Anatomy of an Artificial Intelligence System

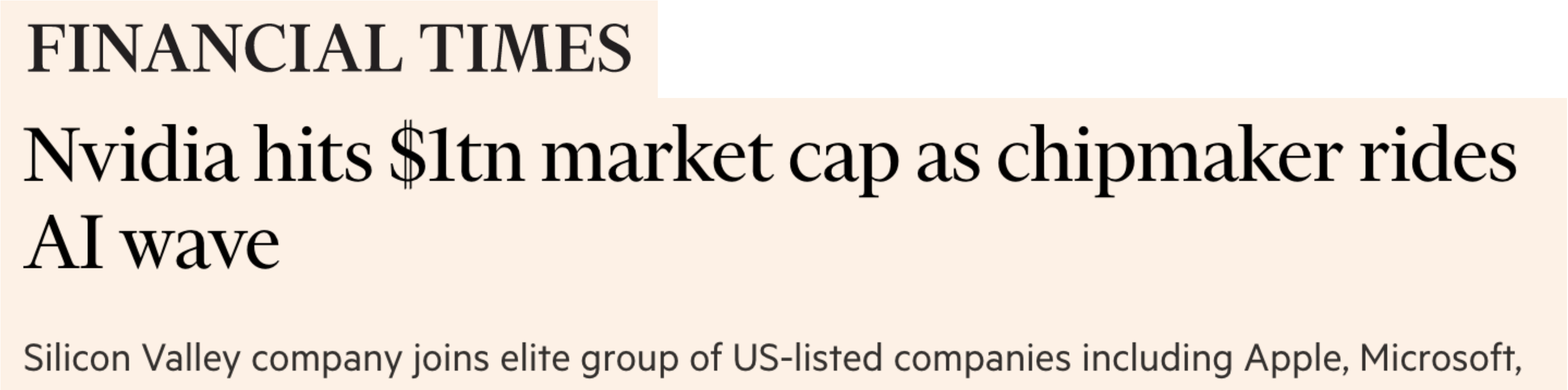
and what it means for policymakers

Sorelle Friedler
Shibulal Family Associate Professor

HAVERFORD
COLLEGE

DEPARTMENT OF COMPUTER SCIENCE

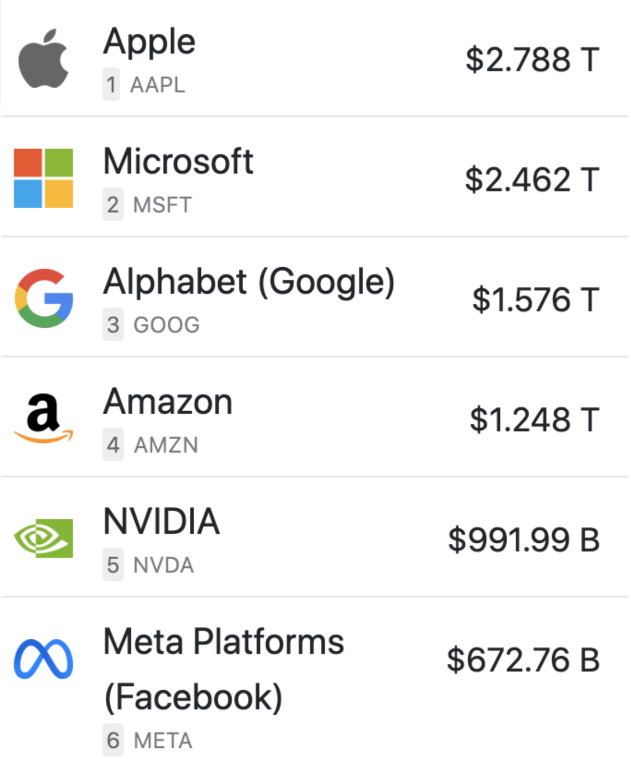
Why is AI important?



FINANCIAL TIMES

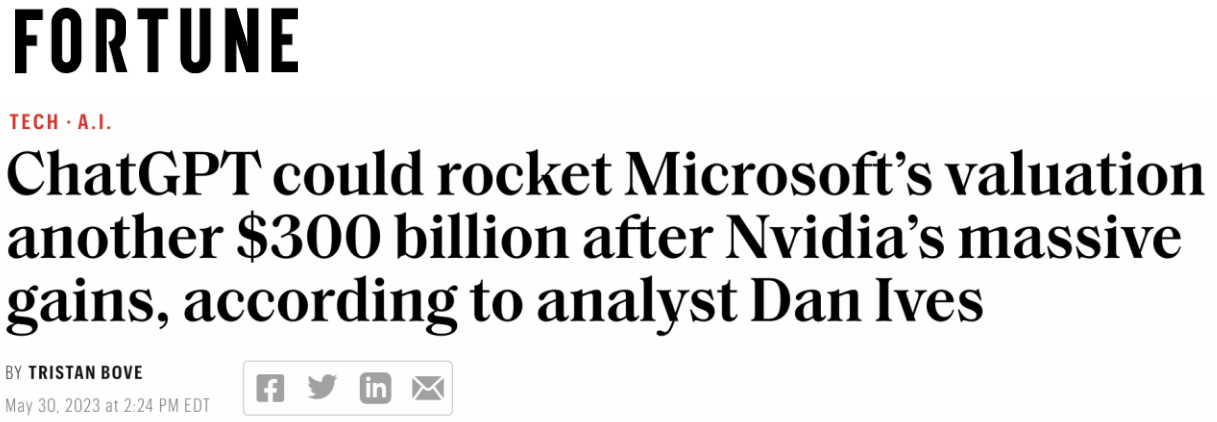
Nvidia hits $1tn market cap as chipmaker rides AI wave

Silicon Valley company joins elite group of US-listed companies including Apple, Microsoft, Amazon and Alphabet



🇺🇸 USA

Apple	1 AAPL	$2.788 T
Microsoft	2 MSFT	$2.462 T
Alphabet (Google)	3 GOOG	$1.576 T
Amazon	4 AMZN	$1.248 T
NVIDIA	5 NVDA	$991.99 B
Meta Platforms (Facebook)	6 META	$672.76 B



