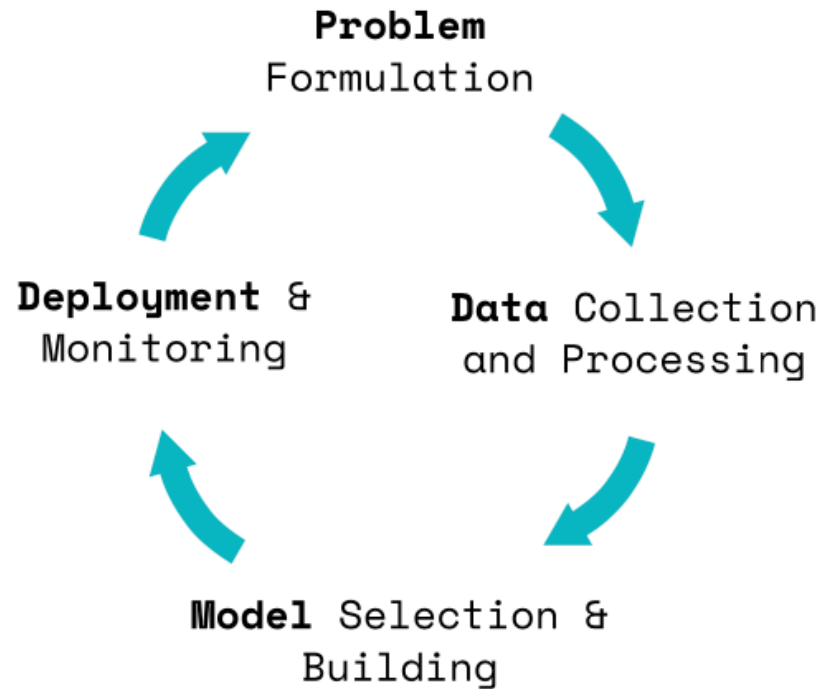


2. What factors contribute to bias and systemic inequities in artificial intelligence?

