

Accountable, Responsible and Transparent AI: How to Get the Algorithms Right

Özgür Şimşek
University of Bath

Amsterdam, October 3, 2019
17th Gender Summit – Europe



Amazon scraps secret AI recruiting tool that showed bias against women.

“Everyone wanted this holy grail. They literally wanted it to be an engine where I’m going to give you 100 resumes, it will spit out the top five, and we’ll hire those.”

The system was not gender neutral: it penalized resumes that included the word “women’s”.

AI system was trained on resumes submitted over 10 years.



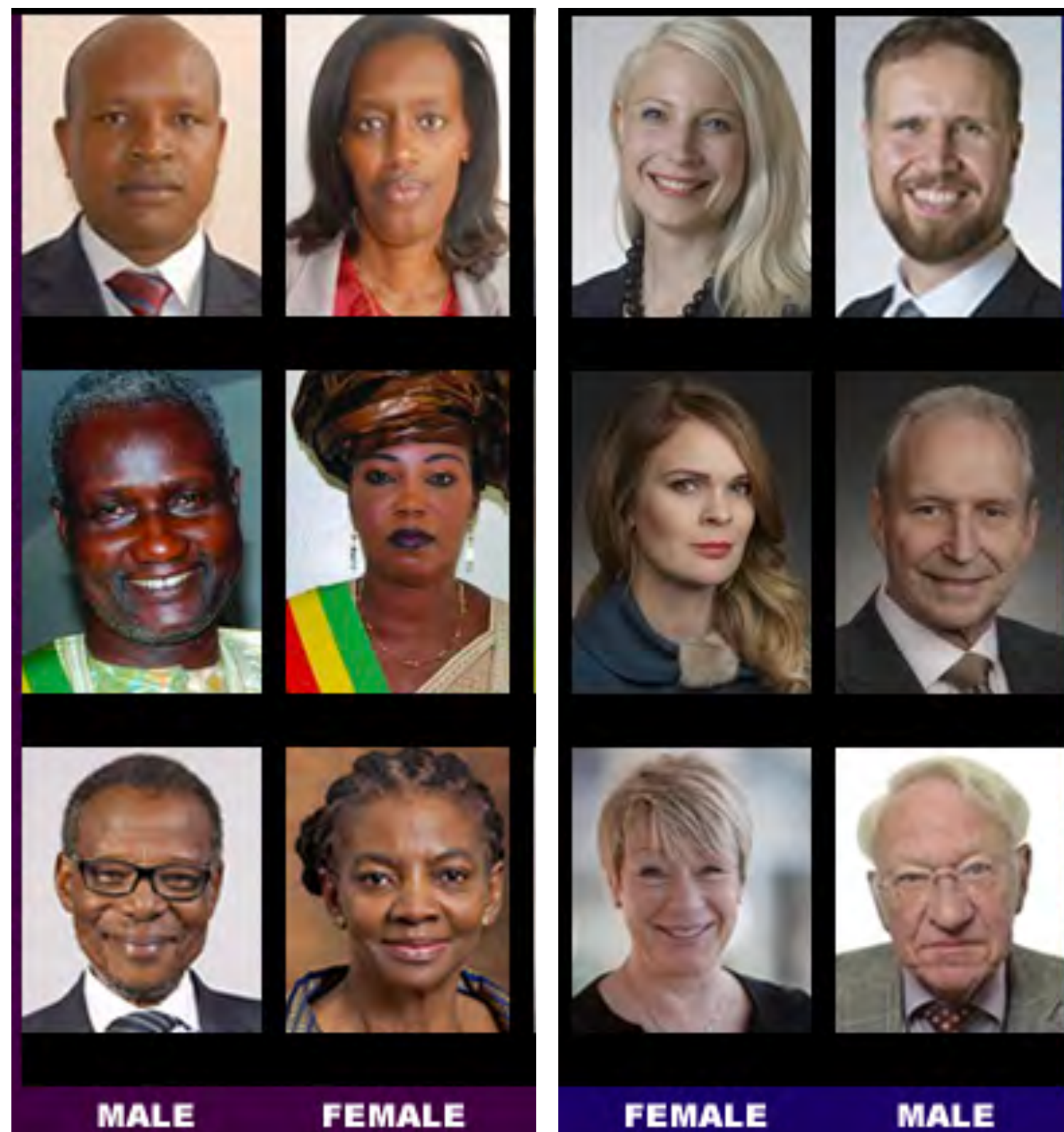
Note: Amazon does not disclose the gender breakdown of its technical workforce.

