CMSC 373 Artificial Intelligence Fall 2025 17-EthicalAI

Deepak Kumar
Bryn Mawr College

1

Comparing ML to Human Learning

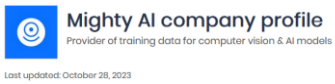
- ConvNet Learning is not human like
- ConvNet Learning uses *supervised learning* of fixed set of categories.
- ConvNet Learning requires a lot of human effort to collect, curate, and label data.
- ConvNet architecture goes through a trial-and-error process of hyperparameter tuning. Akin to “alchemy”, “network whispering”, art form, etc.
- Children learn open-ended tasks,
  Can recognize most objects after seeing a handful of instances
  Ask questiond, deman information, explore, etc.

2

2

The Perils of Big Data

- Acquired from a large amount of human effort
- Also, free online services ("you are the data")
- Self-driving cars: data from mounted cameras on dashboards
  Tesla's EULA to allow users to "share" all driving data
  Uses an offshore outsourcing center (e.g. MightyAI)

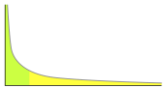


- *Supervised Learning* is not a viable path to general AI.

3

3

The Long Tail



- For ML, the vast range of unexpected situations cannot be captured by a supervised training dataset.
- Proposal to use a combination of *supervised* and *unsupervised* learning. However, not many successful methods exist for unsupervised learning.

  "Unsupervised learning is the dark matter of AI"
  -: Yann LeCun
- Humans have a fundamental competence that is lacking in all AI systems: *commonsense.*

4

4

What does a ConvNet Learn?

- Animal versus No Animal classification task example.
- ConvNets overfit on training data.

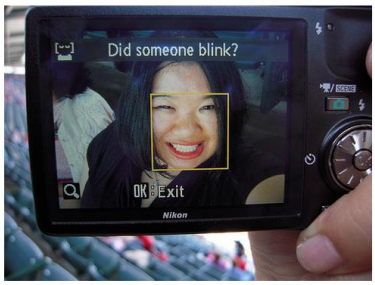
Biased AI

- Bias in Google's photo tagging feature in Photos app
- Also, Nikon…

Nikon Camera Says Asians: People Are Always Blinking

Gwen Sharp, PhD on May 29, 2009

Toban B., Elisabeth, and Mark sent us a link to a post at jozjozjoz about the Nikon S630 digital camera. As Joz explains, "As I was taking pictures of my family, it kept asking 'Did someone blink?' even though our eyes were always open."



Apparently the camera perceives "Asian" eyes as closed.

THE WALL STREET JOURNAL.

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr Follow

Updated July 1, 2015 3:41 pm ET

Share Resize

Google is a leader in artificial intelligence and machine learning. But the company's computers still have a lot to learn, judging by a major blunder by its Photos app this week.

Biased AI

- Datasets for training are biased
  Face recognition dataset was 77.5% male, 83.5% white

  Leads to biases against racial groups, propagates stereotypes, civil rights violations, etc.

  Fairness and transparency in AI is a huge area nowadays.


ACLU About

NEWS & COMMENTARY

Why Amazon's Automated Hiring Tool Discriminated Against Women

Rachel Goodman, Former Staff Attorney, ACLU Racial Justice Program

Share This Page

October 12, 2018

In 2014, a team of engineers at Amazon began working on a project to automate hiring at their company. Their task was to build an algorithm that could review resumes and determine which applicants Amazon should bring on board. But, according to a Reuters report this week, the project was canned just a year later, when it became clear that the tool systematically discriminated against women applying for technical jobs, such as software engineer positions.

7

7

Explanable AI

- Transparent AI
- Interpretable Machine Learning
- A Deep Learning system cannot yet successfully explain itself in human terms. Open area of research in AI.

8

8

Trustworthy AI

- If ML systems are being deployed to make decisions affecting our lives what assurances do we have that machines creating our newsfeed, diagnosing diseases, evaluating loan applications, recommending length of prison sentence, etc. are trustworthy?
- There are a lot of downsides to using AI in society.

9

9

Beneficial AI

- Many useful applications of AI

  Speech transcription
  GPS Navigation
  Trip Planning
  Email Spam Filters
  Language Translation
  Credit Fraud Alerts
  Book/Music Recommendations
  Protection against computer viruses
  Optimizing energy usage in buildings
  Creative media and arts
  Intelligent Tutoring Systems
  Data Analytics
  Phone Computer Vision Apps for Visually Impaired
  Plant/Bird Recognition Apps
  Realtime speech transcription/translation
  *Numerous uses of LLMs
  etc.

10

10

The Great AI Trade-Off

- Should we embrace abilities of AI systems that can improve our lives, help save lives?
- Should we be more cautious, given AI's unpredictable errors, susceptibility to bias, vulnerability to hacking, lack of transparency in decision making, etc.?
- What is required of an AI system in order to trust it enough to let it operate autonomously?
- Pew Research, 2018:

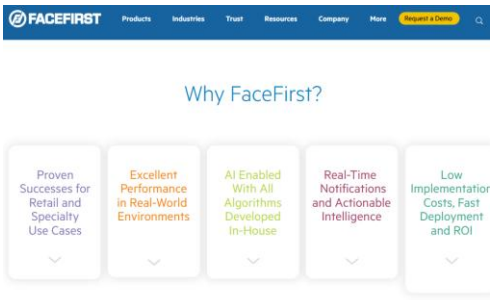
  63% AI scientists predicted that progress in AI would leave us better off by 2030.
  37% disagreed.