The AI Development Lifecycle

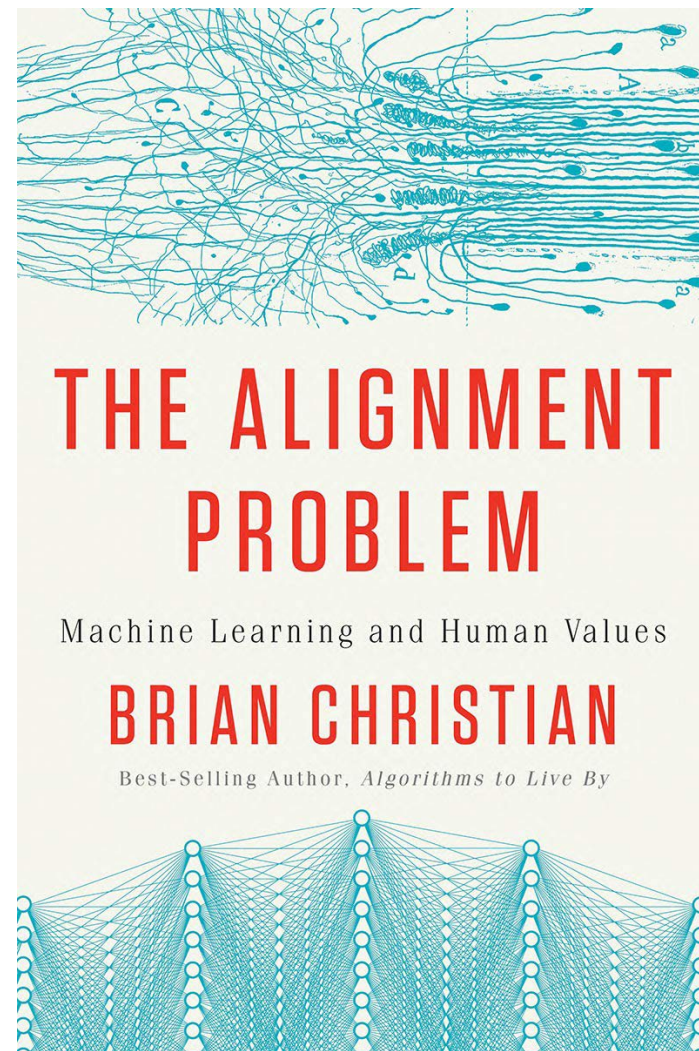
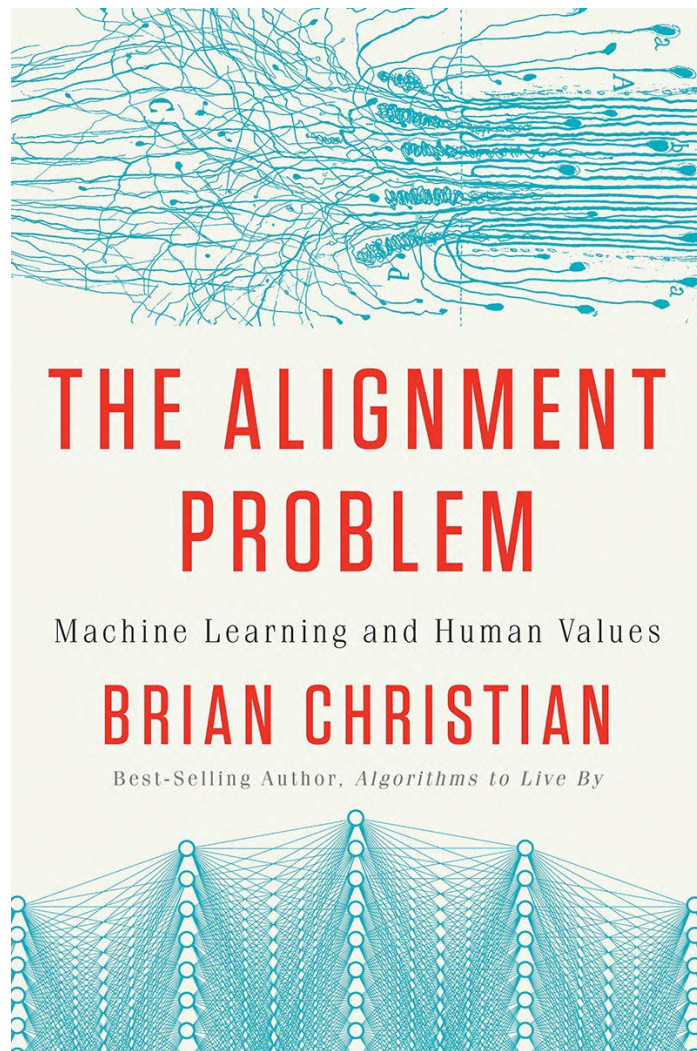
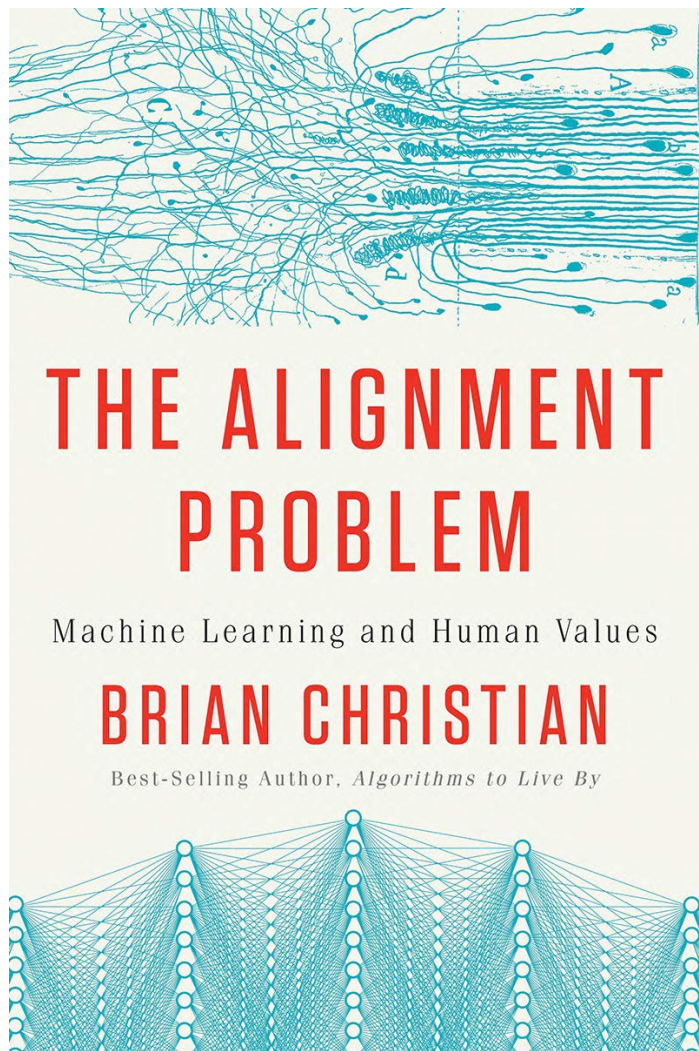


