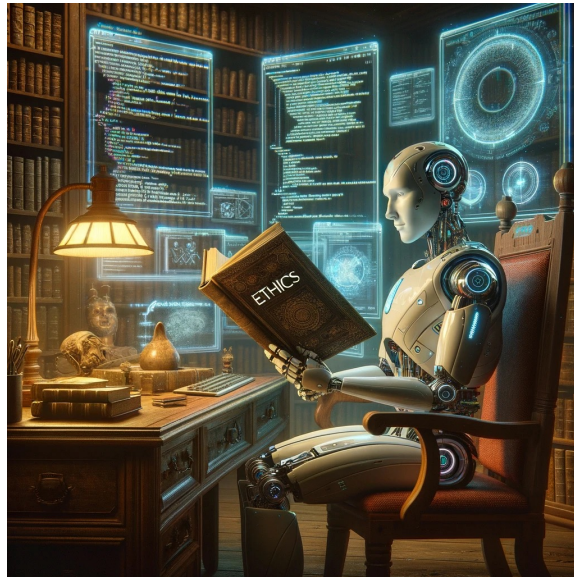


# AI: Representation and Problem Solving

## AI and Ethics



Instructors: Nihar Shah and Tuomas Sandholm

Slide credits: CMU AI with some drawings from [ai.berkeley.edu](http://ai.berkeley.edu)

## AI is in a fairly unique position with respect to ethics

- Not only must AI developers behave ethically
- The AI systems, themselves, must behave ethically



**WARNING**

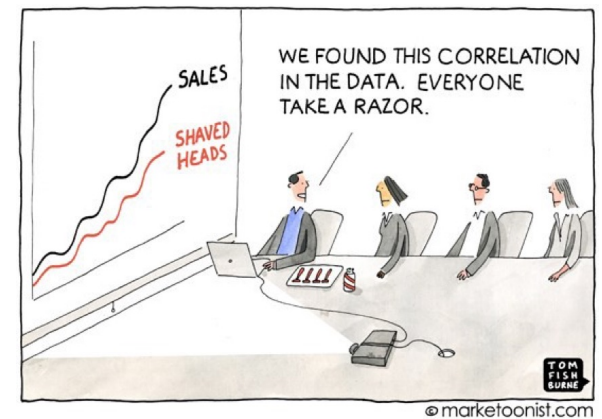
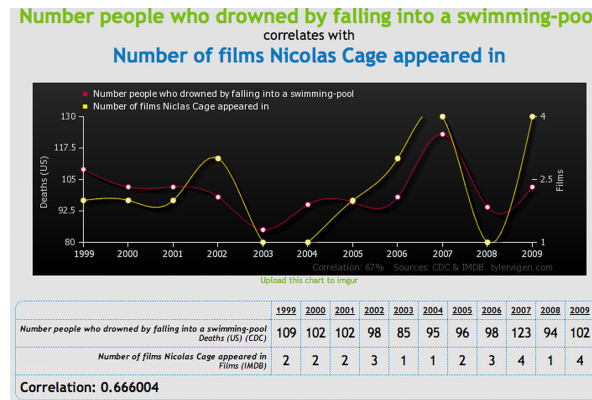
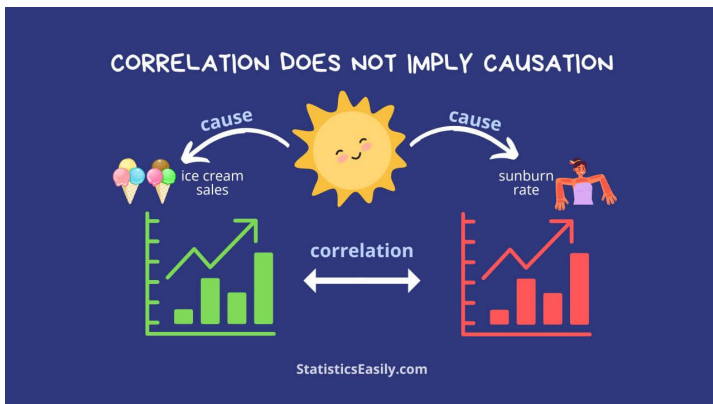
Adobe Stock | #569060502



**(Approximately) True Story**

# Causality

## Classical: Correlation is not causation



## Prediction is not causation

<https://statisticseasily.com/correlation-vs-causality/>

Any intervention changes the system

As in our story...

