Machine Learning Systems

Lecture 3: Trustworthy AI

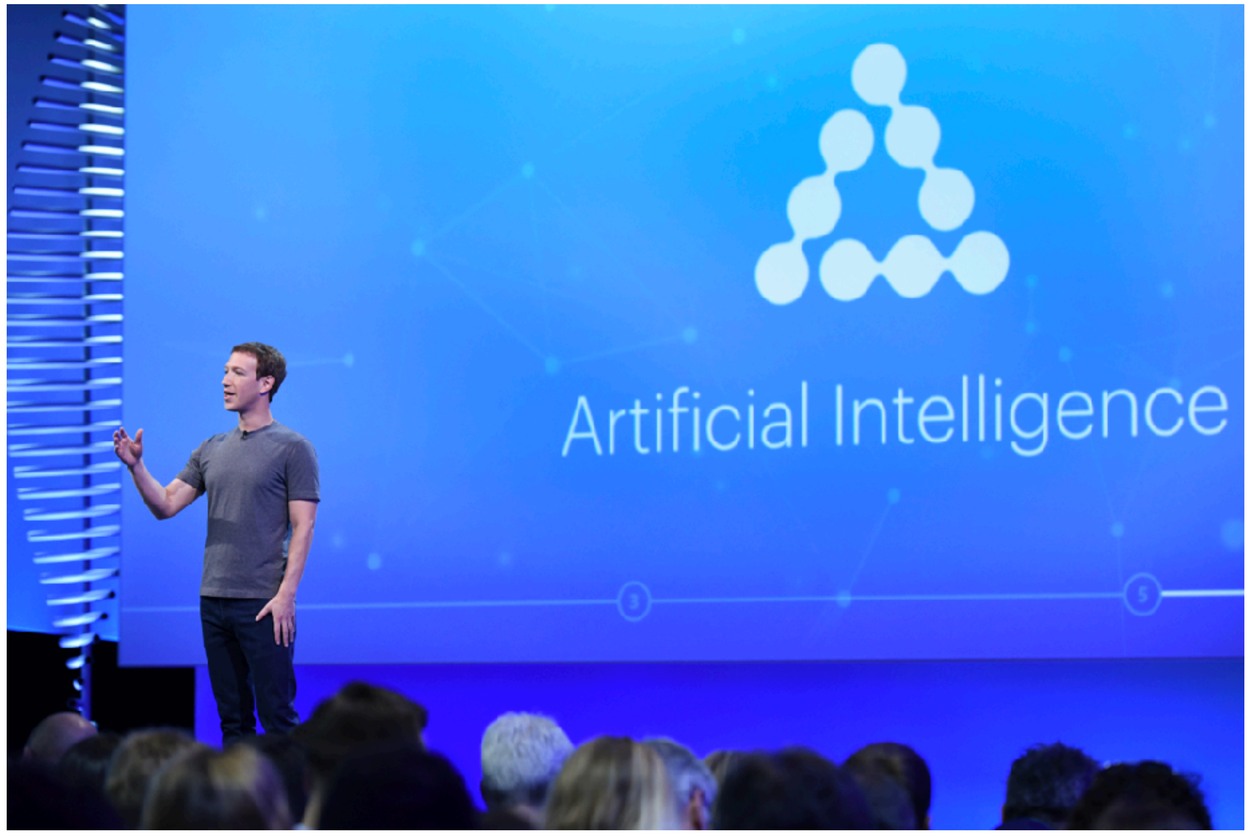
Pooyan Jamshidi

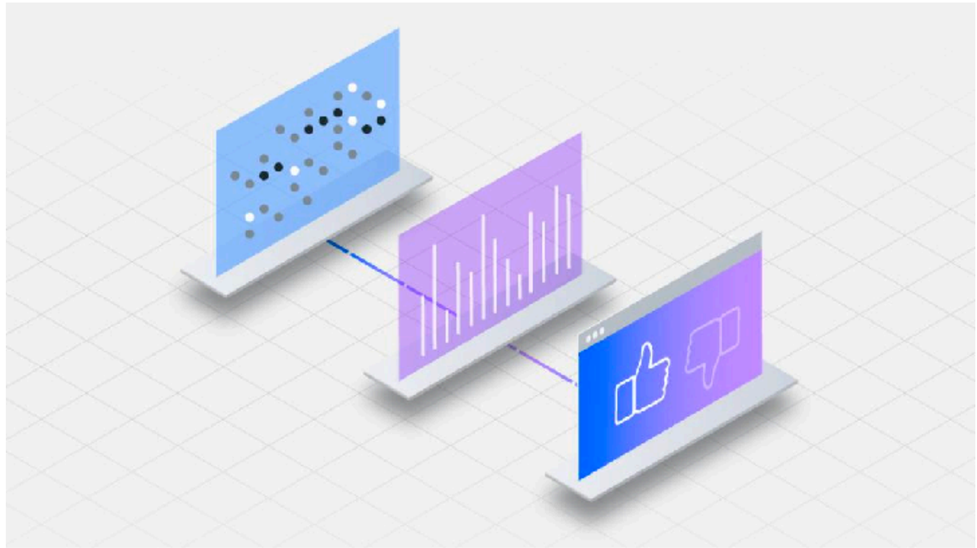
AI is becoming the integral part of our everyday life

They are taking over our society too!

AI is becoming the integral part of our everyday life

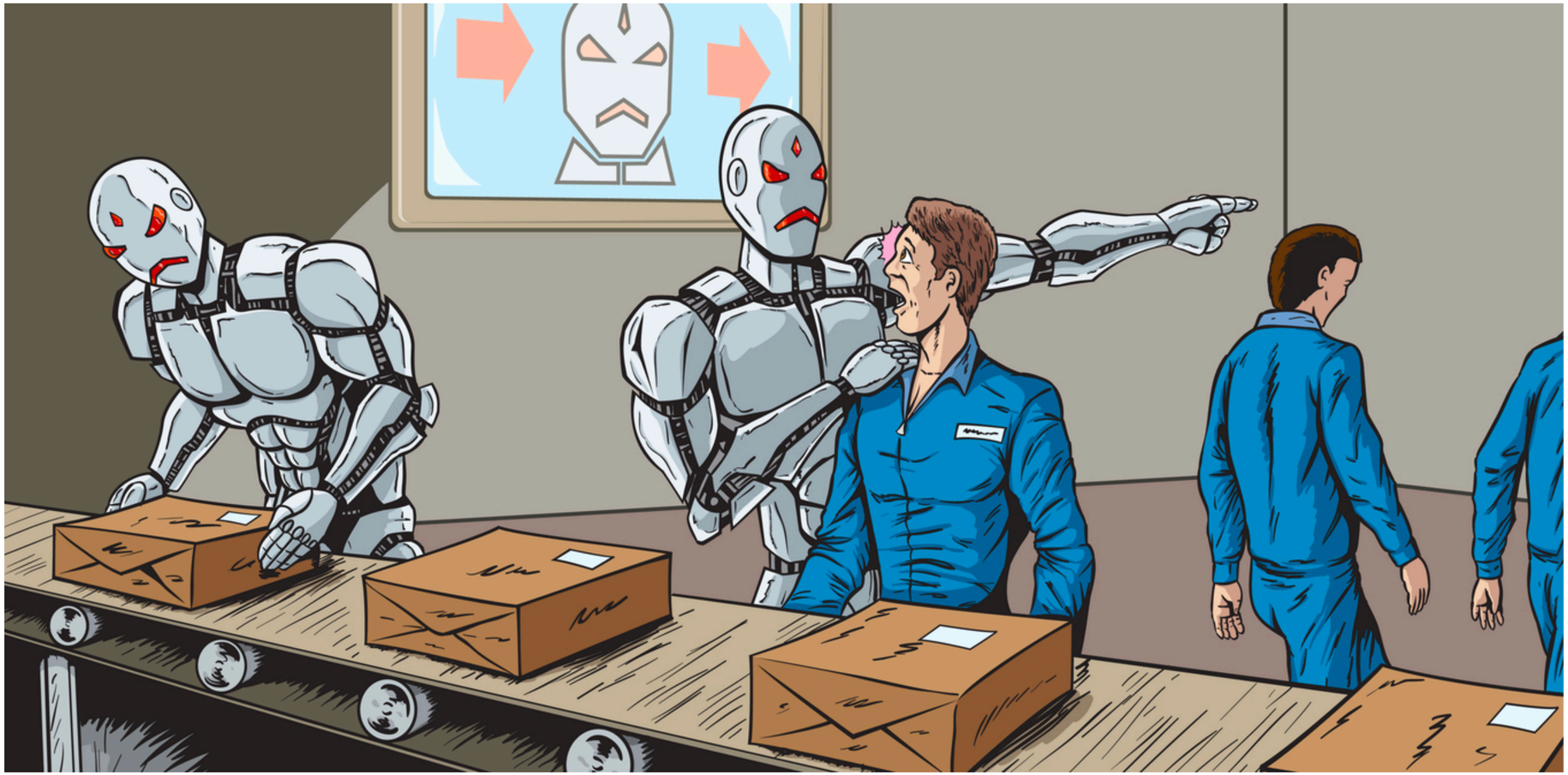
Should we be worried?