Problem: What/Whose problem is AI solving?

Data: Is the data used to train AI models bias?

Model: What tasks do optimize the model for?

Deployment: Is the model monitored for disparate impacts?





Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

COMPAS

Training data based on historical data

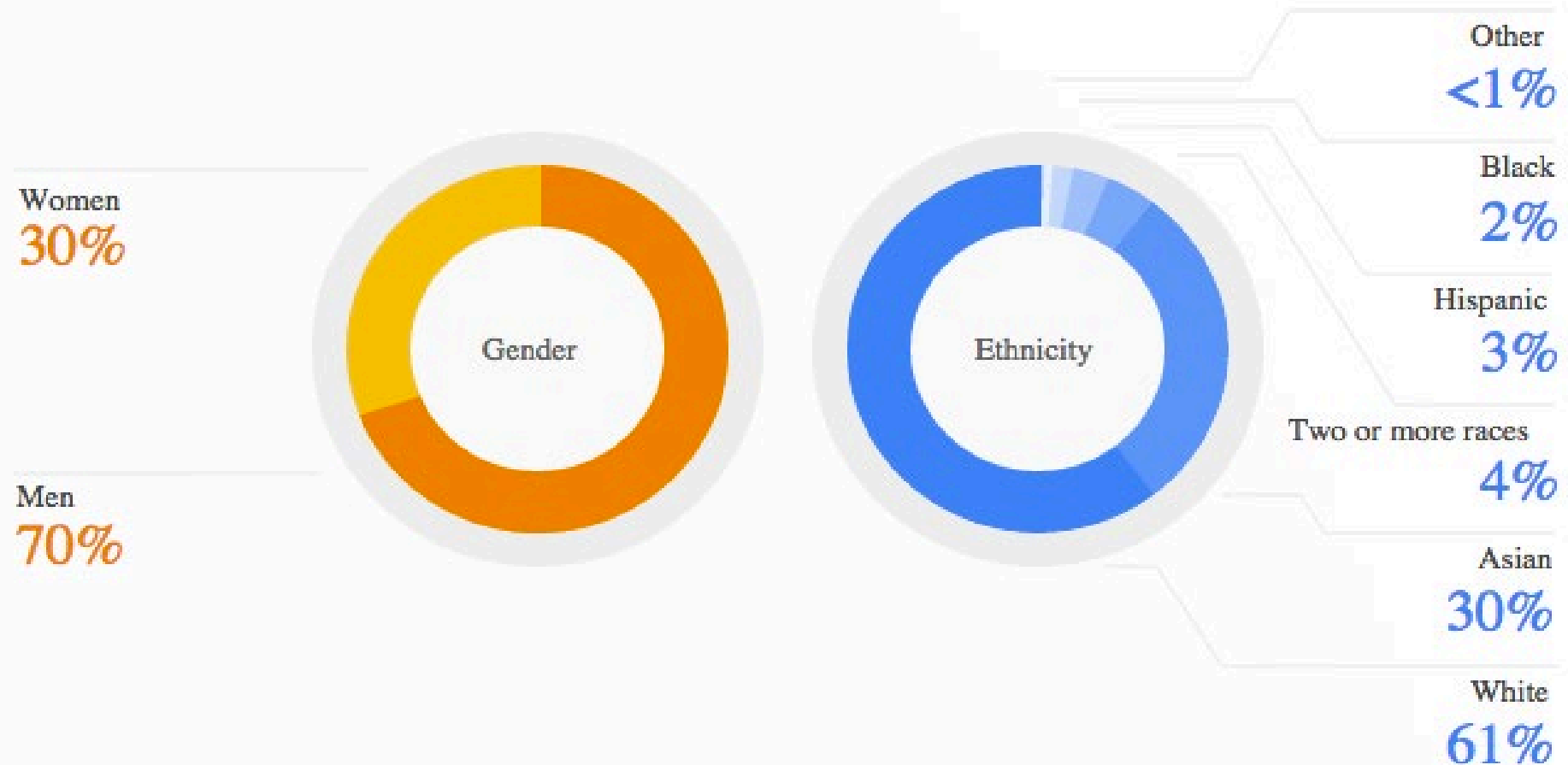
Variable Selection (Zip Code)