Past data may not reflect the future

# Any intervention can also change people's behavior

WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

SIGN IN

SUBSCRIBE

ARIELLE PARDES

BUSINESS FEB 16, 2022 8:00 AM

## How Job Applicants Try to Hack Résumé-Reading Software


Recruiters increasingly use machines to screen applicants. So crafty hopefuls are devising tactics to outwit the machines. But is it worth it?

## How to get your resume past Artificial Intelligence (AI) screening tools: 5 tips

Most resumes must pass through an automated test – based on AI or RPA – before humans ever see them. Career experts share tips, from language to formatting, on how to build a resume that clears that first AI screening tool hurdle

By [Kevin Casey](#)

March 4, 2021

 160 readers like this 

# Explanability/interpretability

What is the AI doing? How does it really work?

Why is it giving the output it is giving?



- Are “explanations” faithful to the AI’s actual working?
- Do explanations make humans complacent?



# Bias/(Un)fairness



REUTERS®

World ▾

Business ▾

Markets ▾

Sustainability ▾

More ▾



World

## Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 10, 2018 8:50 PM EDT · Updated 5 years ago



<https://www.reuters.com/article/idUSKCN1MK0AG/>

One cause (although not the only one): **Data used to create the AI was biased**

ARE EMILY AND GREG MORE EMPLOYABLE  
THAN LAKISHA AND JAMAL?  
A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION

Marianne Bertrand  
Sendhil Mullainathan

[https://www.nber.org/system/files/working\\_papers/w9873/w9873.pdf](https://www.nber.org/system/files/working_papers/w9873/w9873.pdf)

**AI mimics the same patterns**

Technology

## Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use

Asian and African American people were up to 100 times more likely to be misidentified than white men, depending on the particular algorithm and type of search. Native Americans had the highest false-positive rate of all ethnicities, according to the study, which found that systems varied widely in their accuracy.

The faces of African American women were falsely identified more often in the kinds of searches used by police investigators where an image is compared to thousands or millions of others in hopes of identifying a suspect.

Algorithms developed in the United States also showed high error rates for “one-to-one” searches of Asians, African Americans, Native Americans and Pacific Islanders. Such searches are critical to functions including cellphone sign-ons and airport boarding schemes, and errors could make it easier for impostors to gain access to those systems.

<https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/>

# Privacy



<https://cybernews.com/security/chatgpt-samsung-leak-explained-lessons/>

# Privacy

What are some tradeoffs we make when giving AI our private information?

Do you have a limit for the type or amount of data you are willing to provide?

How can we make the data usage of AI systems more explicit?

Are there ways that we can still enjoy the benefits of AI systems without giving up our private information?

# Ethical Dimensions of Using Artificial Intelligence in Health Care

Michael J. Rigby

[Citation](#) [PDF](#) [Altmetric](#)

An artificially intelligent computer program can now diagnose skin cancer more accurately than a board-certified dermatologist.<sup>1</sup> Better yet, the program can do it faster and more efficiently, requiring a training data set rather than a decade of expensive and labor-intensive medical education. While it might appear that it is only a matter of time before physicians are rendered obsolete by this type of technology, a closer look at the role this technology can play in the delivery of health care is warranted to appreciate its current strengths, limitations, and ethical complexities.



# WHO outlines principles for ethics in health AI

*They include protecting autonomy and ensuring equity*

By [Nicole Wetsman](#) | Jun 30, 2021, 12:50pm EDT

[f](#) [t](#) [SHARE](#)



# Right to be forgotten?

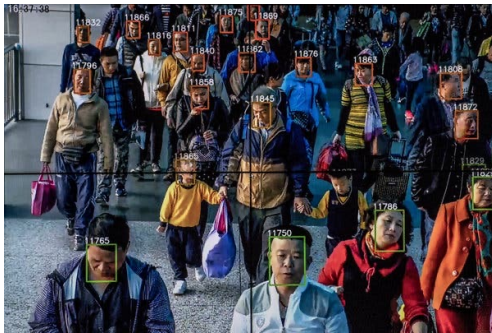
LinkedIn or Facebook can delete a post

Google can delete a search result

But how to make the AI un-learn?

