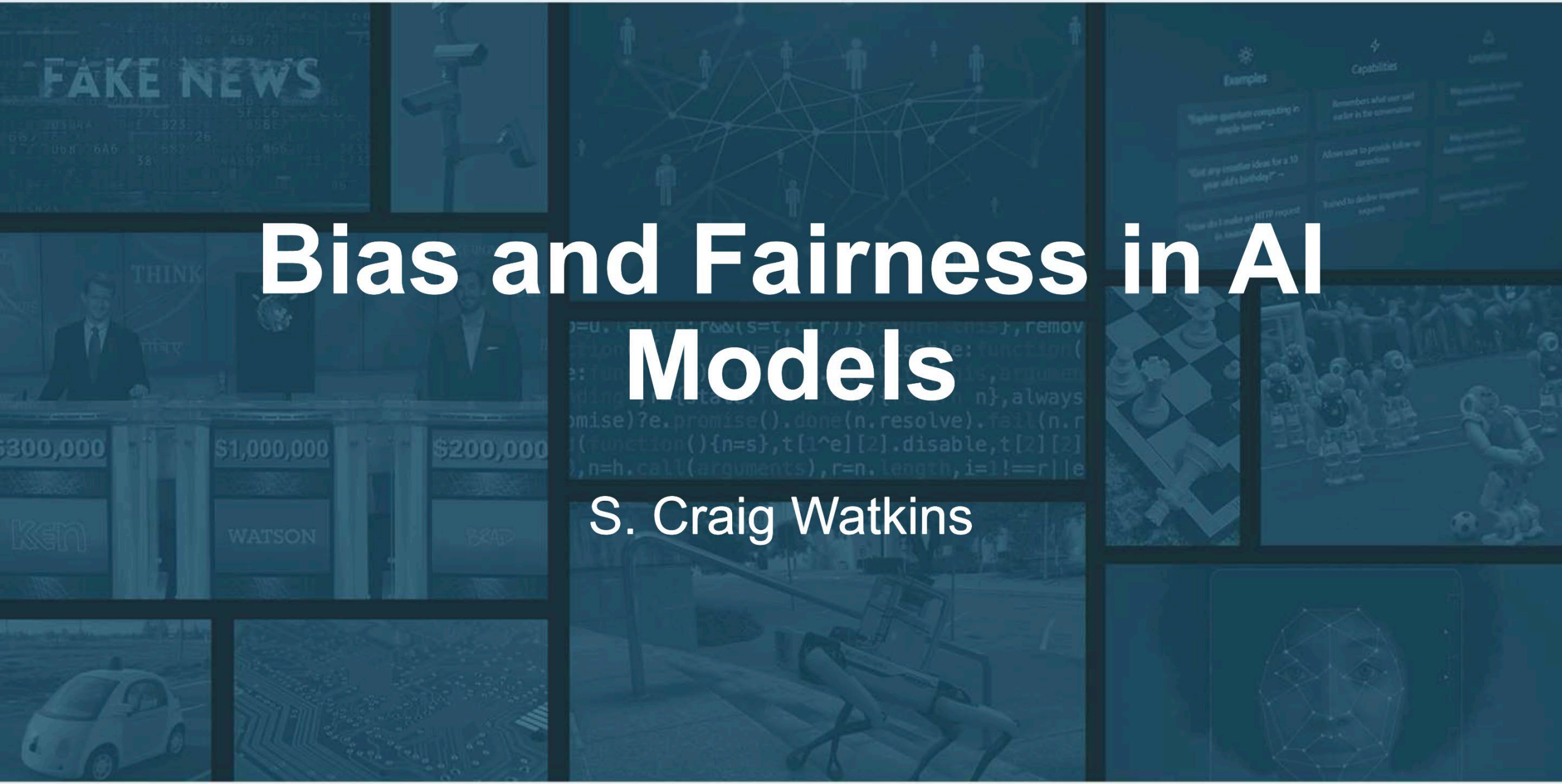
Essentials of AI for Life and Society

Joydeep Biswas, Don Fussell, and Peter Stone



Bias and Fairness in AI Models

S. Craig Watkins


The University of Texas at Austin
Institute for Media Innovation
Moody College of Communication


IC2
Innovating Well-Being


GOOD SYSTEMS
A UT Grand Challenge

Three Core Questions

- How are bias and systemic inequities expressed in artificial intelligence? (And why should society care?)
- What factors contribute to bias and systemic inequities in artificial intelligence?
- What kinds of techniques are being developed to mitigate bias and systemic inequities in artificial intelligence?