Source: Latest data available from the companies, since 2017.

By Han Huang | REUTERS GRAPHICS

Reuters, Oct 10, 2018

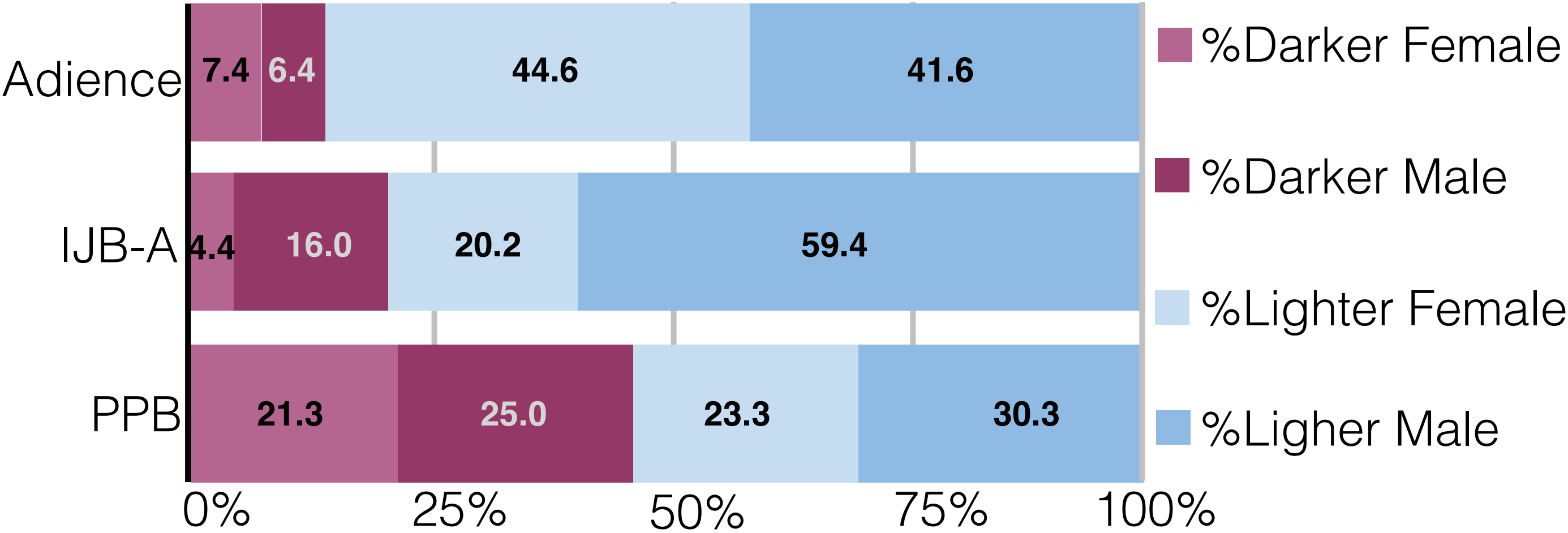
Facial recognition performs better at detecting men than women— and at detecting light-skinned people than dark-skinned people.