# Surveillance

Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras



## *Facial Recognition: Dawn of Dystopia, or Just the New Fingerprint?*



### Huawei tested AI software that could recognize Uighur minorities and alert police, report says

An internal report claims the face-scanning system could trigger a 'Uighur alarm,' sparking concerns that the software could help fuel China's crackdown on the mostly Muslim minority group



## *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.





# Surveillance



NEWS & ANALYSIS

## EU proposes new AI rules that would ban China-style 'social scoring'

April 22, 2021 | 5 Min Read | Marek Grzegorzcyk

The New York Times

This month, [Amazon](#), [Microsoft](#) and [IBM](#) announced they would stop or [pause](#) their facial recognition offerings for law enforcement. The gestures were largely symbolic, given that the companies are not big players in the industry. The technology police departments use is supplied by companies that aren't household names, such as Vigilant Solutions, Cognitec, NEC, Rank One Computing and [Clearview AI](#).

# Liability

- How do we determine what went wrong (or right) in deployed AI systems?
- Which techniques that we learned about have explainable results?
- Who is responsible for “mistakes” that AI systems make?
- What if someone or another AI misuses AI technology? Who is responsible?

Think about:

Self-driving cars

Healthcare decision-making

Air traffic control

<https://cms.law/en/gbr/publication/artificial-intelligence-who-is-liable-when-ai-fails-to-perform>

# Employment

What types of jobs should AI do instead of people?

What will people do if their jobs are taken?

How can AI systems improve or change people's jobs even if it doesn't take the jobs?

<https://www.nytimes.com/2023/08/24/upshot/artificial-intelligence-jobs.html>

<https://www.bloomberg.com/news/articles/2024-02-08/ai-is-driving-more-layoffs-than-companies-want-to-admit>

The New York Times

Artificial Intelligence > | A.I. Faces Quiz | How the A.I. Race Began | Key Figures in the Field | One Year of ChatGPT

**TheUpshot**

## *In Reversal Because of A.I., Office Jobs Are Now More at Risk*

Technology disruption typically affected blue-collar occupations. Now white-collar workers may feel the brunt of changes.

**Bloomberg**

**Global Layoffs:** Brutal Tech Cuts | Anxious US Workers | RIP the Cushy Tech Job? | Le

Work Shift | Technology & Skills

## **AI Is Driving More Layoffs Than Companies Want to Admit**

# Misinformation

VOA ABOUT LEARNING ENGLISH BEGINNING LEVEL INTERMEDIATE LEVEL ADVANCED LEVEL US HISTORY VIDEO

SCIENCE & TECHNOLOGY 'Deepfakes' of Putin and Zelenskyy Appear Online During Ukraine War  
March 23, 2022

Share  
f t w i e  
See comments  
Print



World / Asia

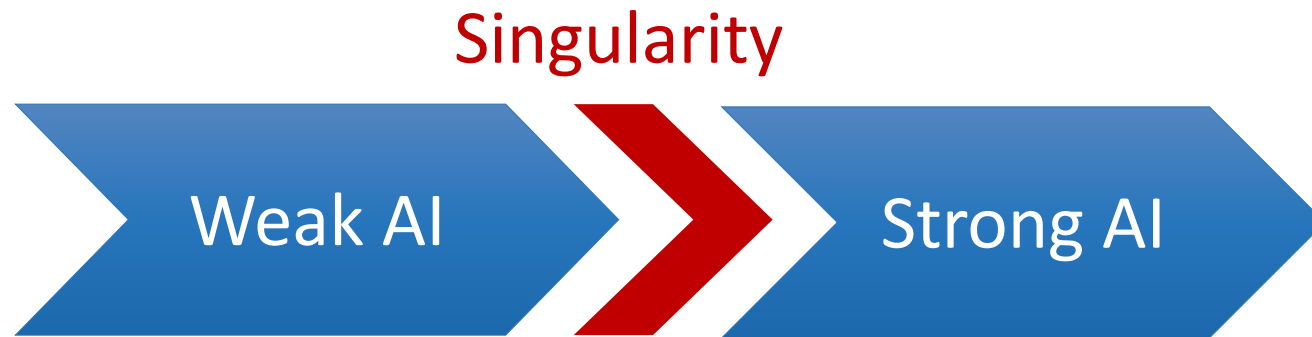
## Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'