Arrest=Crime: Measurement Bias

Social networks

Amazon's Hiring Algorithm

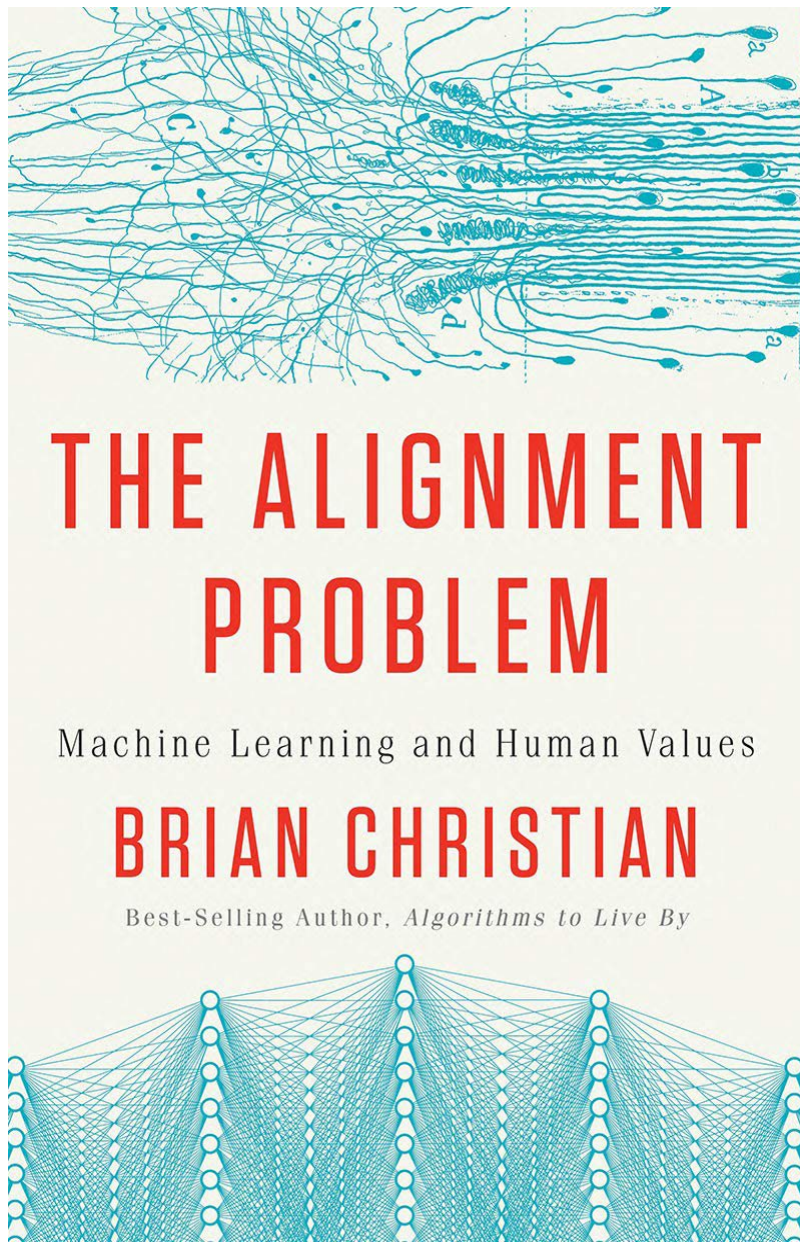
In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges.