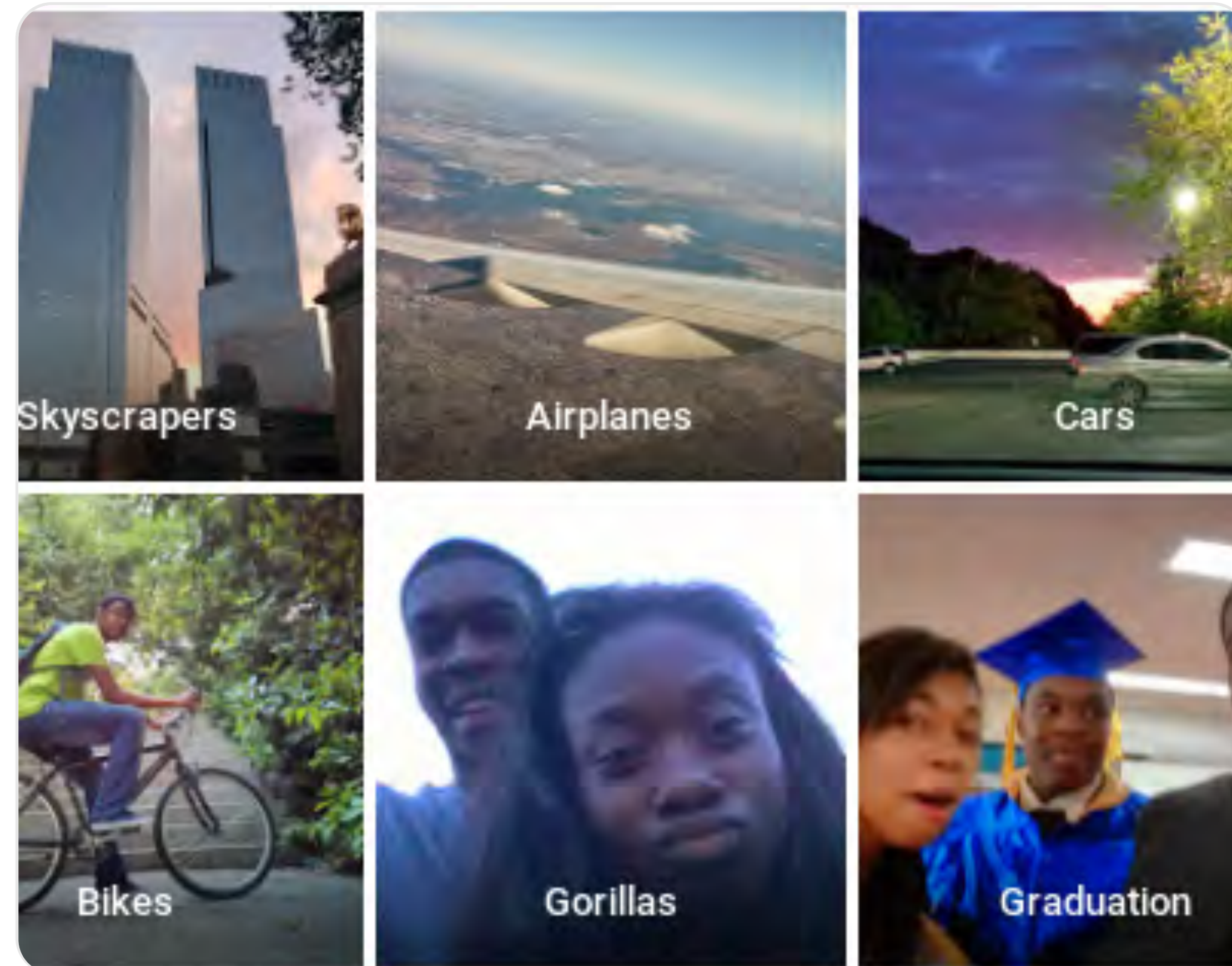
Buolamwini & Gebru (PMLR 2018) tested 3 general-purpose commercial facial-analysis systems.

- All performed better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All performed better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All performed worst on darker female faces (20.8% – 34.7% error rate)

Facial recognition performs better at detecting men than women— and at detecting light-skinned people than dark-skinned people.