AI is becoming the integral part of our everyday life

Should we be worried?



AI could be racist

Algorithmic bias

NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

Heidi Ledford

Black people with complex medical needs were less likely than equally ill white people to be referred to programmes that provide more personalized care. Credit: Ed Kashi/VII/Redux/eyevine

PDF version

RELATED ARTICLES

A fairer way forward for AI in health care

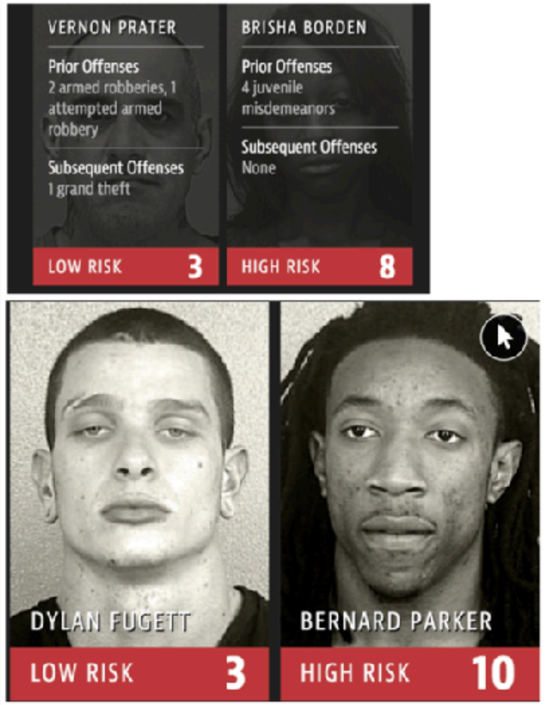
Bias detectives: the researchers striving to make algorithms fair

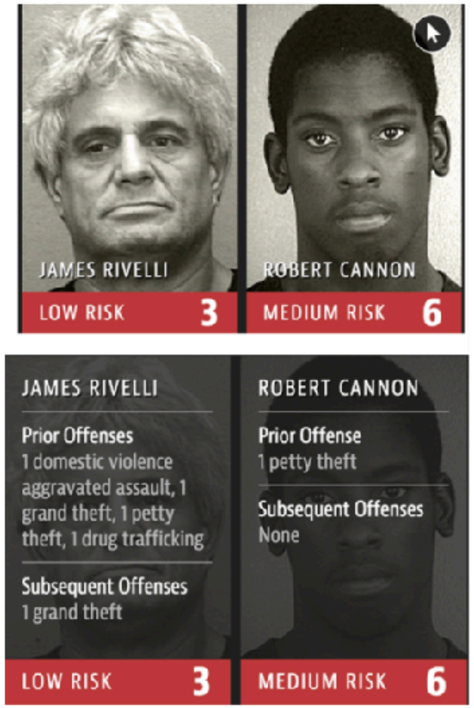
Can we open the black box of AI?

SUBJECTS

Computer science · Health care · Policy · Society

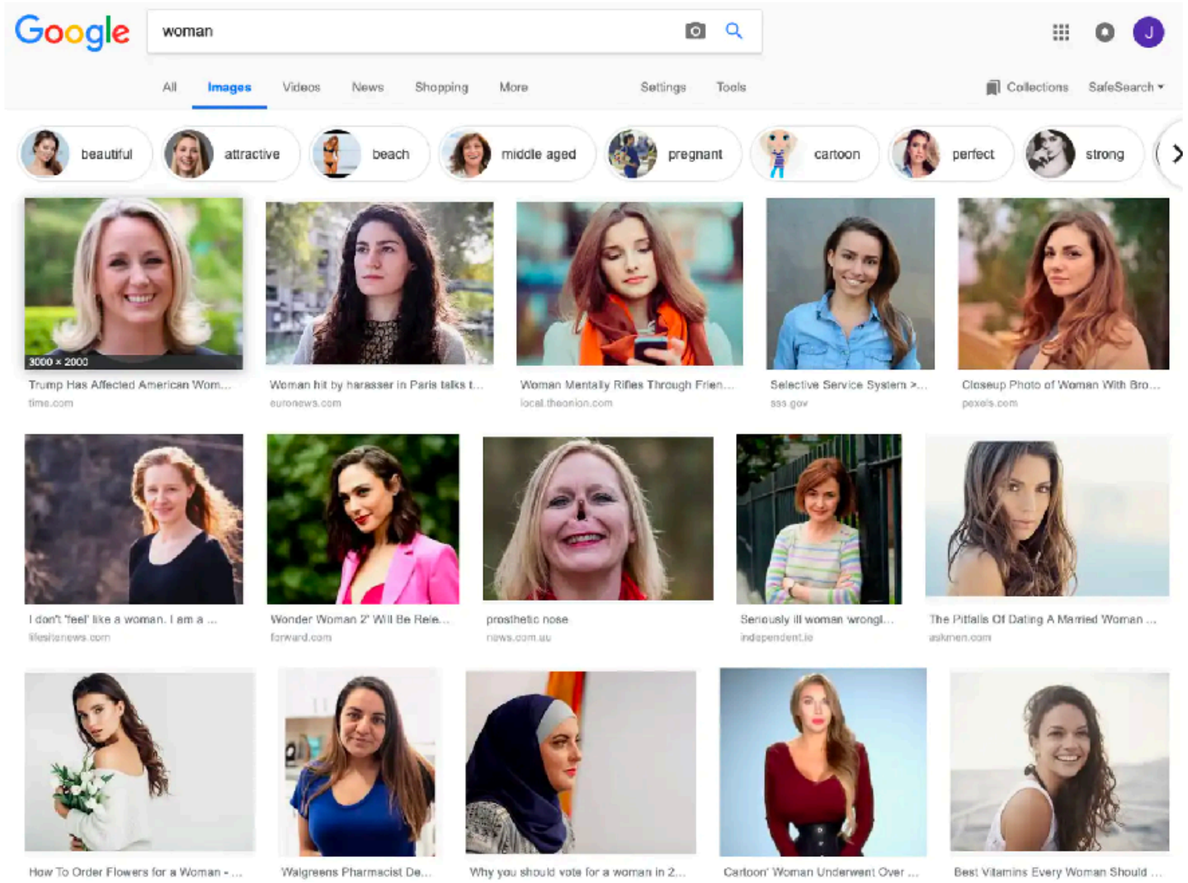
An algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, a sweeping analysis has found.