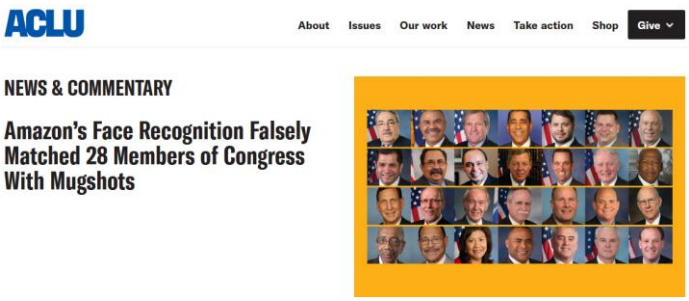
11

11

Ethics of Face Recognition

- Issues of privacy
- Violation of Civil Rights





FaceFirst: Your Fast, Accurate, Ethical Face Matching Platform

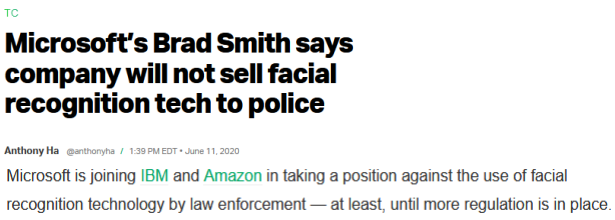
FaceFirst is a global leader in highly effective face matching systems for retailers, hospitals, casinos, airports, stadiums, and arenas. FaceFirst's software leverages artificial intelligence and human oversight to prevent violence, theft, and fraud. With FaceFirst, you can provide safer environments for your valued customers, patients, guests, employees, and associates. We design our patented video analytics platform to be scalable, fast, accurate, and ethical while maintaining high levels of security, privacy, and accountability. FaceFirst is based in Austin, Texas.

12

12

Calls for Regulation

- Problems surrounding AI- trustworthiness, explanability, bias, vulnerability to attack, and morality of use- are social and political issues as much as they are technical ones.

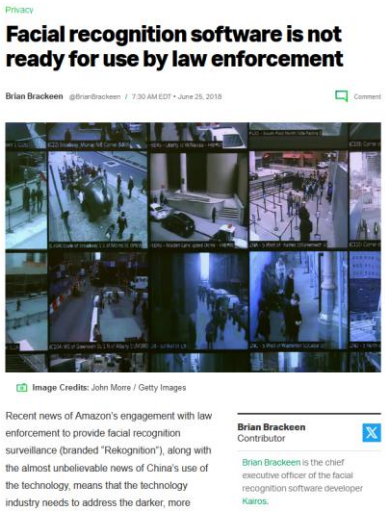


TC

Microsoft's Brad Smith says company will not sell facial recognition tech to police

Anthony Ha @anthonyha / 1:39 PM EDT • June 11, 2020

Microsoft is joining IBM and Amazon in taking a position against the use of facial recognition technology by law enforcement — at least, until more regulation is in place.



Privacy

Facial recognition software is not ready for use by law enforcement

Brian Brackeen @BrianBrackeen / 7:30 AM EDT • June 25, 2018

Comment

Image Credits: John Moore / Getty Images

Recent news of Amazon's engagement with law enforcement to provide facial recognition surveillance (branded "Rekognition"), along with the almost unbelievable news of China's use of the technology, means that the technology industry needs to address the darker, more

Brian Brackeen
Contributor

Brian Brackeen is the chief executive officer of the facial recognition software developer Kairos.

13

Biden-Harris Executive Order on AI (October 30, 2023)

- New standards for AI Safety & Security
- Protecting American's privacy
- Advancing Equity & Civil Rights
- Standing Up for Consumers, Patients, Students
- Supporting Workers
- Promoting Innovation and Competition
- Advancing American Leadership Abroad
- Ensuring Responsible and Effective Government of AI
- See: https://ai.gov/

THE WHITE HOUSE

Administration

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM › STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy development of AI.

14

The EU AI Act (AIA), 2024

- Uses a risk assessment-based approach to AI regulation.
  - Unacceptable risk (banned)
    (social scoring by governments, exploiting vulnerabilities, and other manipulative techniques)
  - High risk (must be compliant)
    (systems used in credit scoring, employment, law enforcement must undergo risk assessment & mitigation, require high quality data, documentation, and have appropriate human oversight)
  - Limited risk
    (chatbots must meet transparency obligations: users must know they are interacting with an AI).
  - Minimal risk
    (Spam filters, bird recognition apps, etc.)

15

15

References

- Melanie Mitchell. *Artificial Intelligence: A Guild for Thinking Humans*. Farrar, Strouss, and Giroux, 2019.

16

16