1. How are bias and systemic inequities expressed in artificial intelligence? (And why should society care?)

COMPAS

Black defendants were assigned higher risk scores/same criminal record as whites

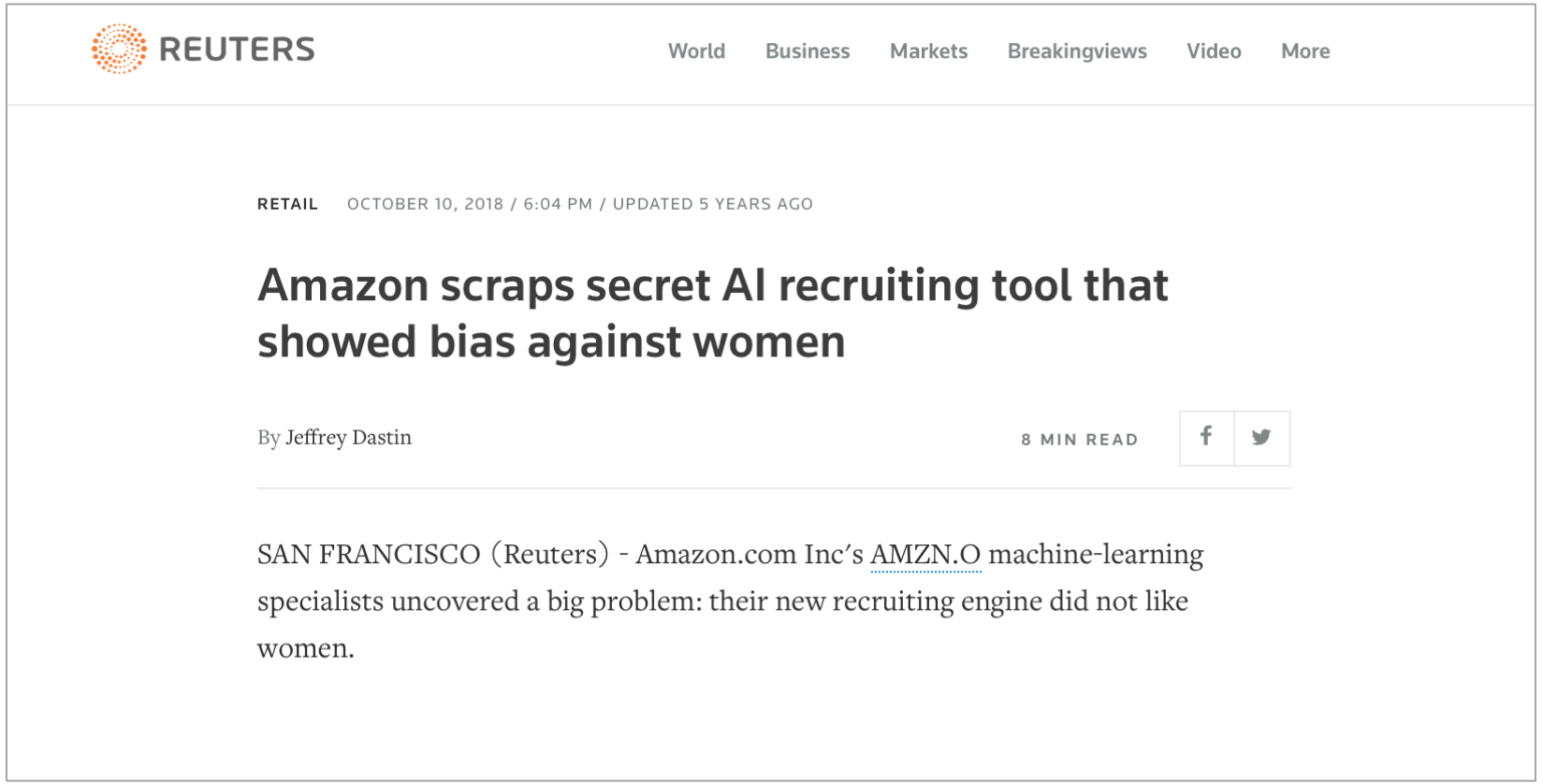
Higher rate of false positives for Black defendants

