation we focus on binary targets  $Y, \hat{Y}$  and protected attributes  $A$ , in which case equalized odds is equivalent to:

$$\Pr\{\hat{Y} = 1 \mid A = 0, Y = y\} = \Pr\{\hat{Y} = 1 \mid A = 1, Y = y\}, \quad y \in \{0, 1\} \quad (2.1)$$

- [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..."]

## Mitigating model bias



- [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..."]

## Engaging with the social context

### 50 Years of Test (Un)fairness: Lessons for Machine Learning

Ben Hutchinson and Margaret Mitchell  
 {benhutch, mmitchellai}@google.com

#### STRACT

Initiative definitions of what is *unfair* and what is *fair* have been induced in multiple disciplines for well over 50 years, including education, hiring, and machine learning. We trace how the notion of fairness has been defined within the testing communities of education and hiring over the past half century, exploring the cultural and social contexts in which different fairness definitions emerged. In some cases, earlier definitions of fairness are similar to definitions of fairness in current machine learning research, and foreshadow current formal work. In other cases, it is in the evolution of what fairness means and how to measure it have largely

the educational and employment testing communities, often with focus on race. The period of time from 1966 to 1976 in particular gave rise to fairness research with striking parallels to ML fairness research from 2011 until today, including formal notions of fairness based on population subgroups, the realization that some fairness criteria are incompatible with one another, and pushback on quantitative definitions of fairness due to their limitations.

Into the 1970s, there was a shift in perspective, with researchers moving from defining how a test may be *unfair* to how a test may be *fair*. It is during this time that we see the introduction of mathematical criteria for fairness identical to the mathematical criteria

- [B. Hutchinson, M. Mitchell FAT\*19 "50 years..."]  
 [Solon Barocas et al. "Fairness and ML" book]  
 [Michael Bernstein et al. arxiv'21 "ESR: Ethics..."]  
 [Angelina Wang et al. FAccT'22 "Towards inters..."]