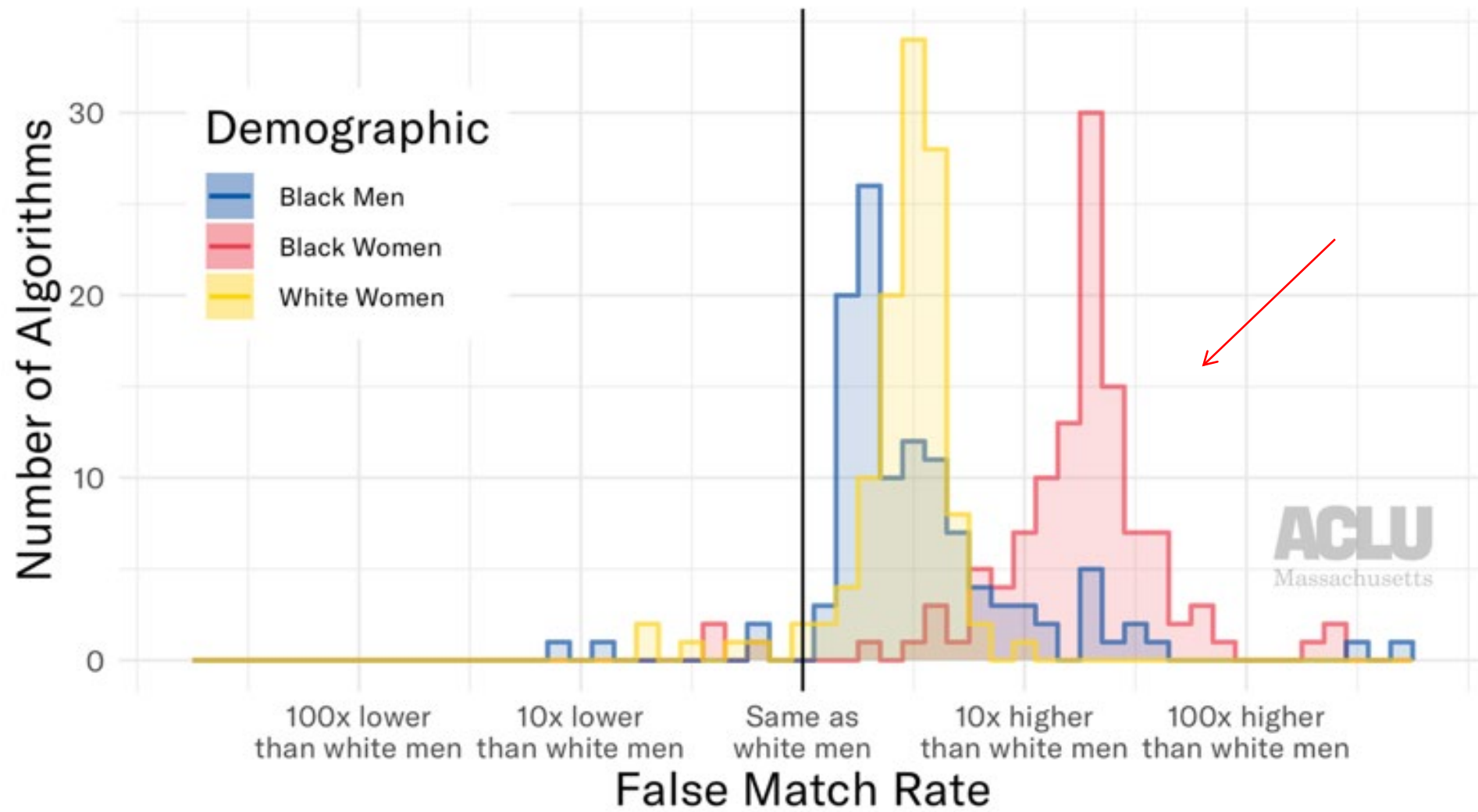
* Data from Jan 2014 - Gender data are global, ethnicity data are US only.

