

ADVERTISEMENT



Menu

Weekly edition

Search

My account

### Graphic detail

Jun 5th 2021 edition

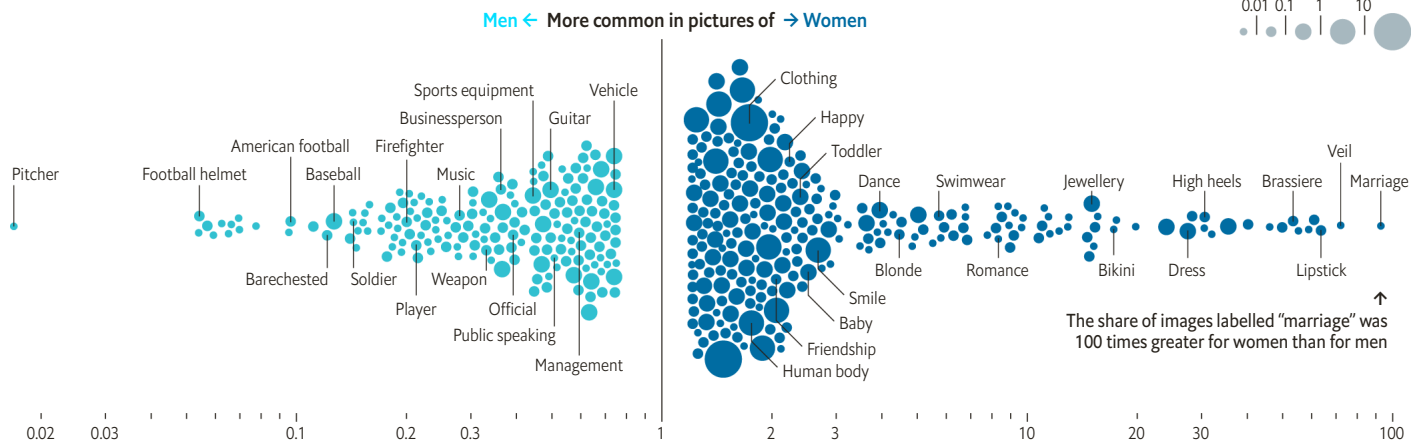
Bias in, bias out

# Demographic skews in training data create algorithmic errors

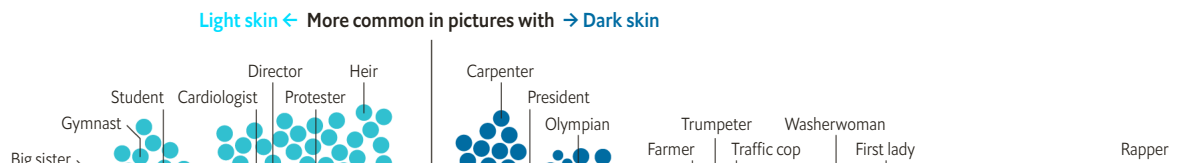
## Women and people of colour are underrepresented and depicted with stereotypes

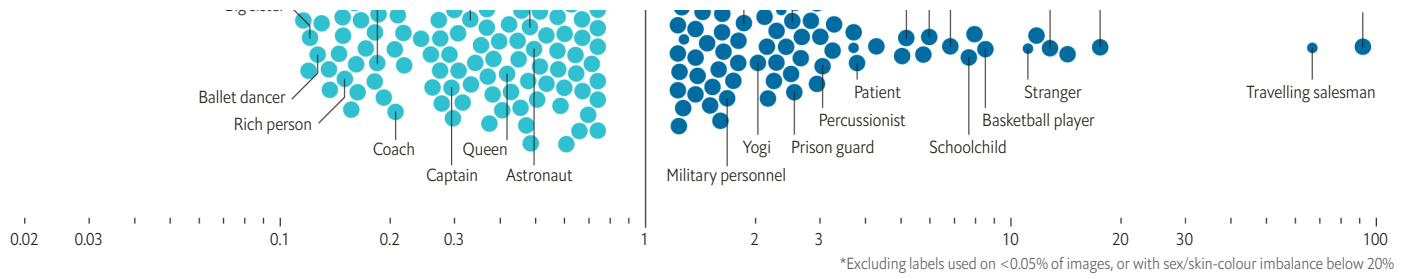
JUN 5TH 2021

Google Open Images, ratio of label frequency\* in images of women v men, log scale



ImageNet, ratio of label frequency\* in images of dark-skinned v light-skinned people, log scale

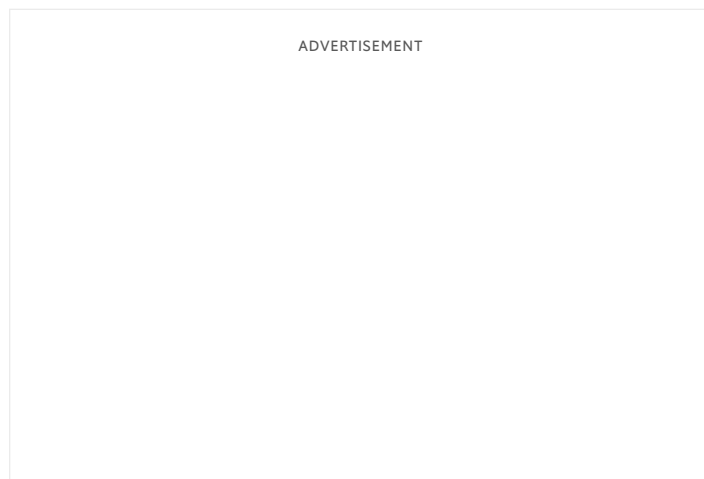




**A**LGORITHMIC BIAS is often described as a thorny technical problem. Machine-learning models can respond to almost any pattern—including ones that reflect discrimination. Their designers can explicitly prevent such tools from consuming certain types of information, such as race or sex. Nonetheless, the use of related variables, like someone’s address, can still cause models to perpetuate disadvantage.