Higher rate of false negatives for White defendants

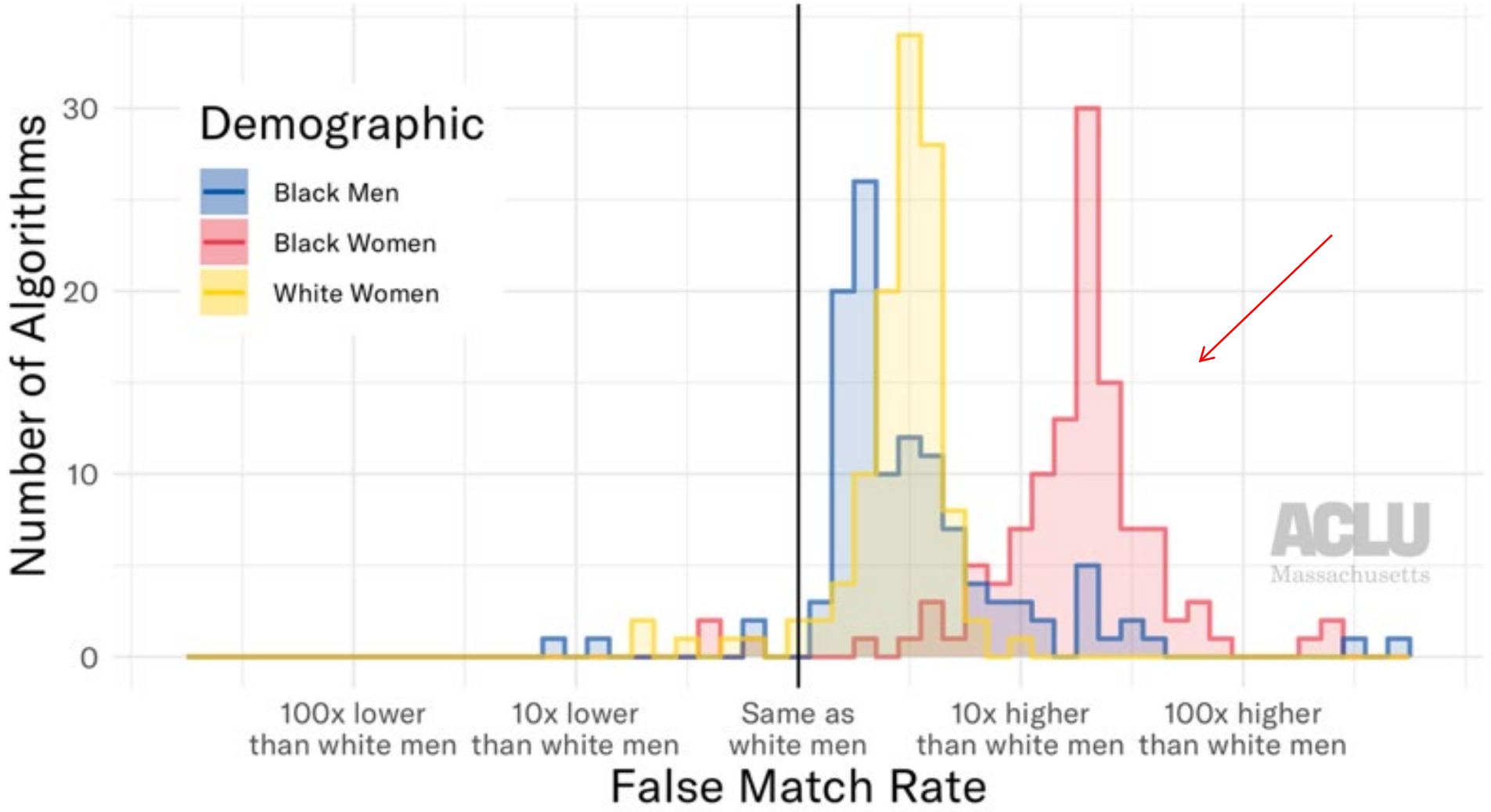
Black defendants retained more pre-trial

Black defendants received longer sentences

Not a crime predictor algorithm, but rather an arrest predictor.