**See our [EEO-1 report](#) for more information. Ethnicity refers to the EEO-1 categories which we know are imperfect categorizations of race and ethnicity, but reflect the US government reporting requirements.

***Other includes American Indian/Alaskan Native and Native Hawaiian/Pacific Islander.





Is this an alignment problem or a problem of systemic bias?



[Follow this preprint](#)

Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare

 Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E. Abdunour, Atul J. Butte,  Emily Alsentzer

doi: <https://doi.org/10.1101/2023.07.13.23292577>

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.



Abstract

Full Text

Info/History

Metrics

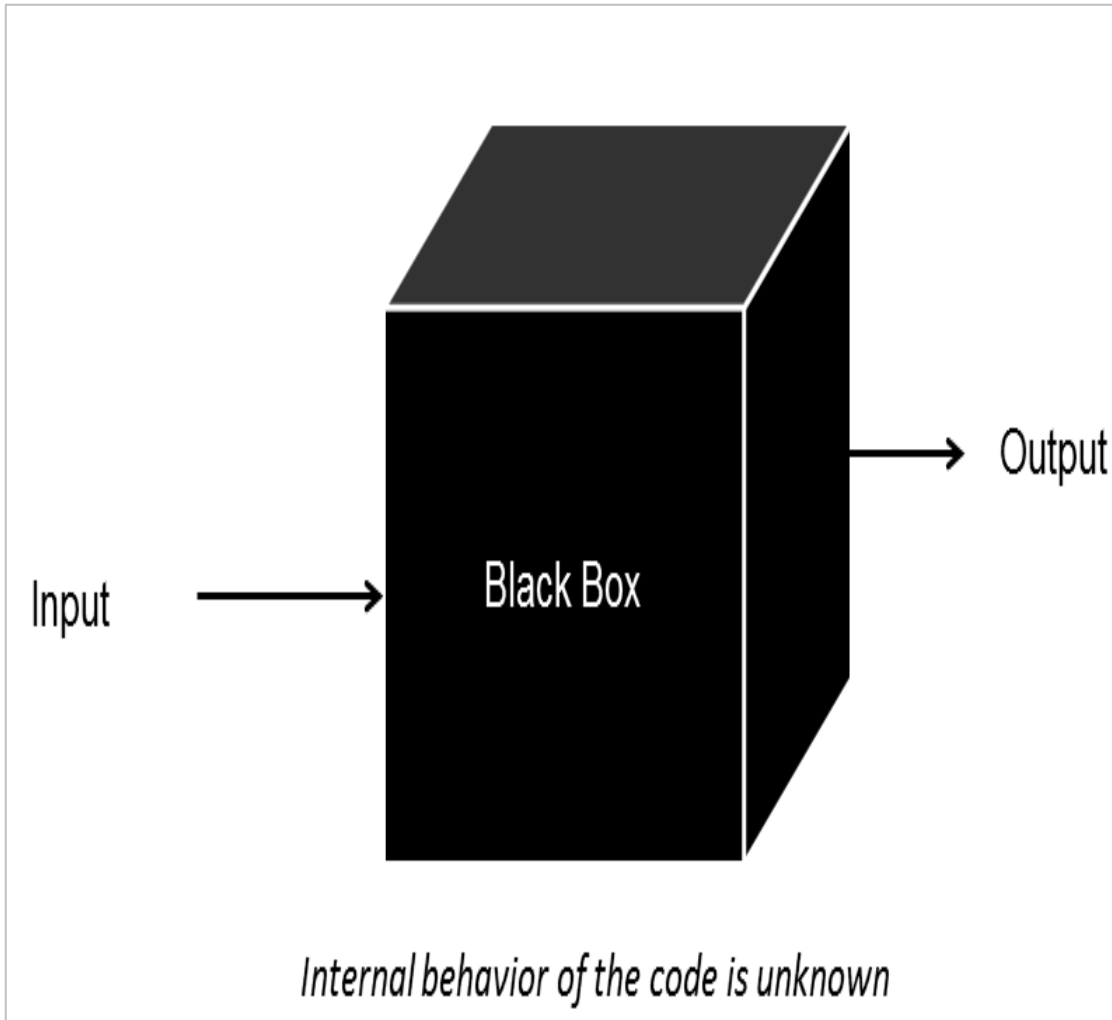
[Preview PDF](#)