Ironing out all traces of bias is a daunting task. Yet despite the growing attention paid to this problem, some of the lowest-hanging fruit remains unpicked.

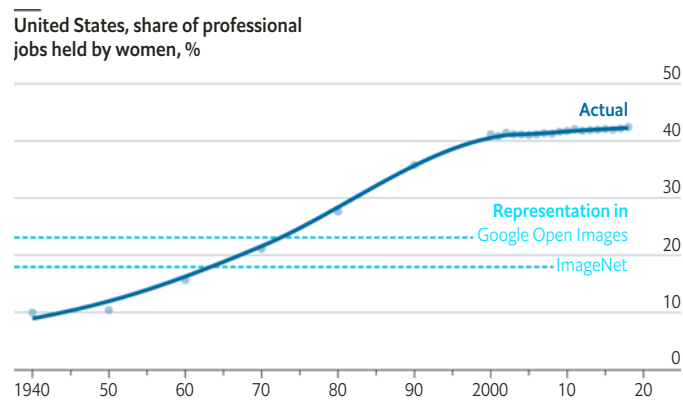
Every good model relies on training data that reflect what it seeks to predict. This can sometimes be a full population, such as everyone convicted of a given crime. But modellers often have to settle for non-random samples. For uses like facial recognition, models need enough cases from each demographic group to learn how to identify members accurately. And when making forecasts, like trying to predict successful hires from recorded job interviews, the proportions of each group in training data should resemble those in the population.



Many businesses compile private training data. However, the two largest public image archives, Google Open Images and ImageNet—which together have 725,000 pictures labelled by

sex, and 27,000 that also record skin colour—are far from representative. In these collections, drawn from search engines and image-hosting sites, just 30-40% of photos are of women. Only 5% of skin colours are listed as “dark”.

Sex and race also sharply affect how people are depicted. Men are unusually likely to appear as skilled workers, whereas images of women disproportionately contain swimwear or undergarments. Machine-learning models regurgitate such patterns. One study trained an image-generation algorithm on ImageNet, and found that it completed pictures of young women’s faces with low-cut tops or bikinis.



Similarly, images with light skin often displayed professionals, such as cardiologists. Those with dark skin had higher shares of rappers, lower-class jobs like “washerwoman” and even generic “strangers”. Thanks to the Obamas, “president” and “first lady” were also overrepresented.

ImageNet is developing a tool to rebalance the demography of its photos. And private firms may use less biased archives. However, commercial products do show signs of skewed data. One study of three programs that identify sex in photos found far more errors for dark-skinned women than for light-skinned men.


Making image or video data more representative would not fix imbalances that reflect real-world gaps, such as the high number of dark-skinned basketball players. But for people trying to clear passport control, avoid police stops based on security cameras or break into industries run by white men, correcting exaggerated demographic disparities would surely help. ■

Sources: ImageNet; Google Open Images; IPUMS

*This article appeared in the Graphic detail section of the print edition under the headline “Bias in, bias out”*

Reuse this content

The Trust Project

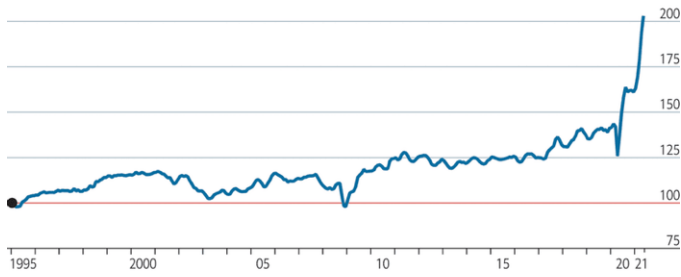


**OFF THE CHARTS**  
**Taking you behind the scenes of our data journalism**  
 Directly to your inbox every Tuesday

Sign up

More from Graphic detail

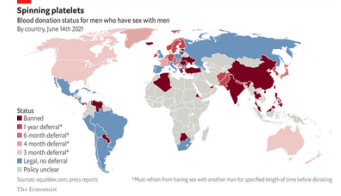
**Over-inflated tyres**  
Manheim used-vehicle value index  
Wholesale prices, January 1995=100



Source: Cox Automotive  
The Economist

**Daily chart**  
**Why the used-car market is sending American inflation soaring**

**Daily chart**  
**Which countries prohibit gay or bisexual men from donating blood?**



**House of unrepresentatives**  
**How the Voting Rights Act limits gerrymanders**



Subscribe

Keep updated

Group subscriptions



Reuse our content



Help and contact us

Published since September 1843 to take part in  
*“a severe contest between intelligence, which presses forward, and an unworthy, timid ignorance obstructing our progress.”*

**The Economist**

**The Economist Group**

[About](#)

[The Economist Group](#)

[Advertise](#)

[The Economist Intelligence Unit](#)

[Press centre](#)

[The Economist Events](#)

[The Economist Store](#)

[Careers](#)

[Which MBA?](#)

[GMAT Tutor](#)

[GRE Tutor](#)

[Executive Jobs](#)

[Executive Education Navigator](#)

[Executive Education](#)