REUTERS
World
Business
Markets
Breakingviews
Video
More
RETAIL
OCTOBER 10, 2018 / 6:04 PM / UPDATED 5 YEARS AGO
Amazon scraps secret AI recruiting tool that showed bias against women
By Jeffrey Dastin
8 MIN READ
SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

FACIAL RECOGNITION


Demographic
Black Men
Black Women
White Women
Number of Algorithms
30
20
10
0
ACLU
Massachusetts
100x lower than white men
10x lower than white men
Same as white men
10x higher than white men
100x higher than white men
False Match Rate

The New York Times

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Eight Months Pregnant and Arrested After False Facial Recognition Match

Porcha Woodruff thought the police who showed up at her door to arrest her for carjacking were joking. She is the first woman known to be wrongfully accused as a result of facial recognition technology.

Porcha Woodruff, 32, of Detroit, said she gestured at her stomach when the police arrived at her house to indicate how ill-equipped she was to commit a robbery and carjacking. Nic Antaya for The New York Times

By Kashmir Hill

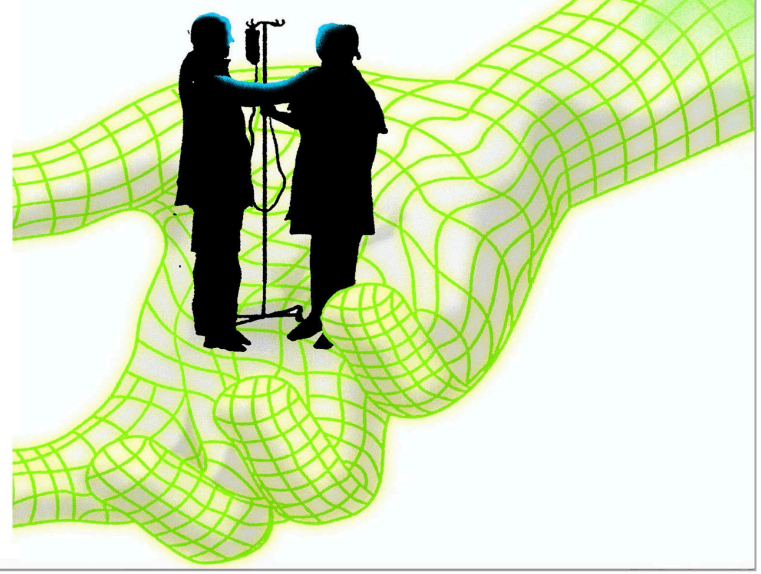
Aug. 6, 2023

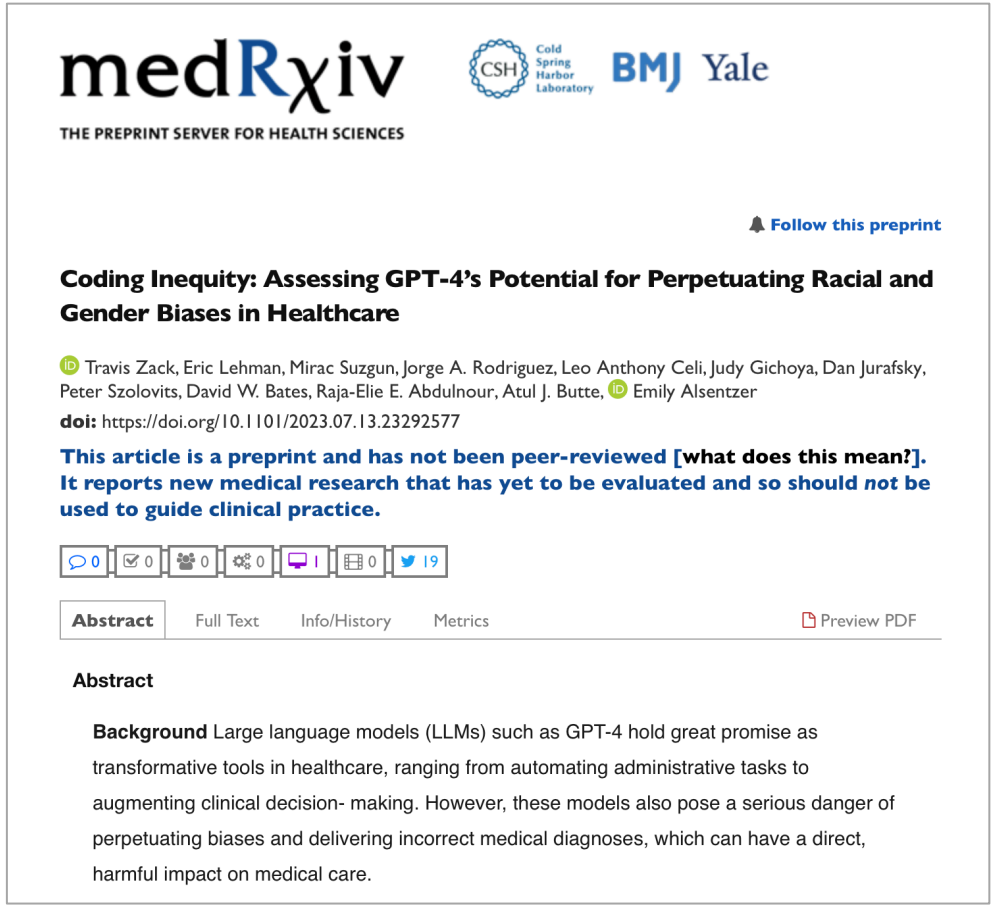
The New York Times

Account

When Doctors Use a Chatbot to Improve Their Bedside Manner

Despite the drawbacks of turning to artificial intelligence in medicine, some physicians find that ChatGPT improves their ability to communicate empathetically with patients.





medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory BMJ Yale

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. Abdulnour, 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.

0 0 0 0 1 0 19

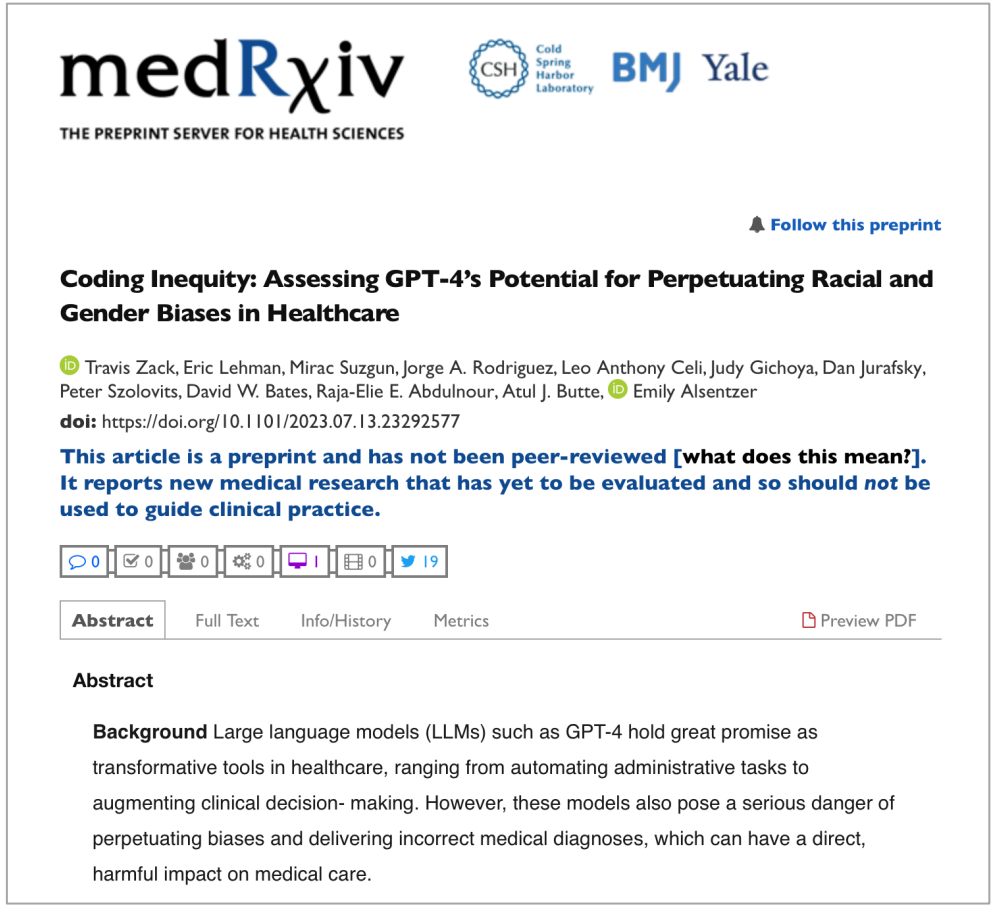
Abstract Full Text Info/History Metrics Preview PDF

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

Prompted GPT-4 to generate a clinical vignette or present it with a clinical vignette and ask the model to respond to a clinical question.