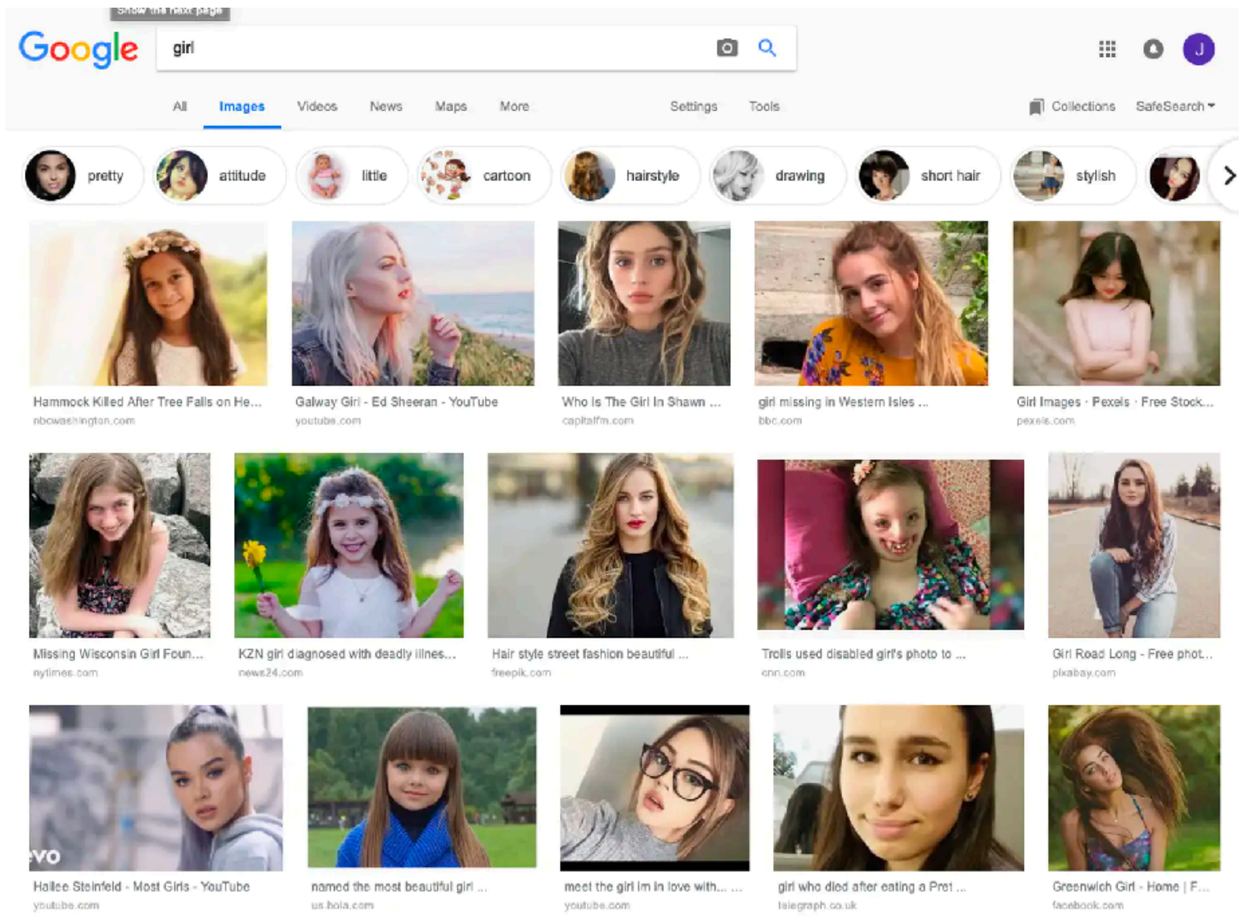




AI could be racist

Algorithmic bias





AI could be also gender biased

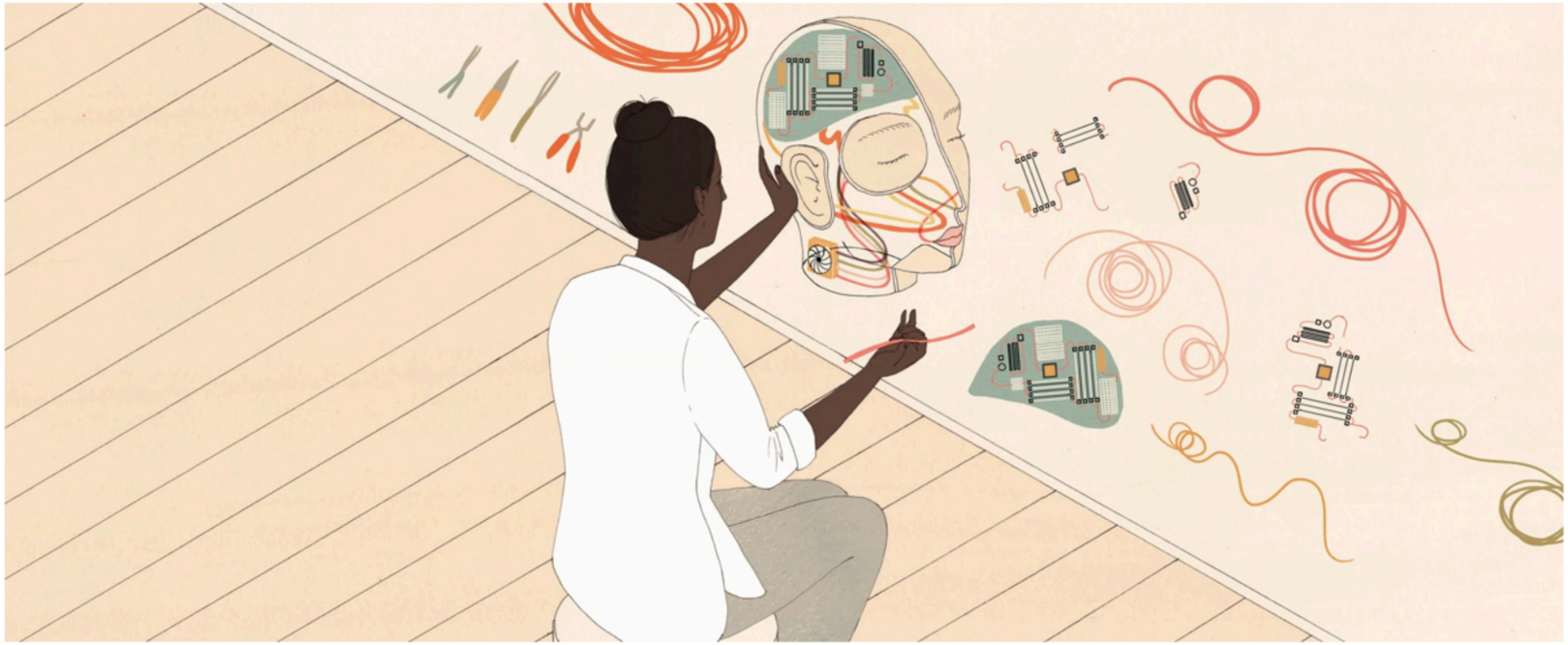
Algorithmic bias

AI could be also gender biased

Algorithmic bias

Dealing With Bias in Artificial Intelligence