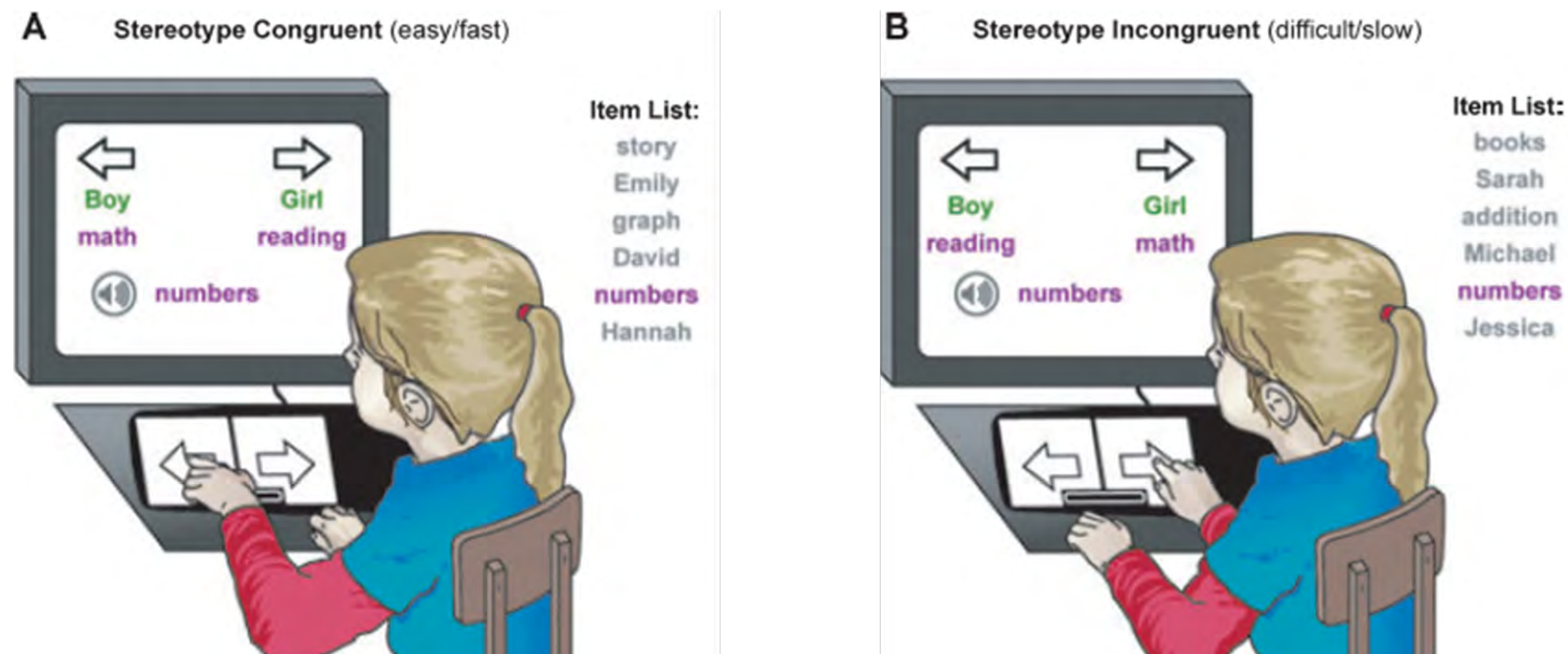
Buolamwini & Gebru, PMLR 2018



2015 tweet by Jacky Alciné, user of Google photos

Wired (01.11.2018) reports:
Google censored “gorilla” from searches and image tags.
“Chimp,” “chimpanzee,” and “monkey” were similarly blocked.

Semantics derived automatically from language corpora contain human-like biases.



Examples of findings:

- Flowers are more pleasant than insects.
- Musical instruments are more pleasant than weapons.
- Men are associated with work, math, and science, and women with family and the arts.

Implicit Association Task: Measures unconscious bias. Concepts that are associated are easier to pair and reaction times are comparatively faster (Greenwald, McGhee, & Schwartz, 1998)

Caliskan, Bryson & Narayanan (Science, 2017) analyzed 840 billion words—from tweets, the US Declaration of Independence, Reddit threads, and many other sources.

They found the same biases using word embeddings alone.