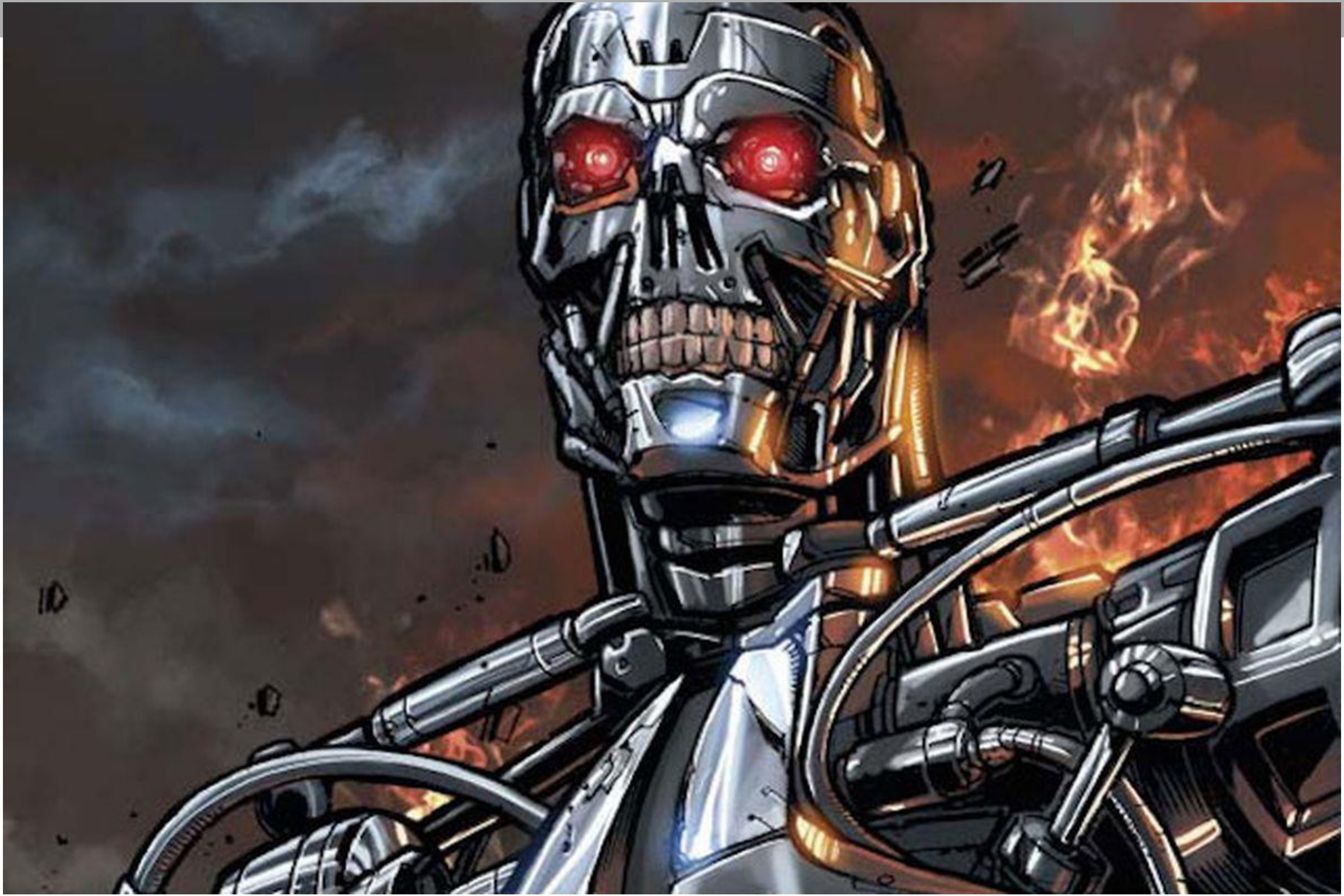
FORTUNE

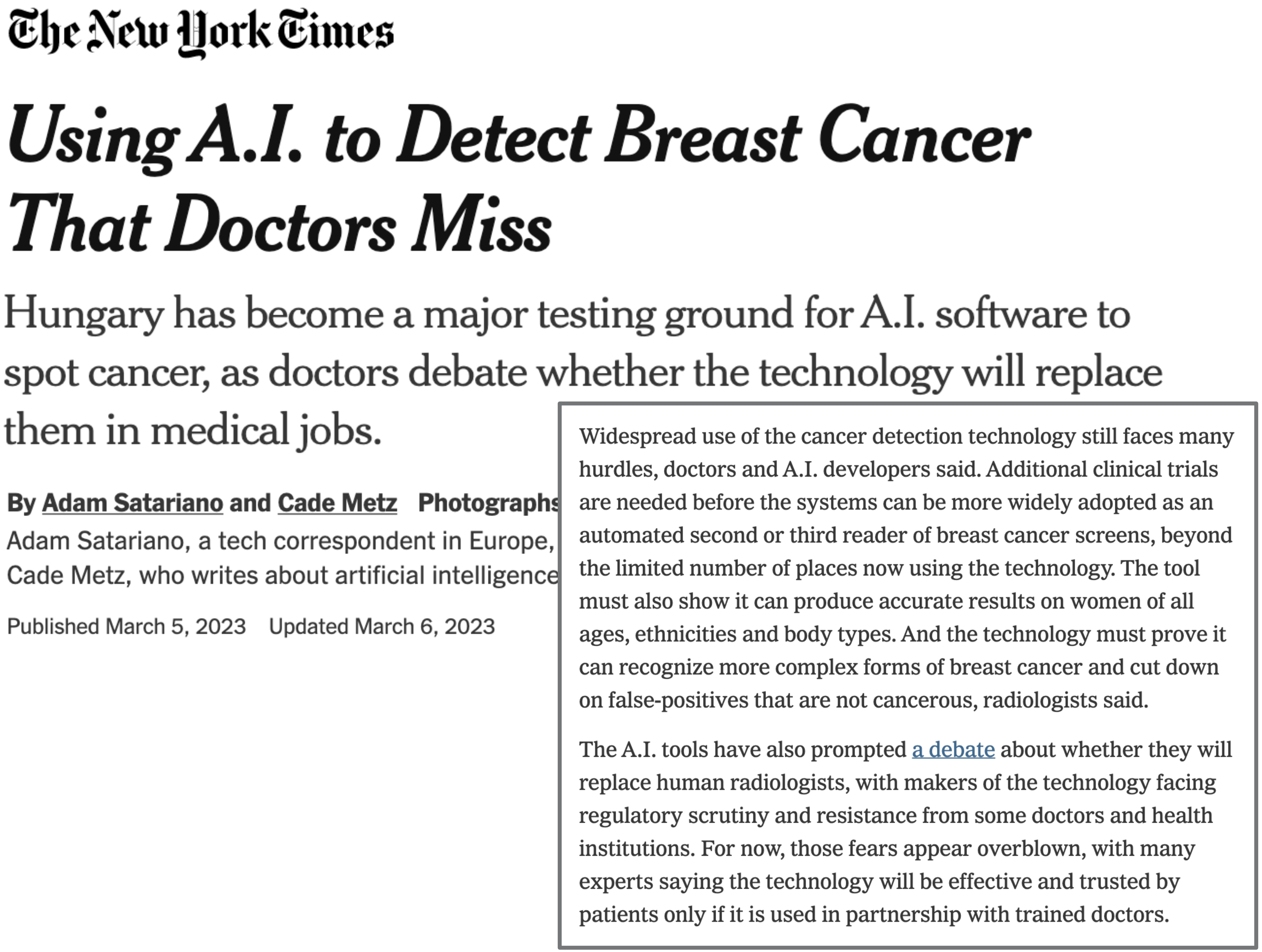
TECH · A.I.

ChatGPT could rocket Microsoft's valuation another $300 billion after Nvidia's massive gains, according to analyst Dan Ives

BY TRISTAN BOVE

May 30, 2023 at 2:24 PM EDT





The New York Times

Using A.I. to Detect Breast Cancer That Doctors Miss

Hungary has become a major testing ground for A.I. software to spot cancer, as doctors debate whether the technology will replace them in medical jobs.