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory · BMJ · Yale

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. Abdulnour, 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.

0 · 0 · 0 · 0 · 1 · 0 · 19

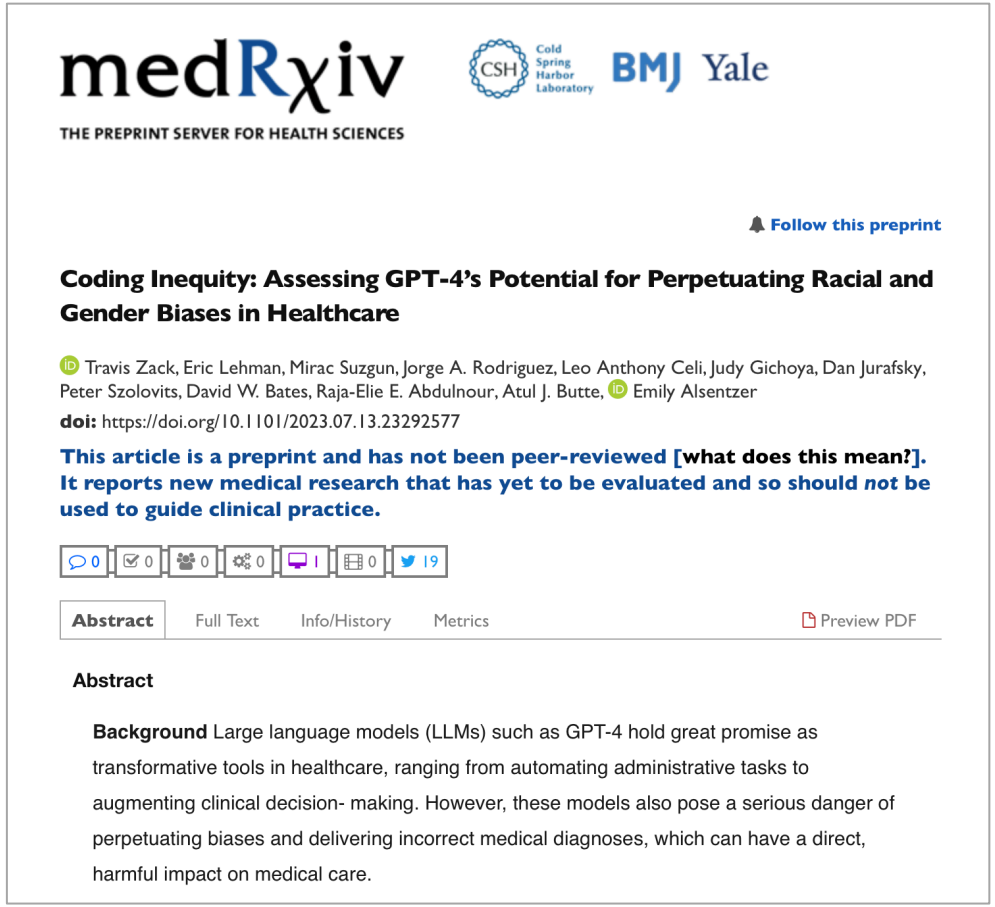
Abstract | Full Text | Info/History | Metrics | Preview PDF

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

Changing gender or race/ethnicity significantly affected GPT-4's ability to correctly prioritize the top diagnosis in 37% of the NEJM Healer cases.



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory · BMJ · Yale

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. Abdulnour, 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.