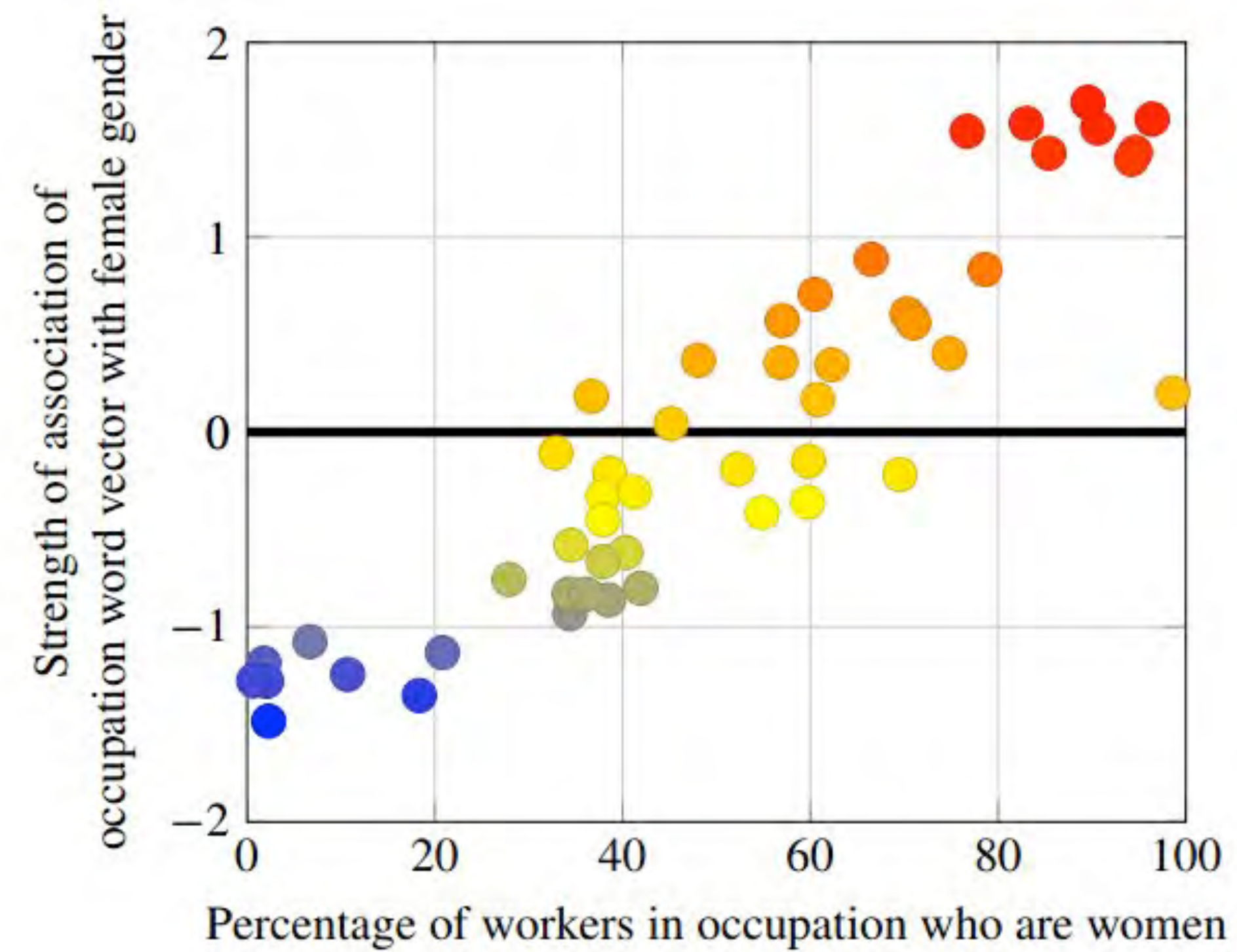


Figure 1. Occupation-gender association
 Pearson's correlation coefficient $\rho = 0.90$ with $p\text{-value} < 10^{-18}$.

Caliskan, Bryson & Narayanan (Science, 2017)

Human culture as expressed in text is sufficient signal for AI to learn regularities about human society and experience.

AI trained by machine learning on culture will inherit our biases.

Caliskan, Bryson & Narayanan (Science, 2017)

“AI is a feminist issue.”