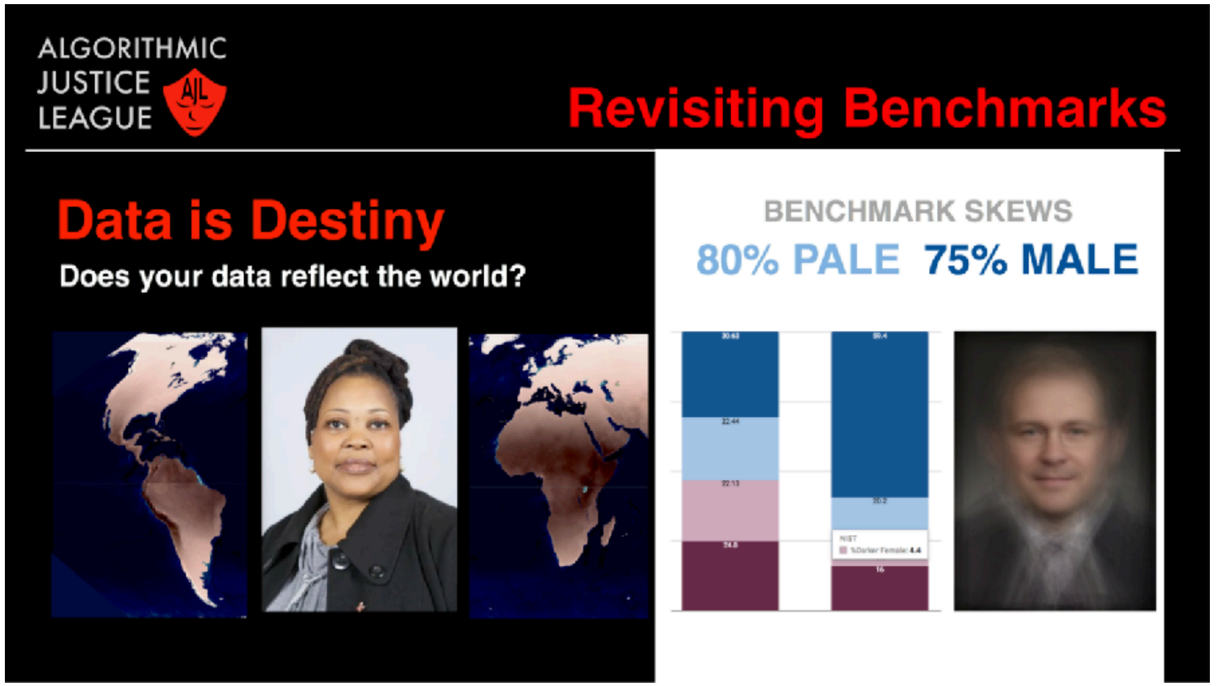
Three women with extensive experience in A.I. spoke on the topic and how to confront it.



Harriet Lee-Merrion

What is the source of the problem?

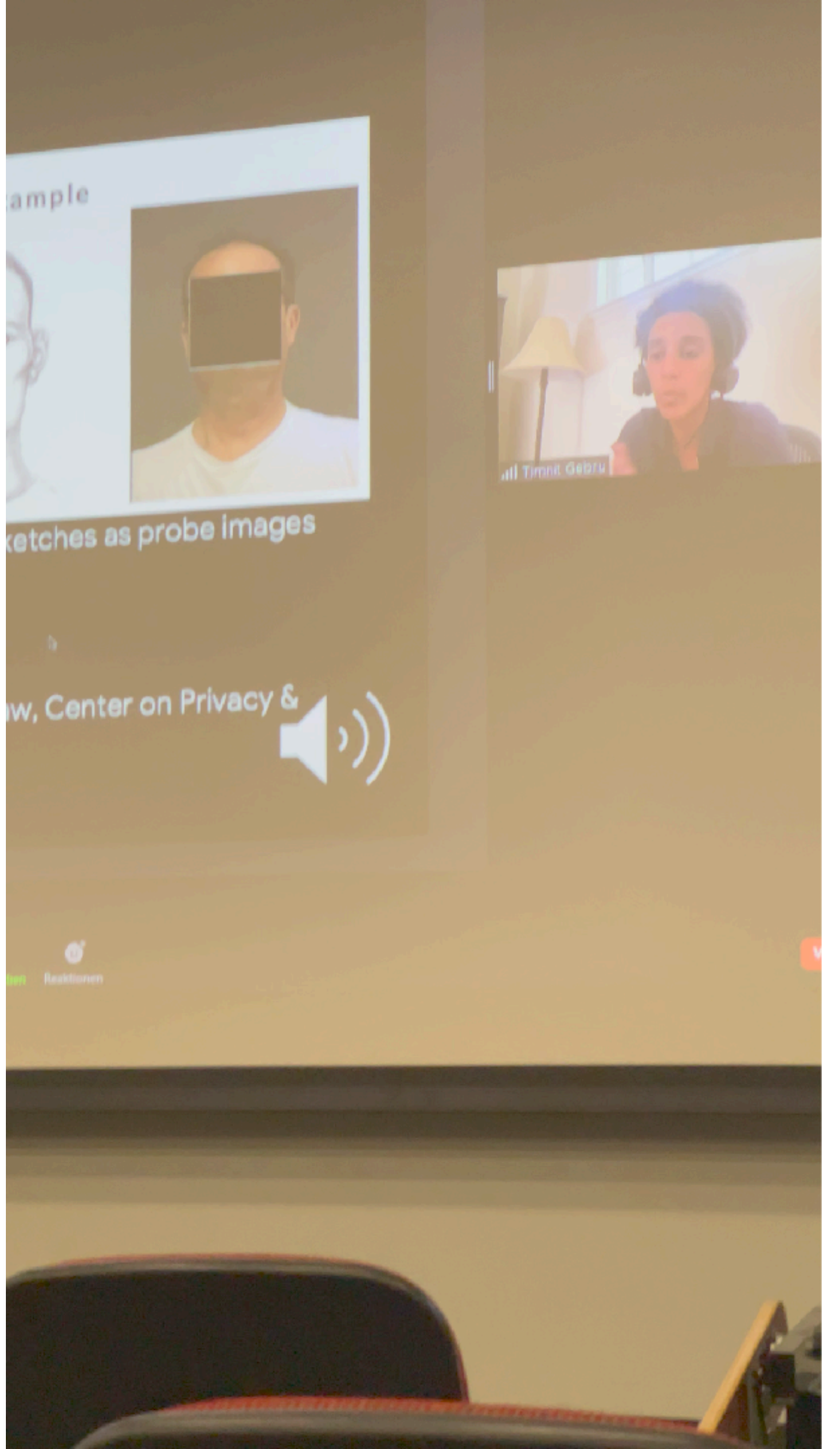
Data or Algorithms or Both?