By Adam Satariano and Cade Metz   Photographs

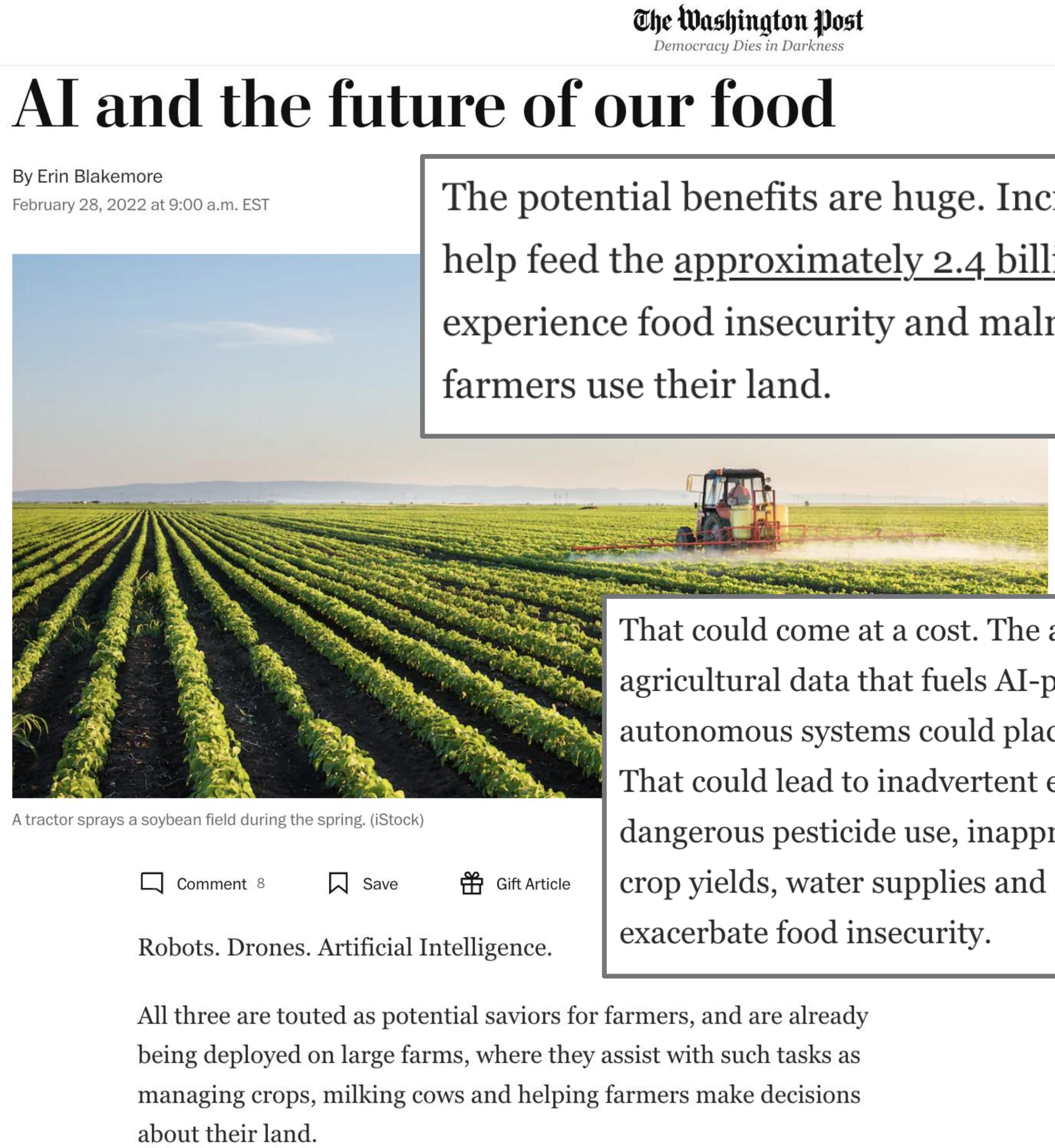
Adam Satariano, a tech correspondent in Europe,
Cade Metz, who writes about artificial intelligence

Published March 5, 2023   Updated March 6, 2023

Widespread use of the cancer detection technology still faces many hurdles, doctors and A.I. developers said. Additional clinical trials are needed before the systems can be more widely adopted as an automated second or third reader of breast cancer screens, beyond the limited number of places now using the technology. The tool must also show it can produce accurate results on women of all ages, ethnicities and body types. And the technology must prove it can recognize more complex forms of breast cancer and cut down on false-positives that are not cancerous, radiologists said.

The A.I. tools have also prompted a debate about whether they will replace human radiologists, with makers of the technology facing regulatory scrutiny and resistance from some doctors and health institutions. For now, those fears appear overblown, with many experts saying the technology will be effective and trusted by patients only if it is used in partnership with trained doctors.

AGRICULTURE



The Washington Post
Democracy Dies in Darkness

AI and the future of our food

By Erin Blakemore
February 28, 2022 at 9:00 a.m. EST

A tractor sprays a soybean field during the spring. (iStock)

Robots. Drones. Artificial Intelligence.

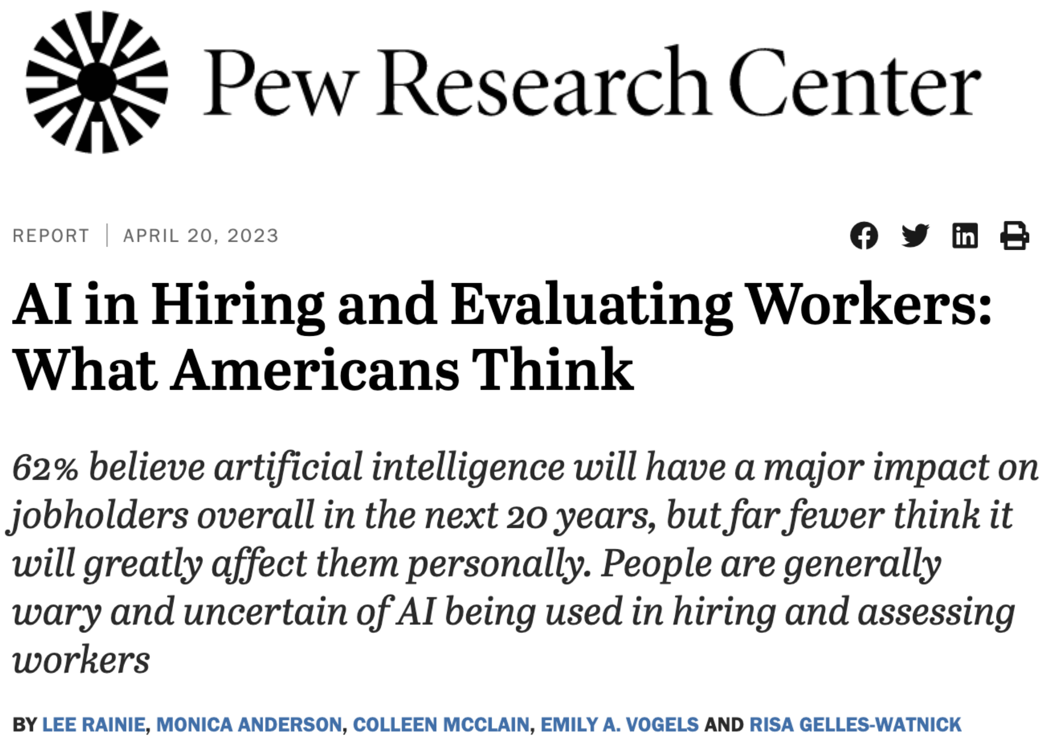
All three are touted as potential saviors for farmers, and are already being deployed on large farms, where they assist with such tasks as managing crops, milking cows and helping farmers make decisions about their land.

> The potential benefits are huge. Increases in farm productivity could help feed the approximately 2.4 billion people around the world who experience food insecurity and malnutrition and revolutionize the way farmers use their land.