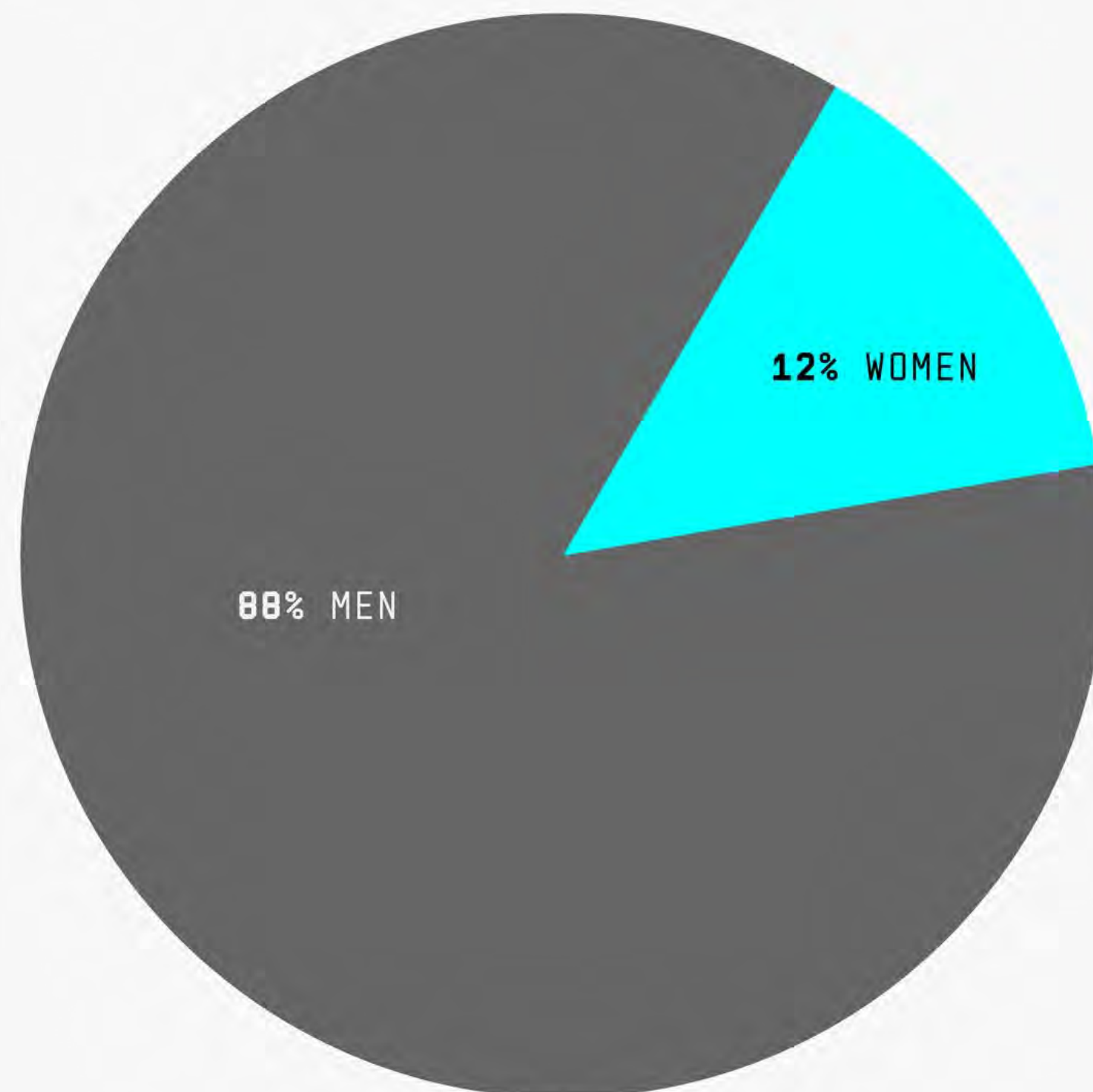
“Progress in fairness is not a simple matter of neutrality that falls out of machine intelligence. Progress must be achieved through explicit effort.”

—Joanna Bryson, 2017

“Like all technologies before it, artificial intelligence will reflect the values of its creators. So inclusivity matters — from who designs it to who sits on the company boards and which ethical perspectives are included.” — Kate Crawford, 2016

“Over-representation of men in the design of these technologies could quietly undo decades of advances in gender equality” —Susan Leavy, 2018