Abstract

Background Large language models (LLMs) such as GPT-4 hold great promise as transformative tools in healthcare, ranging from automating administrative tasks to augmenting clinical decision-making. However, these models also pose a serious danger of perpetuating biases and delivering incorrect medical diagnoses, which can have a direct, harmful impact on medical care.

KEY FINDINGS

When there are known genetic and biological relationships between a disease and a patient's demographics, GPT-4 exaggerated these prevalence differences when generating clinical vignettes.



Black Box Society (Pasquale 2015)

1. Algorithms are increasingly used to make high-stakes decisions.
2. These decision-making mechanisms are hidden from public view and accountability.
3. There is no transparency in how those decisions are made by machines.
4. Invisible power.
5. Open to manipulation.

AI recognition of patient race in medical imaging: a modelling study



Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimoreddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrras, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Summary

Background Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed to conduct a comprehensive evaluation of the ability of AI to recognise a patient's racial identity from medical images.

Methods Using private (Emory CXR, Emory Chest CT, Emory Cervical Spine, and Emory Mammogram) and public (MIMIC-CXR, CheXpert, National Lung Cancer Screening Trial, RSNA Pulmonary Embolism CT, and Digital Hand Atlas) datasets, we evaluated, first, performance quantification of deep learning models in detecting race from medical images, including the ability of these models to generalise to external environments and across multiple imaging modalities. Second, we assessed possible confounding of anatomic and phenotypic population features by assessing the ability of these hypothesised confounders to detect race in isolation using regression models, and by re-evaluating the deep learning models by testing them on datasets stratified by these hypothesised confounding variables. Last, by exploring the effect of image corruptions on model performance, we investigated the underlying mechanism by which AI models can recognise race.

Findings In our study, we show that standard AI deep learning models can be trained to predict race from medical images with high performance across multiple imaging modalities, which was sustained under external validation conditions (x-ray imaging [area under the receiver operating characteristics curve (AUC) range 0.91–0.99], CT chest imaging [0.87–0.96], and mammography [0.81]). We also showed that this detection is not due to proxies or imaging-related surrogate covariates for race (eg, performance of possible confounders: body-mass index [AUC 0.55], disease distribution [0.61], and breast density [0.61]). Finally, we provide evidence to show that the ability of AI deep learning models persisted over all anatomical regions and frequency spectrums of the images, suggesting the efforts to control this behaviour when it is undesirable will be challenging and demand further study.