> That could come at a cost. The analysis points out potential flaws in the agricultural data that fuels AI-powered systems and the possibility that autonomous systems could place productivity over the environment. That could lead to inadvertent errors causing overfertilization, dangerous pesticide use, inappropriate irrigation or erosion, risking crop yields, water supplies and soil. And wide-scale crop failures could exacerbate food insecurity.

https://www.pewresearch.org/internet/2023/04/20/ai-in-hiring-and-evaluating-workers-what-americans-think/

EMPLOYMENT



Pew Research Center

REPORT | APRIL 20, 2023

AI in Hiring and Evaluating Workers: What Americans Think

62% believe artificial intelligence will have a major impact on jobholders overall in the next 20 years, but far fewer think it will greatly affect them personally. People are generally wary and uncertain of AI being used in hiring and assessing workers

BY LEE RAINIE, MONICA ANDERSON, COLLEEN MCCLAIN, EMILY A. VOGELS AND RISA GELLES-WATNICK

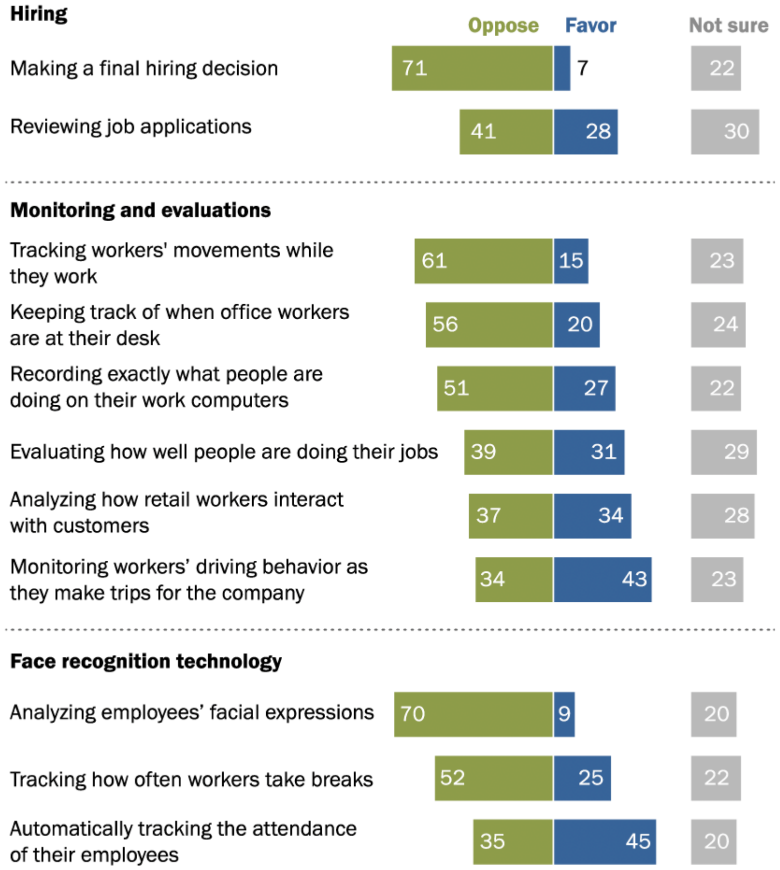
Would you want to apply for a job that uses AI to help make hiring decisions?

% of U.S. adults who say they would or would not want to apply for a job with an employer that uses artificial intelligence to help in hiring decisions

66% say **No** **32%** say **Yes**

Americans widely oppose employers using AI to make final hiring decisions, track workers' movements while they work, and analyze their facial expressions

% of U.S. adults who say they __ employers' use of artificial intelligence for each of the following



Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Dec. 12-18, 2022.
"AI in Hiring and Evaluating Workers: What Americans Think"

PEW RESEARCH CENTER

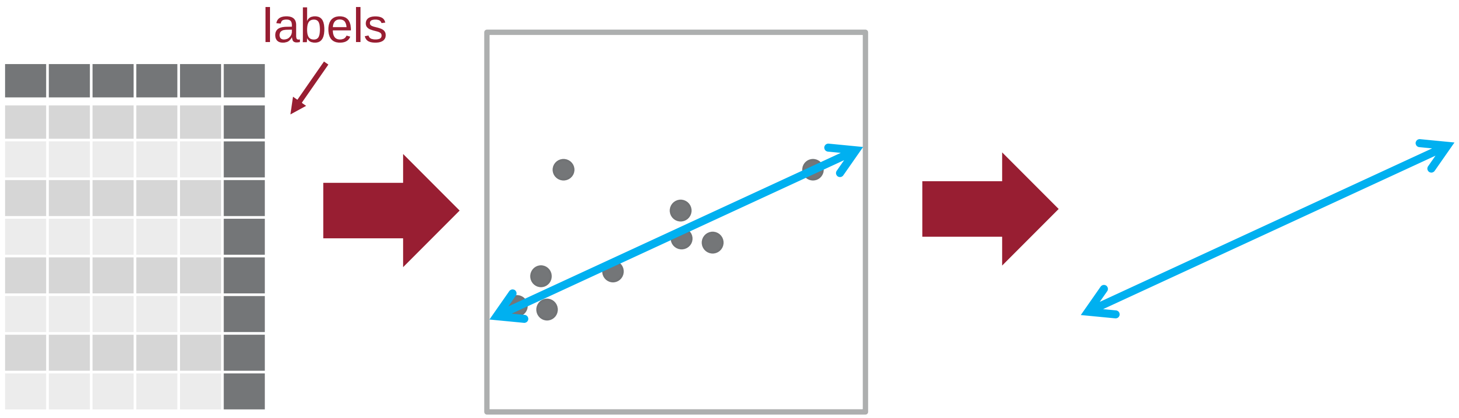
What is AI?

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke

A Basic AI Pipeline

Training



labels

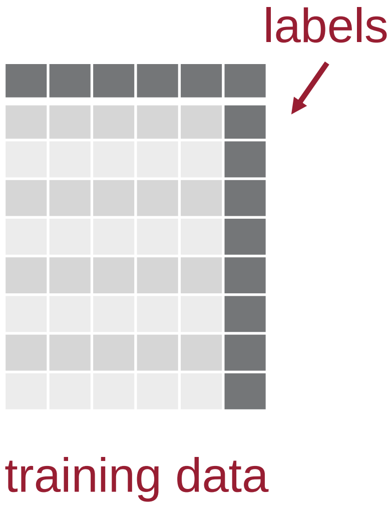
training data

pattern finding algorithm

trained model

A Basic AI Pipeline

Training



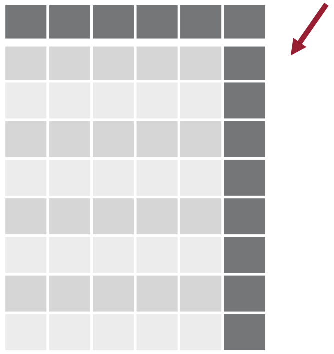
Examples:

- breast cancer scans with radiologist highlighted concerns
- resumes with historical hire / no hire decisions from previous company processes
- text prompts with written responses from specialized contractors