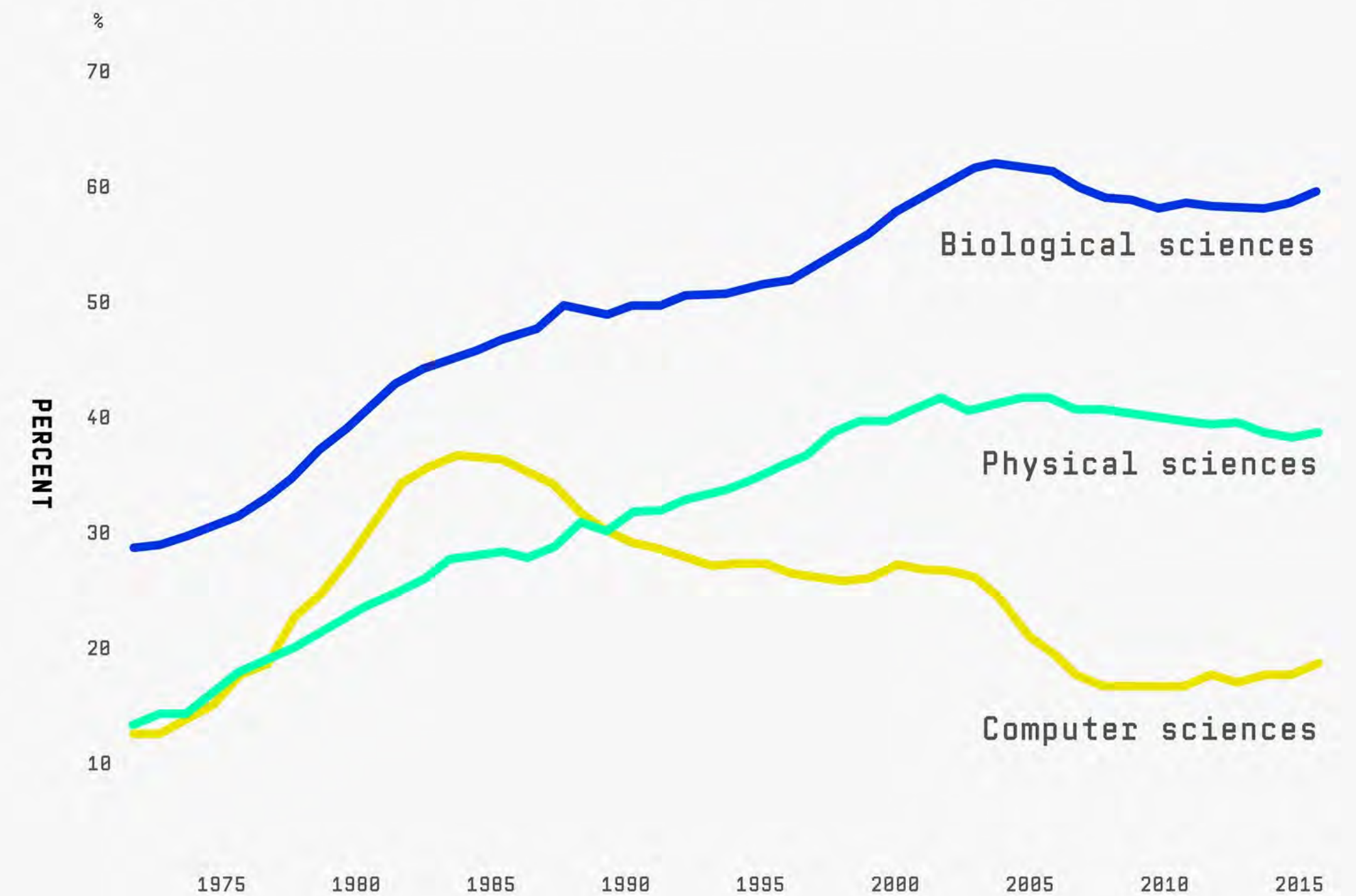
WOMEN ARE A SMALL MINORITY IN MACHINE-LEARNING RESEARCH



Percent of men and women who contributed work to three leading machine learning conferences in 2017.

Source: Element AI

WOMEN EARN A SMALLER SHARE OF COMPUTING DEGREES THAN 30 YEARS AGO



Share of bachelor's degrees earned by women in the US.

Source: NCES

Wired 08.17.2018

What we are doing at Bath:

UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent Artificial Intelligence (ART-AI)

Research Topics:
Algorithms, ethics,
public policy,
transparency,
safety and trust,
policy-making with
and for AI, ...