Do we need systems like face recognition for law enforcement?


US ADULTS INDEXED
130 MILLION
One in two American adults is in a law enforcement face recognition network used in unrestricted searches employing algorithms with unaudited accuracy.
Facial Recognition is the Plutonium of AI
It's dangerous, racializing, and has few legitimate uses; facial recognition needs regulation and control on par with nuclear waste.
By Luke Stark

Timnit Gebru is explaining a few consequences of having face recognition systems deployed in our society


ample
ketches as probe images
w, Center on Privacy &
Timnit Gebru

Hilde Weerts believes such systems are Sociotechnical systems, and we need to Sociotechnical approach


Es spricht: Hilde Weerts

Timnit Gebru is explaining a few consequences of having face recognition systems deployed in our society


Es spricht: Timnit Gebru
Teilnehmer
Chat
Bildschirm freigeben
Reaktionen
Verlassen

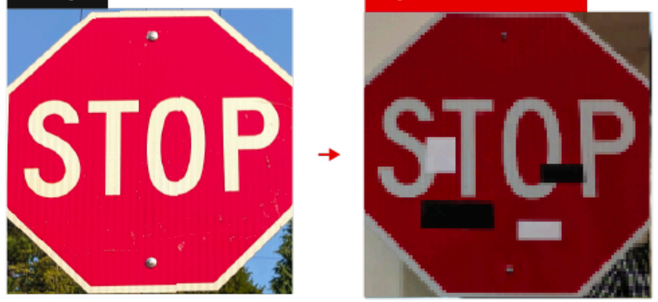
AI/ML Systems can be easily fooled!

What? Yes, it is true, and the implications could be massive!

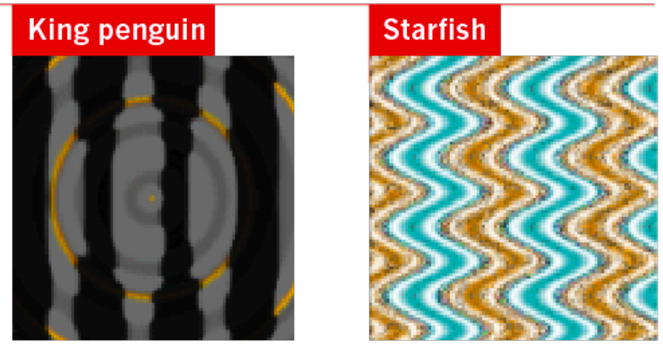
FOOLING THE AI

Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.



Scientists have evolved images that look like abstract patterns — but which DNNs see as familiar objects.



©nature

PERCEPTION PROBLEMS

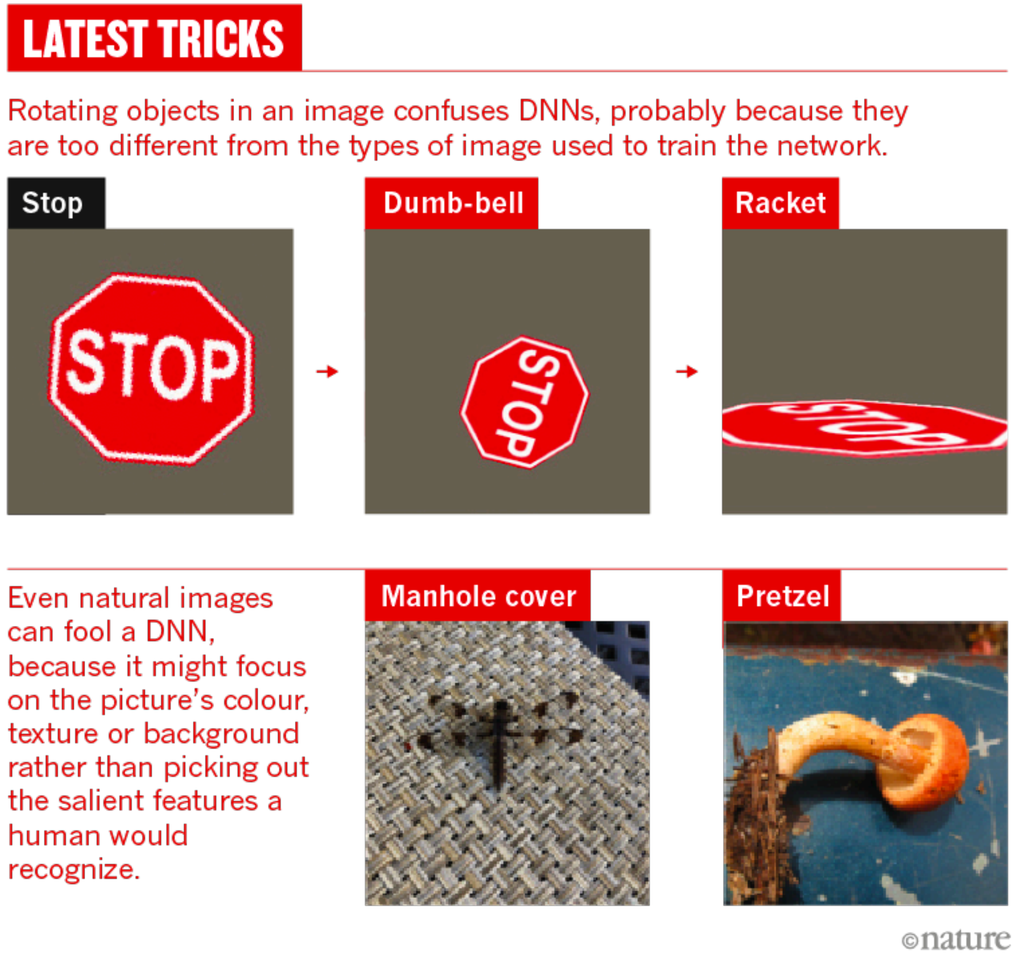
Adding carefully crafted noise to a picture can create a new image that people would see as identical, but which a DNN sees as utterly different.

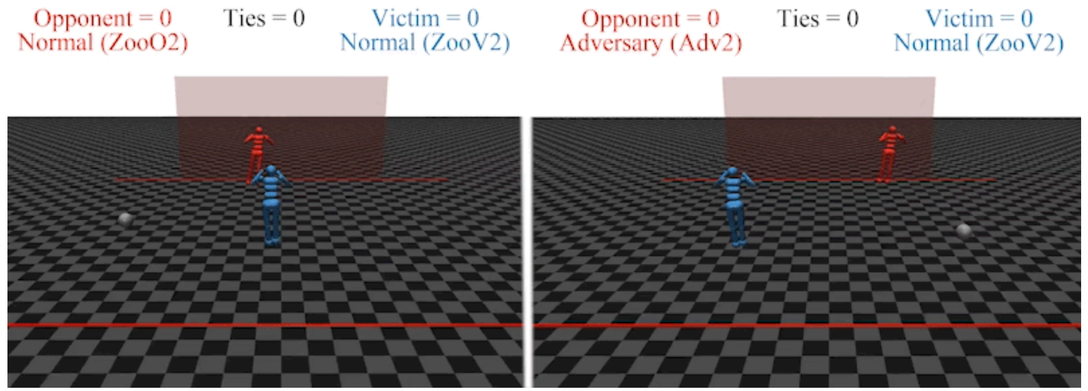
In this way, any starting image can be tweaked so a DNN misclassifies it as any target image a researcher chooses.