A Basic AI Pipeline

Training

labels

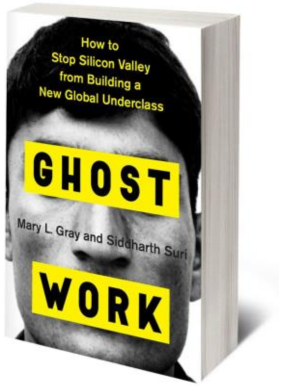


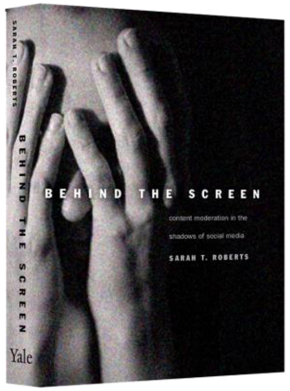
training data

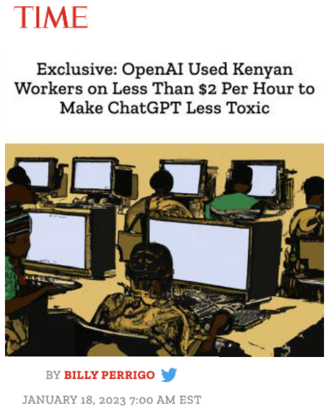
Examples:

- breast cancer scans with radiologist highlighted concerns
- resumes with historical hire / no hire decisions from previous company processes
- text prompts with written responses from specialized contractors

Manual labor from people makes this possible!


How to Stop Silicon Valley from Building a New Global Underclass
GHOST
Mary L. Gray and Siddharth Suri
WORK





TIME

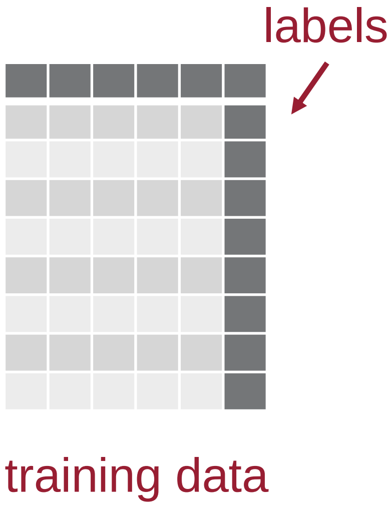
Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

BY BILLY PERRIGO

JANUARY 18, 2023 7:00 AM EST

A Basic AI Pipeline

Training



Examples:

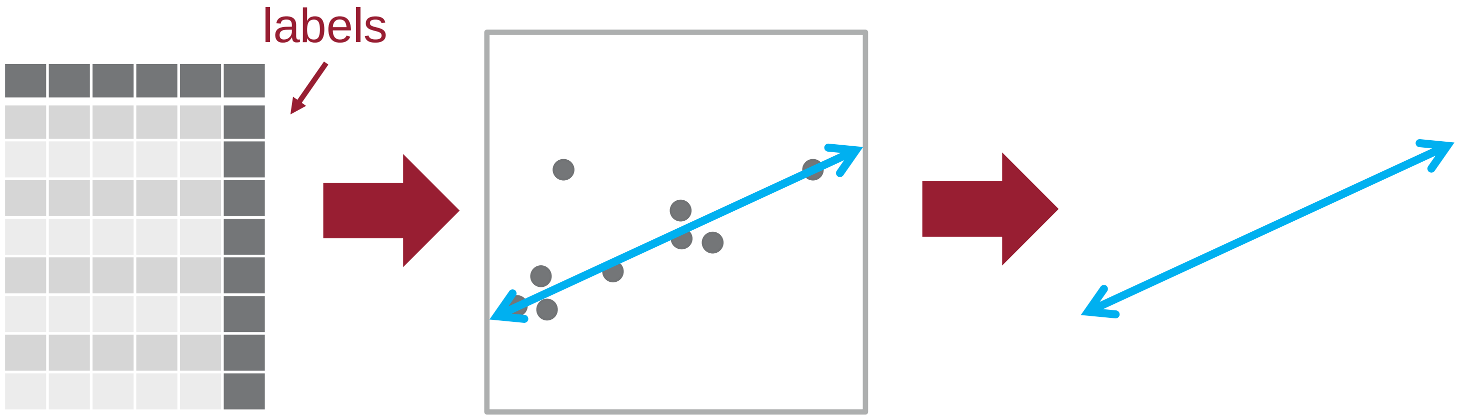
- breast cancer scans with radiologist highlighted concerns
- resumes with historical hire / no hire decisions from previous company processes
- text prompts with written responses from specialized contractors

Data takeaways:

- Requires data that is accurately able to represent the goal – this is **not magic**!
- Uses data collected about people who may have **privacy** concerns with its use.

A Basic AI Pipeline

Training

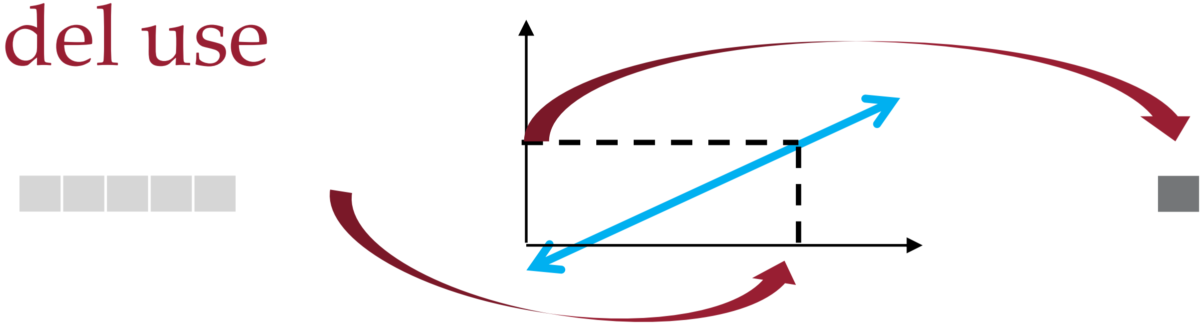


training data

pattern finding algorithm

trained model

Model use



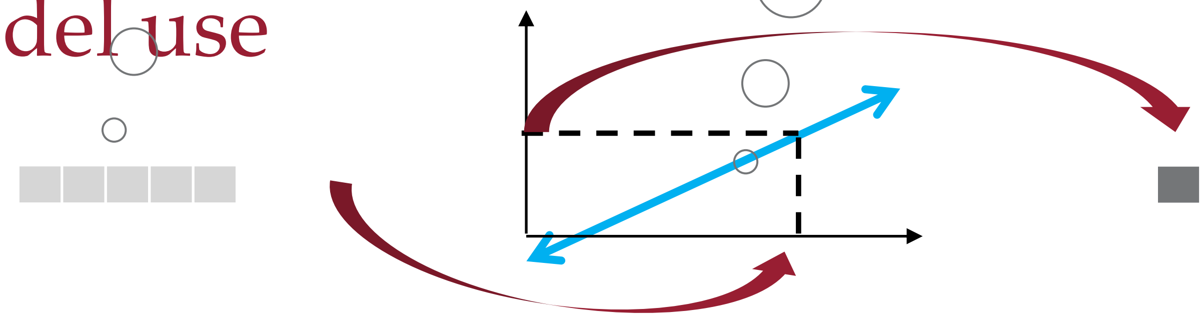
A Basic AI Pipeline

Get a person to fix it!

What if this data doesn't match the original training data well?