By Heather Chen and [Kathleen Magramo](#), CNN

🕒 2 minute read

Published 2:31 AM EST, Sun February 4, 2024

# Should we worry about future A.I.?



- Narrow AI
- Limited number of applications

- Artificial General Intelligence (AGI)
- Recursive self-improvement
- Beyond human control

# AI Weapons/Safety

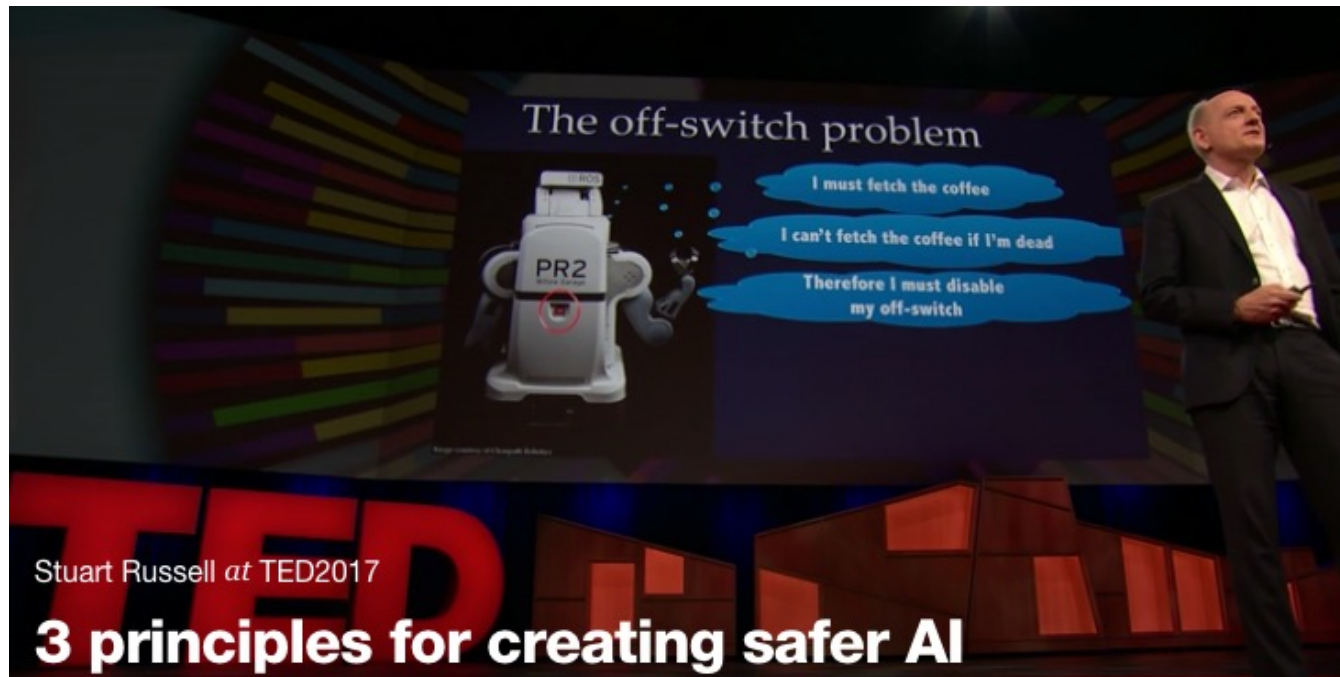


<https://www.youtube.com/watch?v=9CO6M2HsoIA>

# Value (Mis)alignment

**Are the AI's goals same as human goals/preferences?**

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively... we had better be quite sure that the purpose put into the machine is the purpose which we really desire." – Norbert Wiener



[https://www.ted.com/talks/stuart\\_russell\\_how\\_ai\\_might\\_make\\_us\\_better\\_people](https://www.ted.com/talks/stuart_russell_how_ai_might_make_us_better_people)



# AI and Ethics: Careful Thought in AI Research



“In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes.”