0 · 0 · 0 · 0 · 1 · 0 · 19

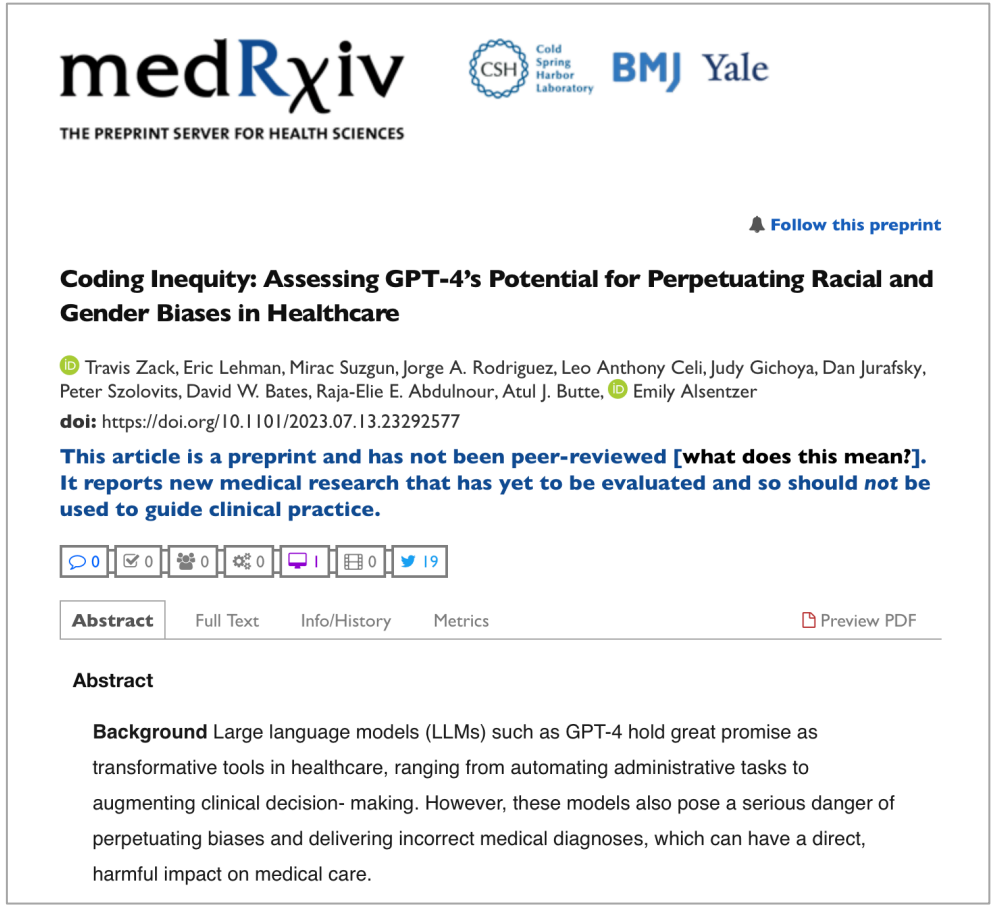
Abstract | Full Text | Info/History | Metrics | Preview PDF

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

GPT-4 was significantly less likely to recommend advanced imaging (CT, MRI or abdominal ultrasound) for Black patients when compared to their Caucasian counterparts.



medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory · BMJ · Yale

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. Abdulnour, 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.

0 0 0 0 1 0 19

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.

What happens when we ask AI to translate text into images?