Then the input to the model may be far from values it can predict confidently

Model use



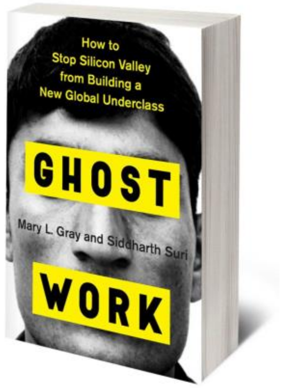
Cars are getting better at driving themselves, but you still can't sit back and nap

December 22, 2021 · 5:00 AM ET

Heard on Morning Edition

Camila Domonoske

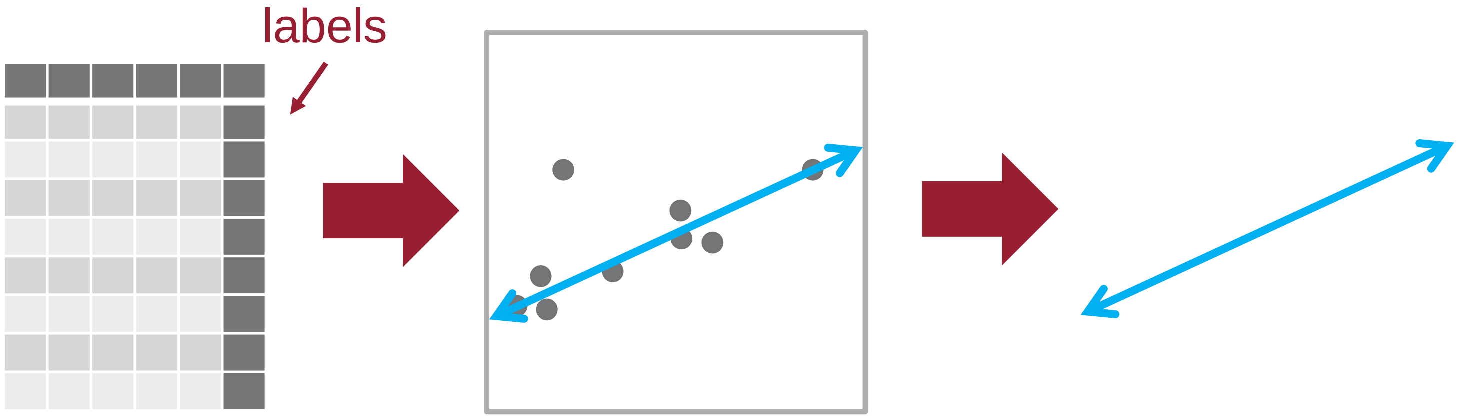
npr



How to Stop Silicon Valley from Building a New Global Underclass
GHOST
Mary L. Gray and Siddharth Suri
WORK