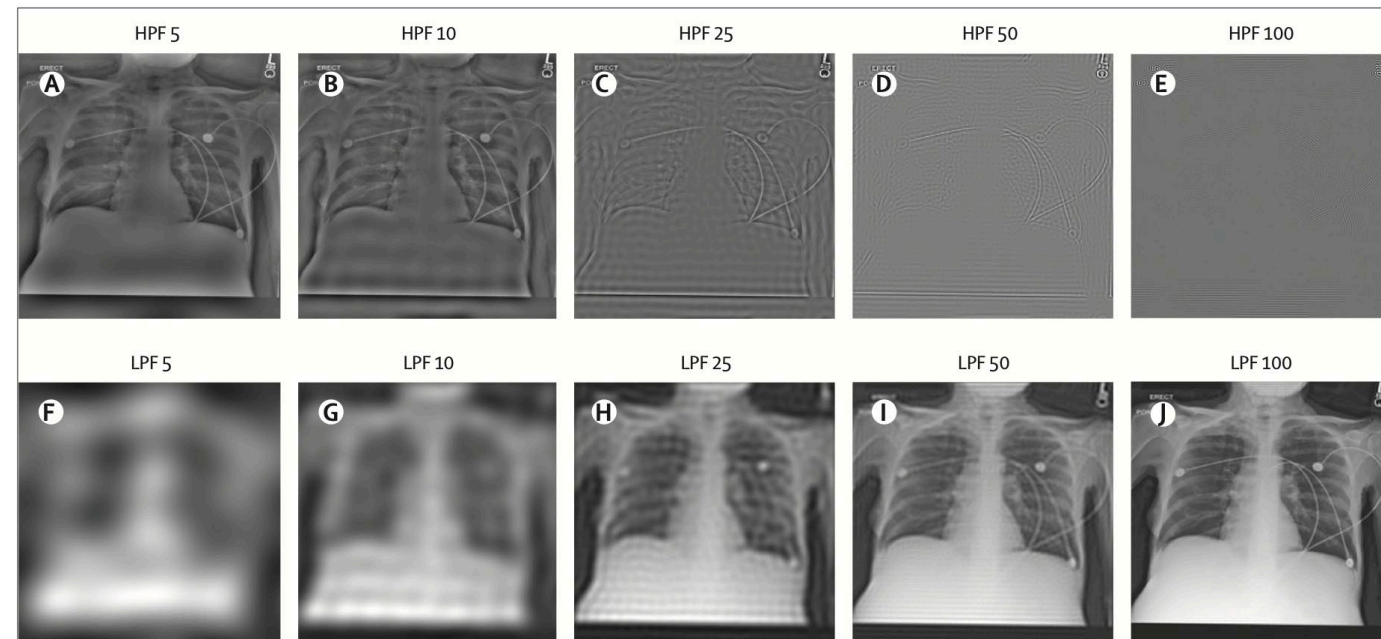
Lancet Digit Health 2022; 4: e406–14

Published Online
May 11, 2022
[https://doi.org/10.1016/S2589-7500\(22\)00063-2](https://doi.org/10.1016/S2589-7500(22)00063-2)

See [Comment](#) page e399

Department of Radiology (J W Gichoya MD, A R Bhimoreddy MS, H Trivedi MD) and Department of Computer Science (Z Zaiman), Emory University, Atlanta, GA, USA; School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA (I Banerjee PhD, R Correa BS); School of Informatics and Computing, Indiana University-Purdue University, Indianapolis, IN, USA (J L Burns MS, S Purkayastha PhD); Institute for Medical Engineering and Science (L A Celi MD, M Ghassemi PhD) and Department of Electrical

Transparency and Explainability: Medical Images





Historical Bias

Word embeddings;
nurse or Engineer are
encoded with gender biases



Representation Bias

When defining the target
population, if it does not
reflect the use population.



Measurement Bias

The accuracy of measurement
varies across groups.