©nature

AI/ML Systems can be easily fooled!

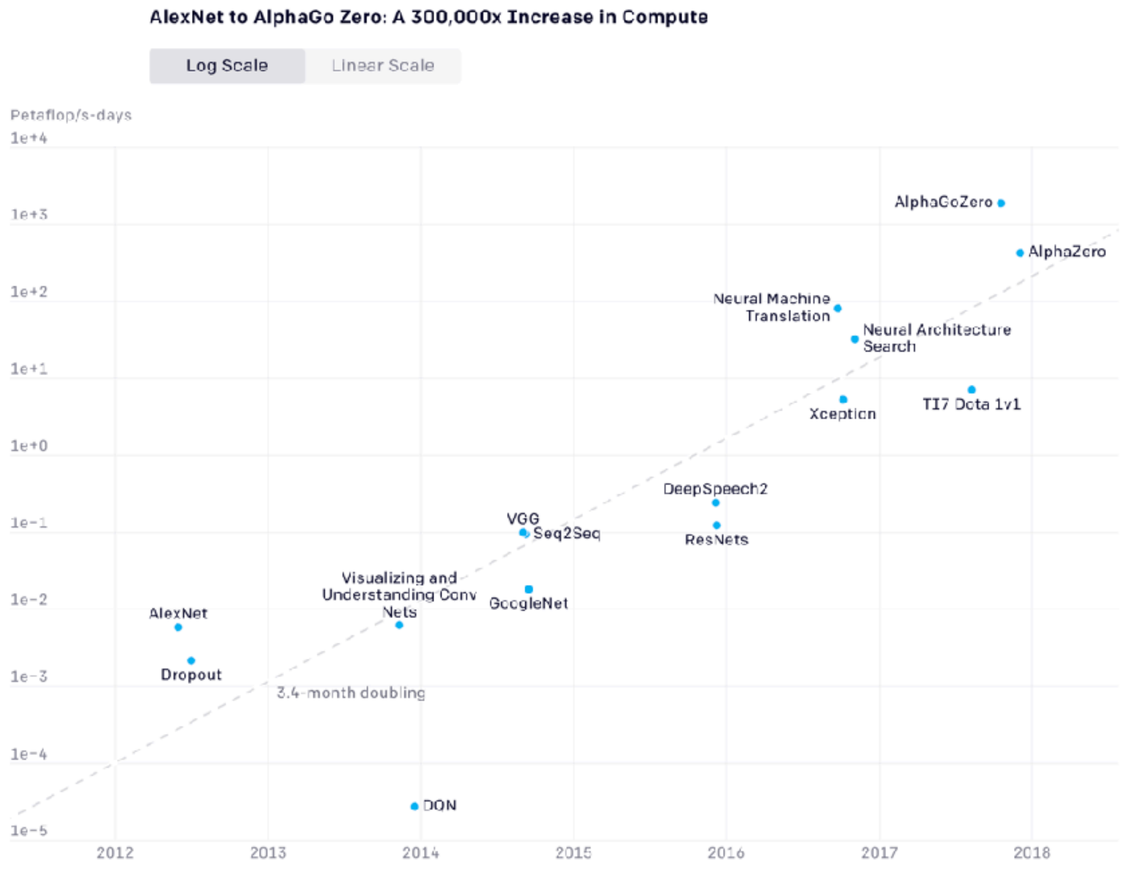
What? Yes, it is true, and the implications could be massive!