Points of Policy Intervention

Training



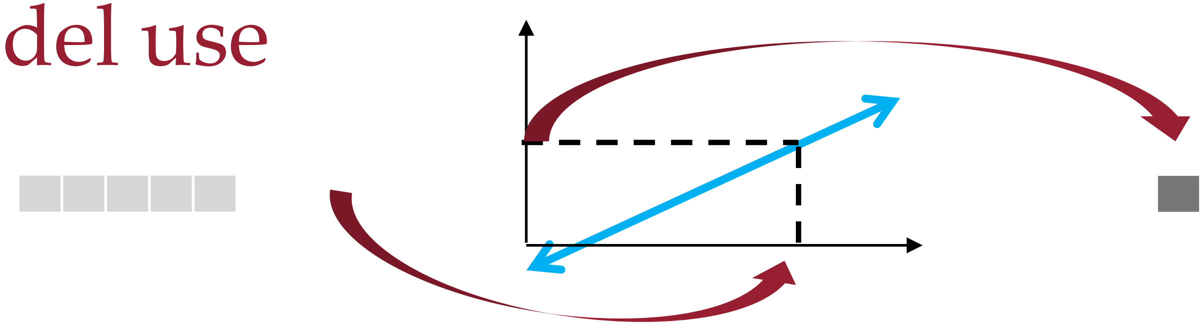
labels

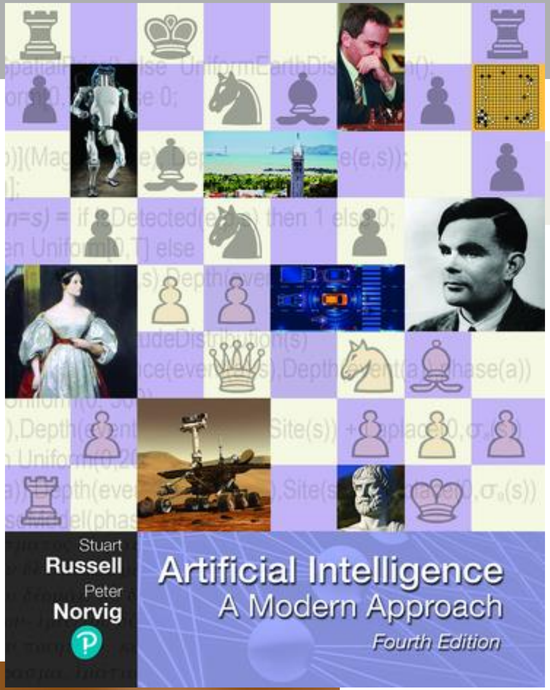
training data

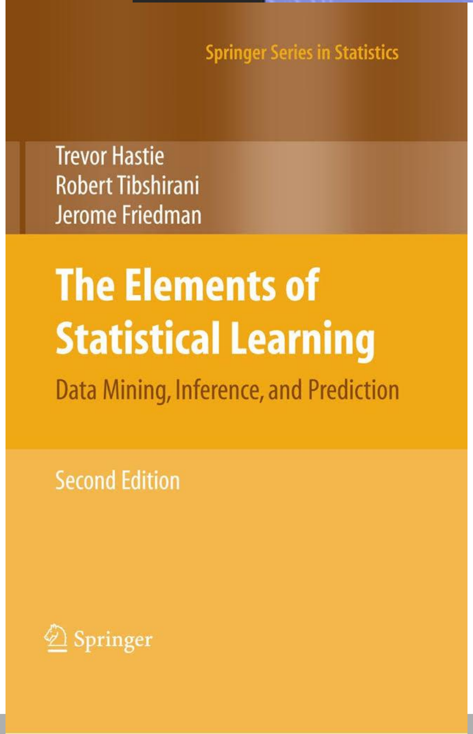
pattern finding algorithm

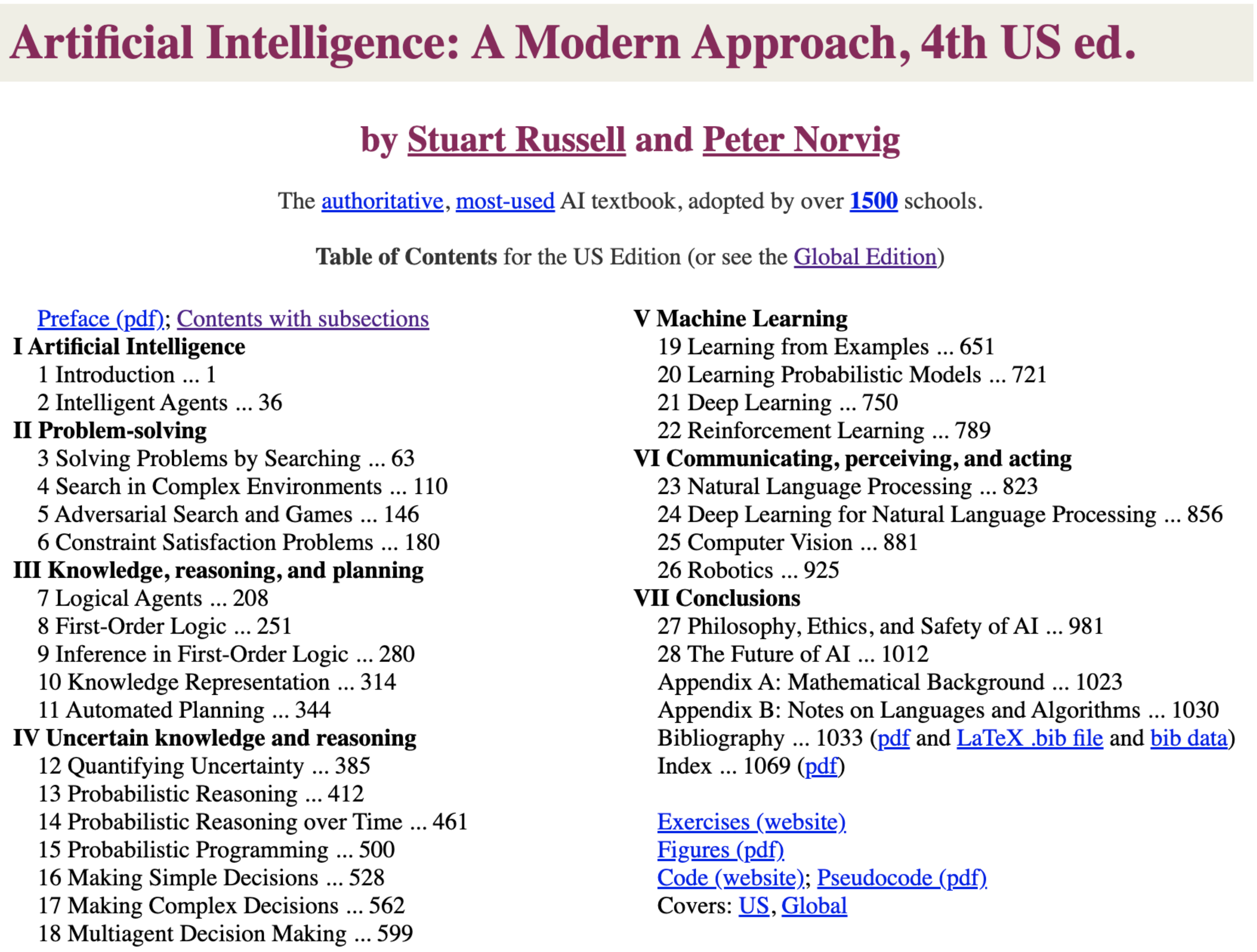
trained model

Model use









Artificial Intelligence: A Modern Approach, 4th US ed.

by Stuart Russell and Peter Norvig

The authoritative, most-used AI textbook, adopted by over 1500 schools.

Table of Contents for the US Edition (or see the Global Edition)

Preface (pdf); Contents with subsections

I Artificial Intelligence

II Problem-solving

III Knowledge, reasoning, and planning

IV Uncertain knowledge and reasoning

V Machine Learning

VI Communicating, perceiving, and acting

VII Conclusions

Exercises (website)
Figures (pdf)
Code (website); Pseudocode (pdf)
Covers: US, Global

Definitions

ARTIFICIAL INTELLIGENCE.—In this section, the term ‘‘artificial intelligence’’ includes the following:

1. Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
2. An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
3. An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
4. A set of techniques, including machine learning, that is designed to approximate a cognitive task.
5. An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting.

2019 NDAA

Definitions

COVERED ALGORITHM.—The term "covered algorithm" means a computational process that uses machine learning, natural language processing, artificial intelligence techniques, or other computational processing techniques of similar or greater complexity and that makes a decision or facilitates human decision-making with respect to covered data, including to determine the provision of products or services or to rank, order, promote, recommend, amplify, or similarly determine the delivery or display of information to an individual.

2022 ADPPA

Definitions

CONSEQUENTIAL DECISION.— “Consequential decision” means a decision or judgment that has a legal, material, or similarly significant effect on an individual’s life relating to the impact of, access to, or the cost, terms, or availability of, any of the following:

(1) Employment, workers management, or self-employment, including, but not limited to, all of the following: (A) Pay or promotion. (B) Hiring or termination. (C) Automated task allocation.

(2) Education and vocational training, including, but not limited to, all of the following: (A) Assessment, including, but not limited to, detecting student cheating or plagiarism. (B) Accreditation. (C) Certification. (D) Admissions. (E) Financial aid or scholarships.

(3) Housing or lodging, including rental or short-term housing or lodging.

(4) Essential utilities, including electricity, heat, water, internet or telecommunications access, or transportation.

(5) Family planning, including adoption services or reproductive services, as well as assessments related to child protective services.

(6) Health care or health insurance, including mental health care, dental, or vision.

(7) Financial services, including a financial service provided by a mortgage company, mortgage broker, or creditor.

(8) The criminal justice system, including, but not limited to, all of the following: (A) Risk assessments for pretrial hearings. (B) Sentencing. (C) Parole.

(9) Legal services, including private arbitration or mediation.

(10) Voting.

(11) Access to benefits or services or assignment of penalties.

2023 CA AB 331

Options

- Sector-specific scoping
  - **Example: "Health and health insurance technologies** such as medical AI systems and devices, AI-assisted diagnostic tools, algorithms or predictive models used to support clinical decision making, medical or insurance health risk assessments, drug addiction risk assessments and associated access algorithms, wearable technologies, wellness apps, insurance care allocation algorithms, and health insurance cost and underwriting algorithms."

    list from: White House AI Bill of Rights: Examples of Automated Systems

- Regulatory refinement
  - Identify "consequential decisions" and staff a state agency to update a list of covered algorithms in those areas.

How can policymakers intervene?

Sector-specific Approaches

Narrow and specific red lines

Examples

- Senate: Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023
- Ban on affective AI in law enforcement

https://www.brookings.edu/blog/techtank/2021/08/04/why-president-biden-should-ban-affective-computing-in-federal-law-enforcement/

Sector-specific approaches

Example: employment

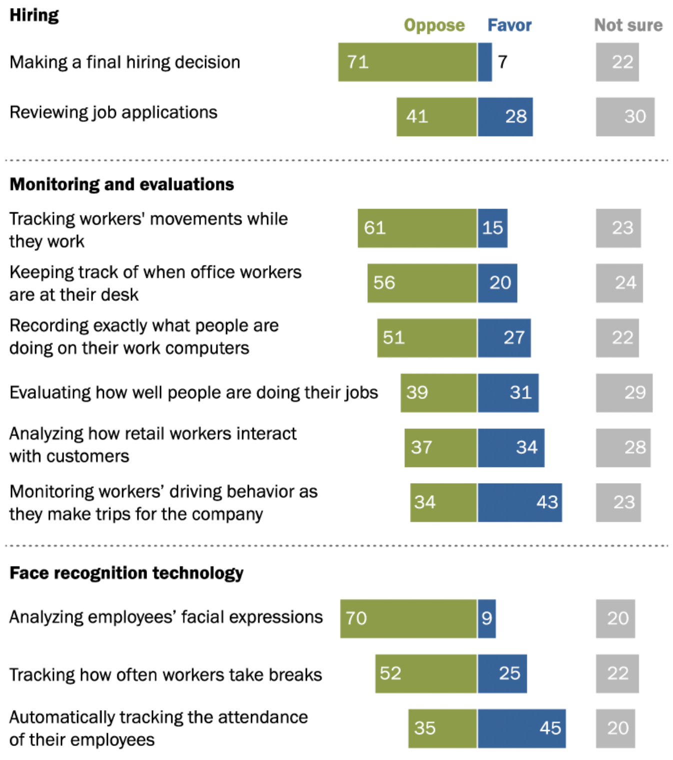
- Americans don’t want employers to track movements or facial expressions
- Americans want to know that a final hiring decision is made by a person