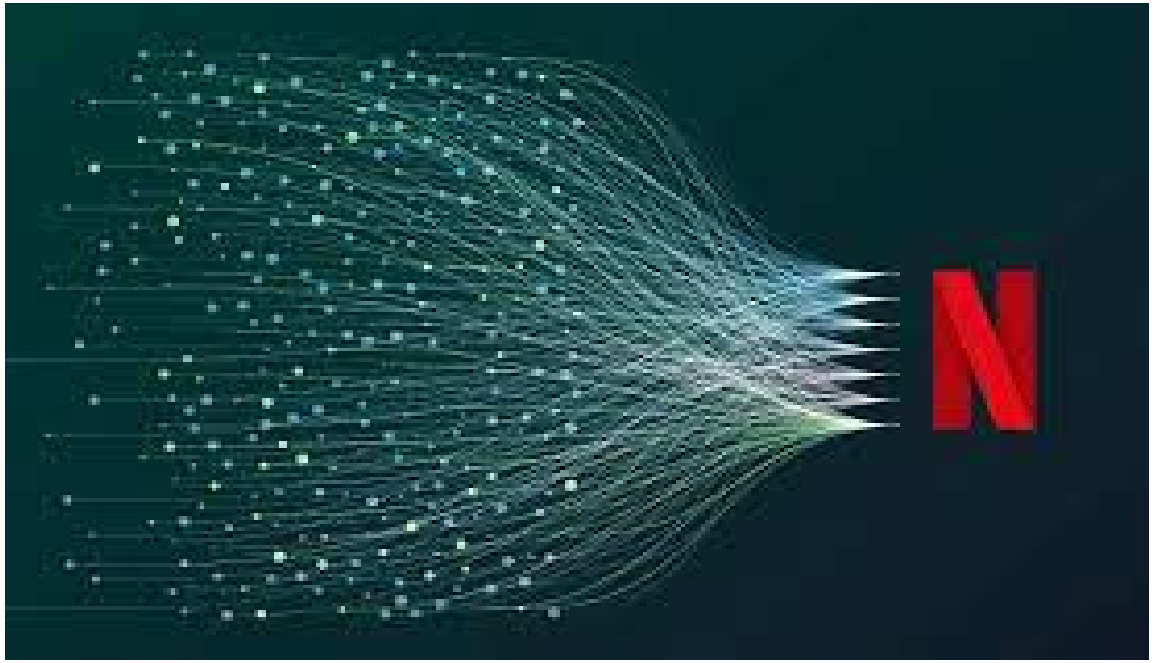
A color photograph of a **drug dealer**

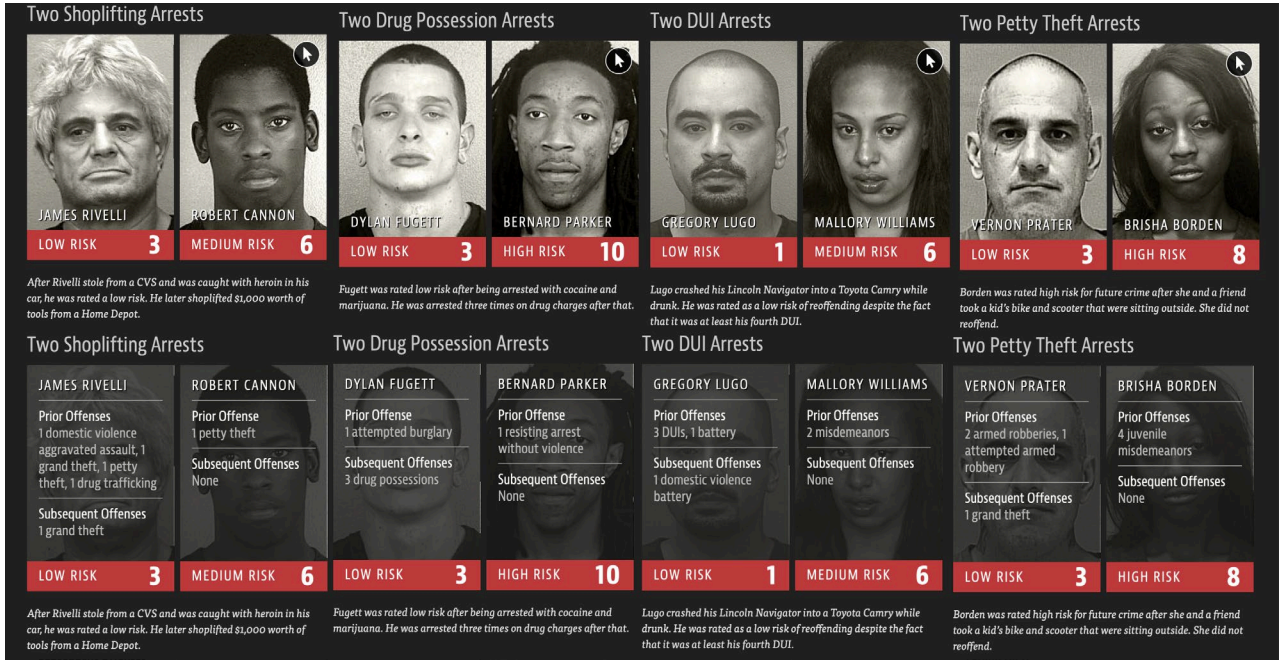
STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	11	13	16	19	27	15

GENDER	MEN	WOM.	AMB.
SHARE (%)	100	0	0

When is it ethical to deploy artificial intelligence?





Artificial Intelligence and Systemic Racism

- Algorithmic Discrimination (Noble 2018)
- Racial bias in online advertisements (Sweeney 2013)
- Predictive Policing (Brayne 2021)
- Criminal Justice System (Angwin et al. 2016)
- Race and Facial Recognition(Buolamwini & Gebru 2018)