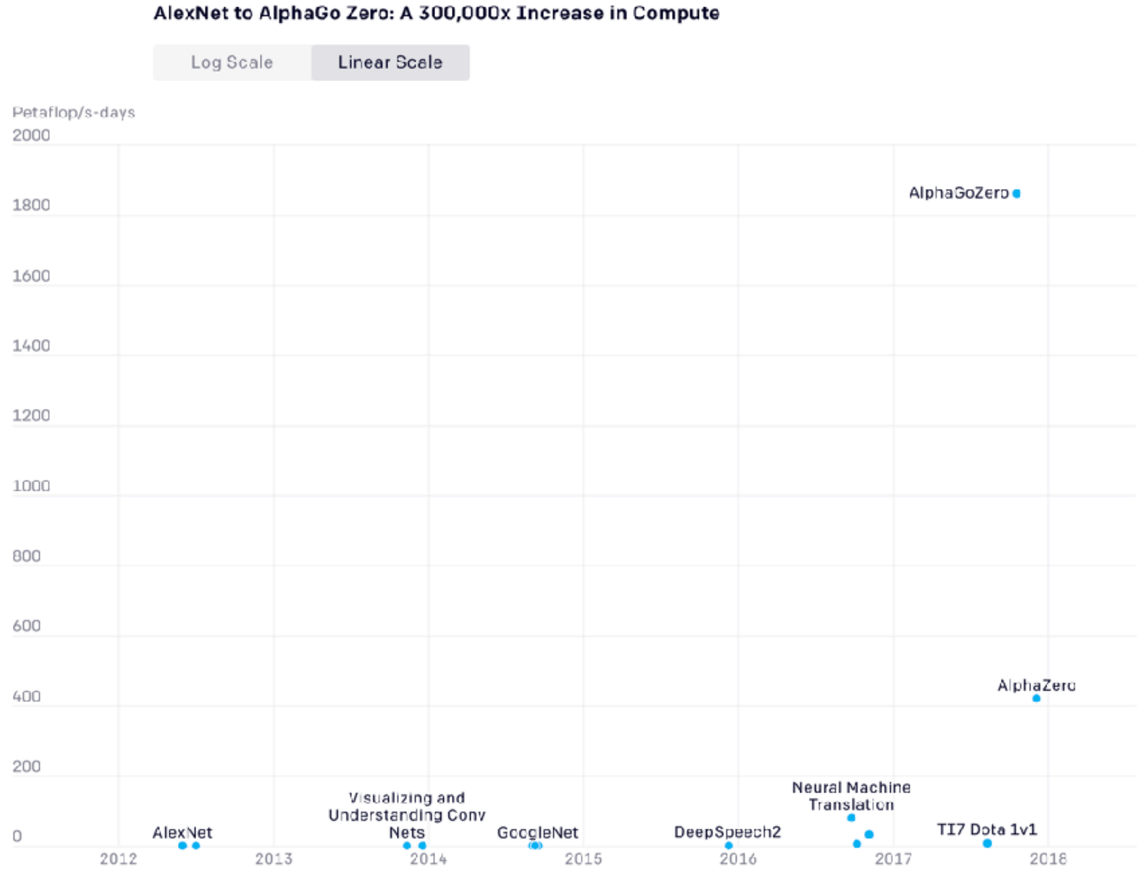




AI and Compute

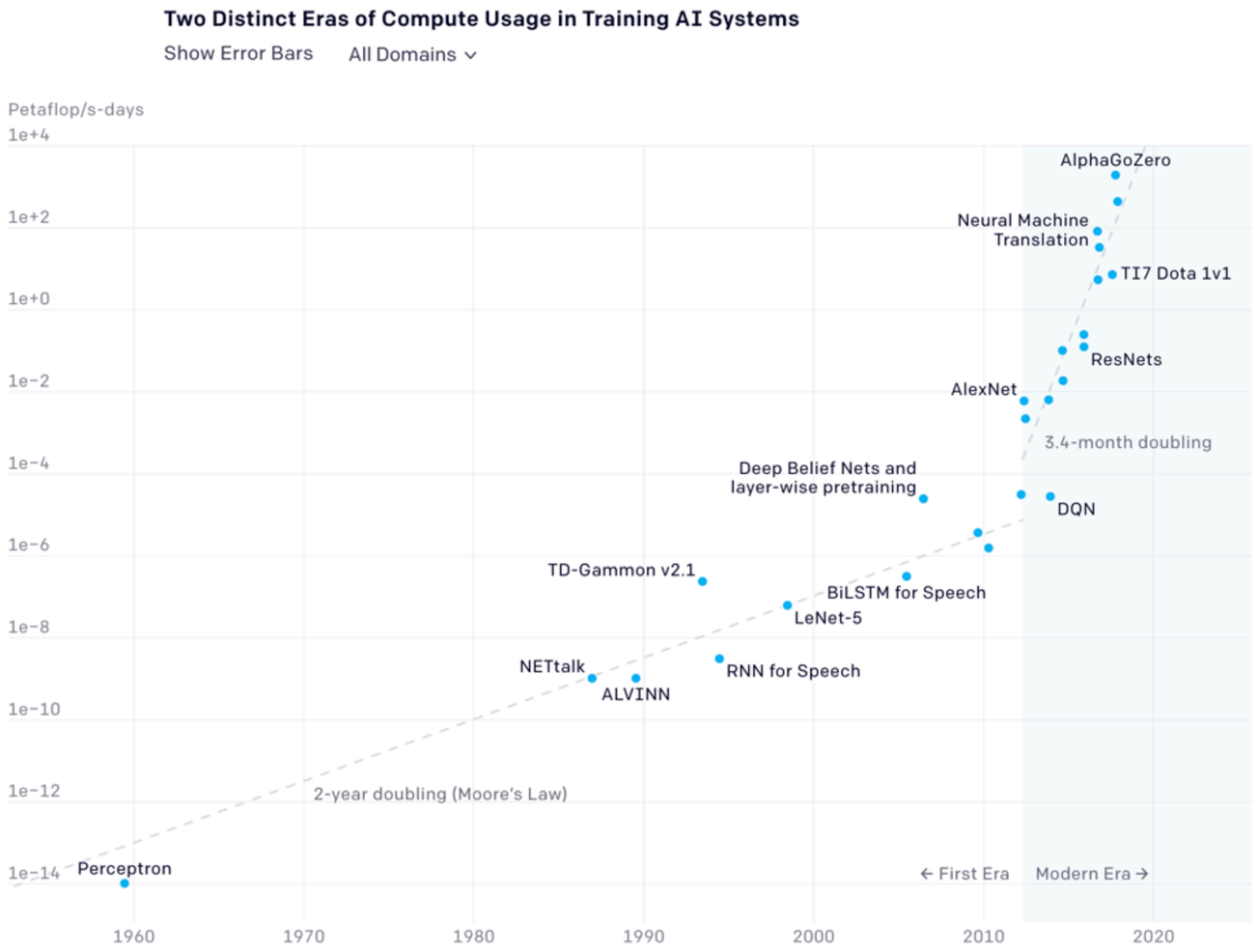
The amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time (by comparison, Moore's Law had a 2-year doubling period).