Options:

- Define a list of employment-specific algorithms
- Set out principles / goals
- Have the state Department of Labor issue guidance on meeting these principles

Americans widely oppose employers using AI to make final hiring decisions, track workers’ movements while they work, and analyze their facial expressions

% of U.S. adults who say they __ employers’ use of artificial intelligence for each of the following



Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Dec. 12-18, 2022.
“AI in Hiring and Evaluating Workers: What Americans Think”

PEW RESEARCH CENTER

Preemptive requirements

Example: employment

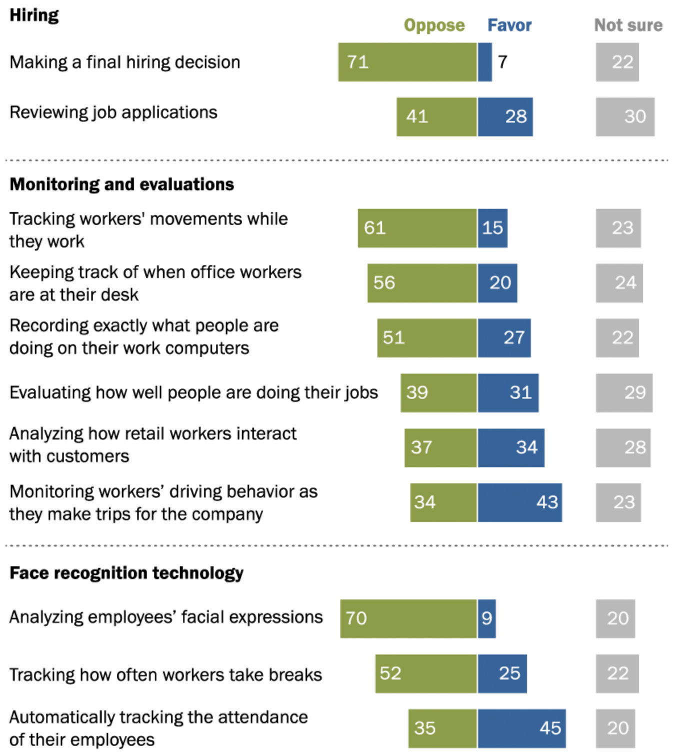
- Americans don’t want employers to track movements or facial expressions
- Americans want to know that a final hiring decision is made by a person

Options:

- Define a list of employment-specific algorithms
- Set out principles / goals
- Have the state Department of Labor issue guidance on meeting these principles
- Require that this guidance is met *before* any such system can be used in the state

Americans widely oppose employers using AI to make final hiring decisions, track workers’ movements while they work, and analyze their facial expressions

% of U.S. adults who say they __ employers’ use of artificial intelligence for each of the following



Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Dec. 12-18, 2022.
“AI in Hiring and Evaluating Workers: What Americans Think”

PEW RESEARCH CENTER

Cross-cutting Approaches

Safety and Efficacy

- Why? Examples:
  - predictive policing technology that incorrectly repeatedly sends police back to where they've been before
  - gunshot detectors that incorrectly alert and send police into neighborhoods incorrectly and dangerously on alert
  - model to predict sepsis that underperforms and causes alert fatigue
  - AI evaluation of delivery drivers' road safety incorrectly cost them a bonus

Safety and Efficacy

- Preemptive and ongoing requirements
  - Sector-specific and/or regulations from a Tech-focused agency
    - e.g., requirements that policing technology be shown to work
  - Set up a mechanism where concentrated technical talent can work with sector-specific agencies

- Liability
  - AI decisions / outputs are not covered by Section 230

Prohibit Algorithmic Discrimination

- Why? Examples:
  - Loan underwriting and pricing model charged **HBCU alums** more
  - Hiring tool rejected applicants with “**women’s**” on their resume
  - Statements “I’m **gay**” and “I’m a **Jew**” were marked as toxic
  - Remote exam proctoring systems incorrectly marked **disabled students** as cheating
  - Healthcare risk assessment incorrectly marked **Black patients** as needing less care

Prohibit Algorithmic Discrimination

- Definition:
  - The term “algorithmic discrimination” refers to instances when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their actual or perceived race, color, ethnicity, sex (including based on pregnancy, childbirth, and related conditions; gender identity; intersex status; and sexual orientation), religion, age, national origin, limited English proficiency, disability, veteran status, genetic information, or any other classification protected by law. EO 14091

Prohibit Algorithmic Discrimination

- How:
  - Private right of action (e.g.,: CA AB 331)
  - Sector-specific requirements and oversight
  - Impact assessments

Impact Assessments

- Why?
  - Safety and Efficacy Protections
  - Algorithmic Discrimination Tests
  - Transparency
  - Oversight and Accountability

Impact Assessments

- What:
  - Detailed, specific questions about the assessment process and results of an algorithmic system
  - Important: public consultation component
  - Example: Algorithmic Accountability Act of 2022

- How:
  - pre-release and ongoing
  - kept in private company records versus submitted to a state agency

Transparency

- Impact assessments
- Notice – to people impacted *before* use
- Explanation – how and why was a decision made
  - such adverse action notices already required for financial decisions

- Environmental impact (kWh)
  - targeted requirement to report on the kWh used for AI

Data-focused Interventions

- Data Privacy Protections
  - Data minimization
  - See, e.g.,: American Data Privacy and Protection Act of 2022 (ADPPA)

- Intellectual Property Protections
  - E.g., permission / contract required to use a song as part of training data

Labor

- Ensuring safety and efficacy
  - Require human review for consequential decision systems
- Providing human alternatives
  - Allow people to opt-out and use a provided human alternative
- Protecting jobs
  - Require that AI augments, not replaces, the existing workforce

Recommendations

- Don’t set up a task force! Pick something specific instead.
  - workplace surveillance limits, ban affective AI for law enforcement, etc
- Focus on impacts, not technical details
  - craft AI definitions that are limited based on impact
- Make use of the sector-specific expertise in state agencies and add (shared) technical expertise as necessary
  - sector-specific regulation can be owned by the relevant existing agency
- Be specific when crafting transparency requirements
  - asking specific questions can lead the evaluations you want done to happen

Resources

- White House AI Bill of Rights
  - www.whitehouse.gov/ostp/ai-bill-of-rights
  - “What should be expected” sections include specific actionable safeguards
  - Appendix includes examples of consequential automated systems
- American Data Privacy and Protection Act (2022)
  - bipartisan enforcement framework
- Algorithmic Accountability Act (2022)
  - useful list of specific questions to ask
- CA AB 331 Automated Decision Tools (2023)
  - consequential decision definition including specific domains

Thanks!

sorelle@cs.haverford.edu

Sorelle Friedler, Haverford College