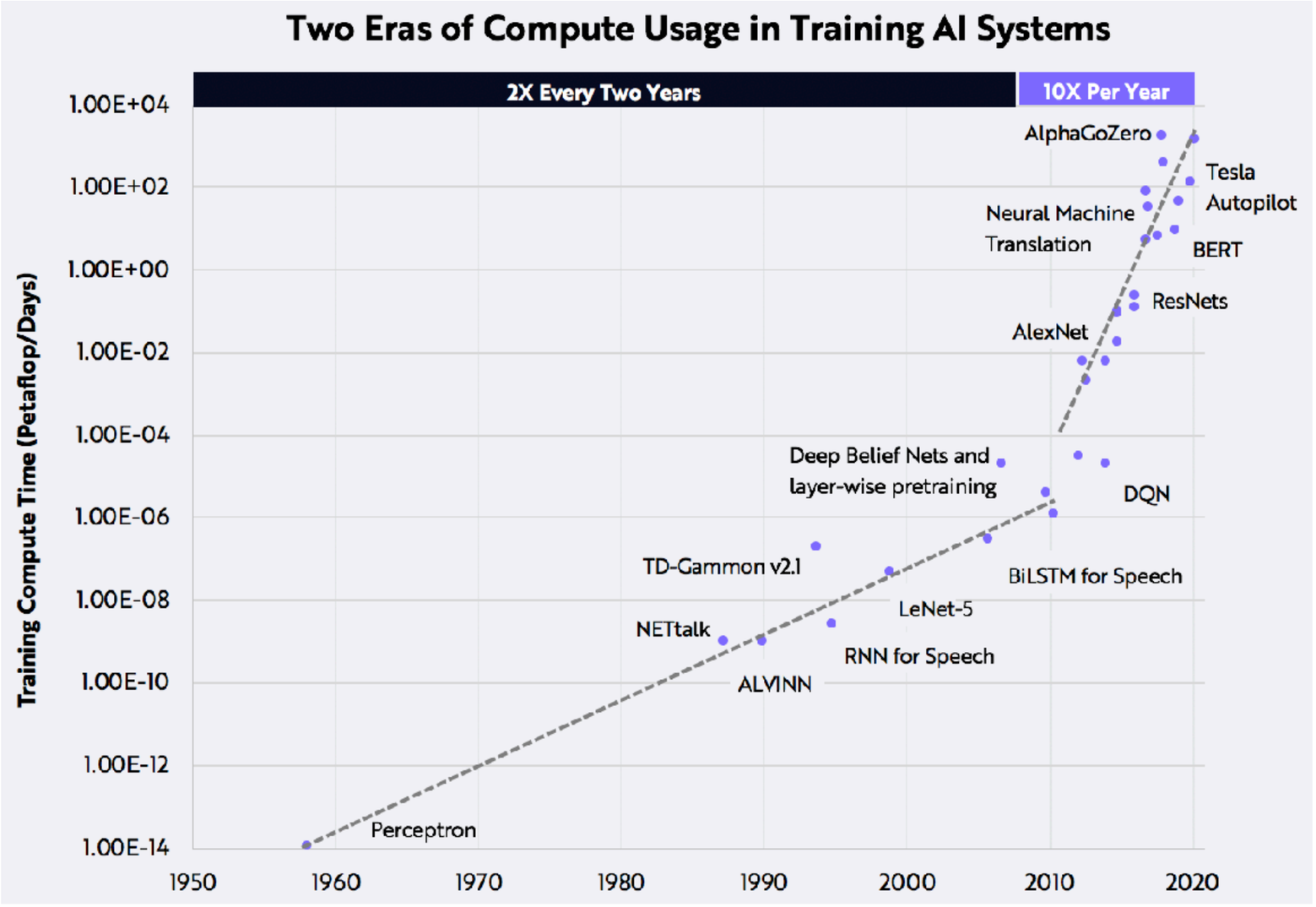
AI and Compute

Two Distinct Eras of Compute Usage in Training AI Systems



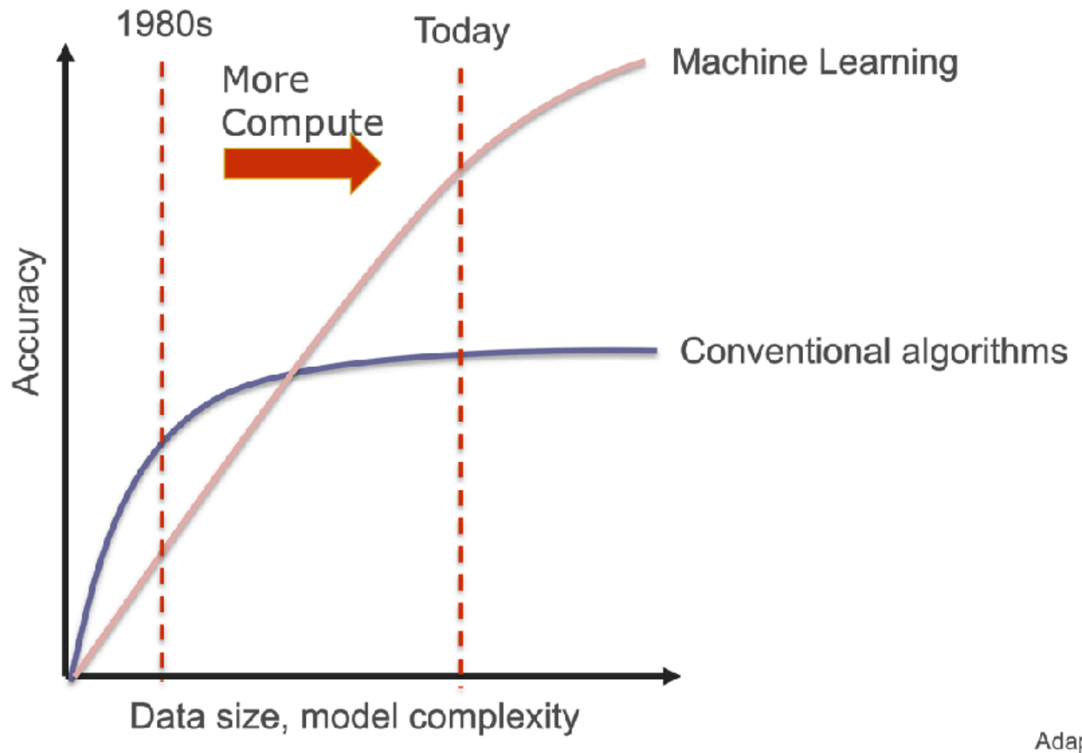
AI and Compute

Two Distinct Eras of Compute Usage in Training AI Systems



AI and Compute

Two Distinct Eras of Compute Usage in Training AI Systems



Adapted from Jeff Dean
HotChips 2017

Machine Learning Systems

Algorithmic Bias

NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

Heidi Ledford

Black people with complex medical needs were less likely than equally ill white people to be referred to programmes that provide more personalized care. Credit: Ed Kashi/VII/Redux/eyevine

An algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, a sweeping analysis has found.

PDF version

RELATED ARTICLES

A fairer way forward for AI in health care

Bias detectives: the researchers striving to make algorithms fair

Can we open the black box of AI?

SUBJECTS

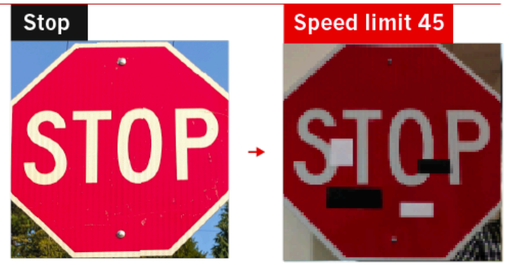
Computer science · Health care · Policy · Society

Fooling AI Systems

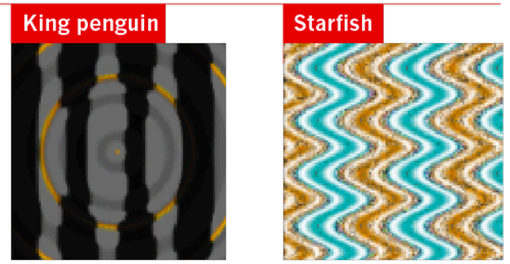
FOOLING THE AI

Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.

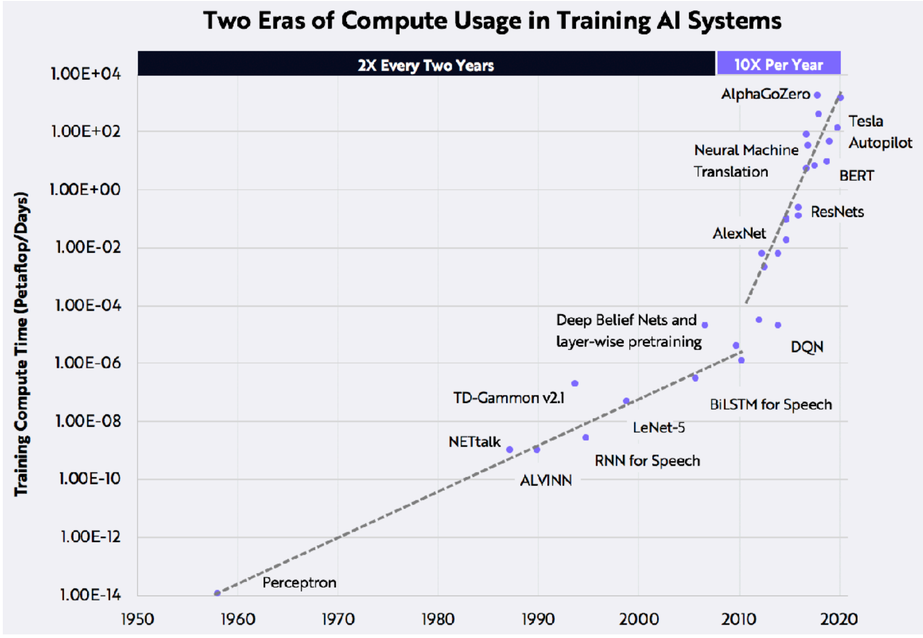


Scientists have evolved images that look like abstract patterns — but which DNNs see as familiar objects.



©nature

AI and Compute



Reading Assignments

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Fairness and Abstraction in Sociotechnical Systems

ANDREW D. SELBST, Data & Society Research Institute
DANAH BOYD, Microsoft Research and
Data & Society Research Institute
SORELLE A. FRIEDLER, Haverford College, PA
SURESH VENKATASUBRAMANIAN, University of Utah
JANET VERTESI, Princeton University

A key goal of the fair-ML community is to develop machine-learning based systems that, once introduced into a social context, can achieve social and legal outcomes such as fairness, justice, and due process. Bedrock concepts in computer science—such as abstraction and modular design—are used to define notions of fairness and discrimination, to produce fairness-aware learning algorithms, and to intervene at different stages of a decision-making pipeline to produce "fair" outcomes. In this paper, however, we contend that these concepts render technical interventions ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems. We outline this mismatch with five "traps" that fair-ML work can fall into even as it attempts to be more context-aware in comparison to traditional data science. We draw on studies of sociotechnical systems in Science and Technology Studies to explain why such traps occur and how to avoid them. Finally, we suggest ways in which technical designers can mitigate the traps through a refocusing of design in terms of process rather than solutions, and by drawing abstraction boundaries to include social actors rather than purely technical ones.

CCS Concepts: • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*;

Additional Key Words and Phrases: Fairness-aware Machine Learning, Sociotechnical Systems, Interdisciplinary

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally "general" broad measures of progress they are set up to be.