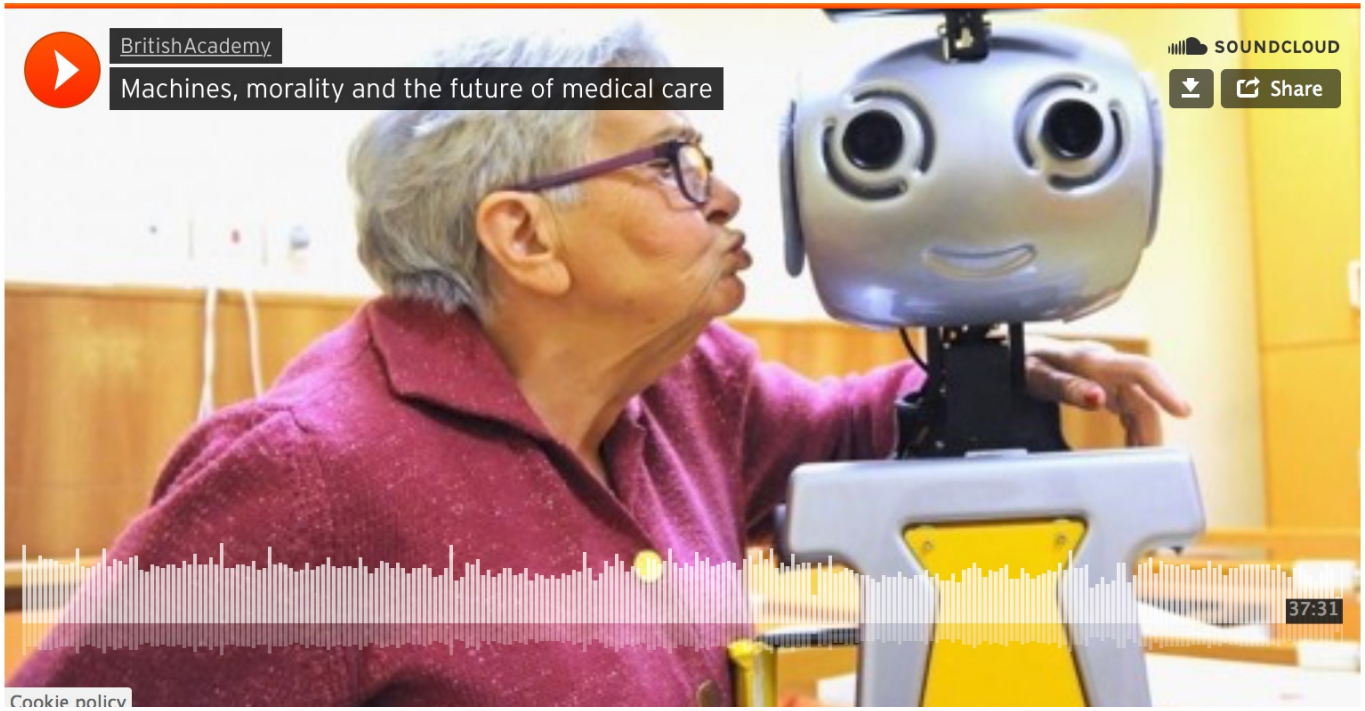
CS440/ECE448 Lecture 25: AI Ethics



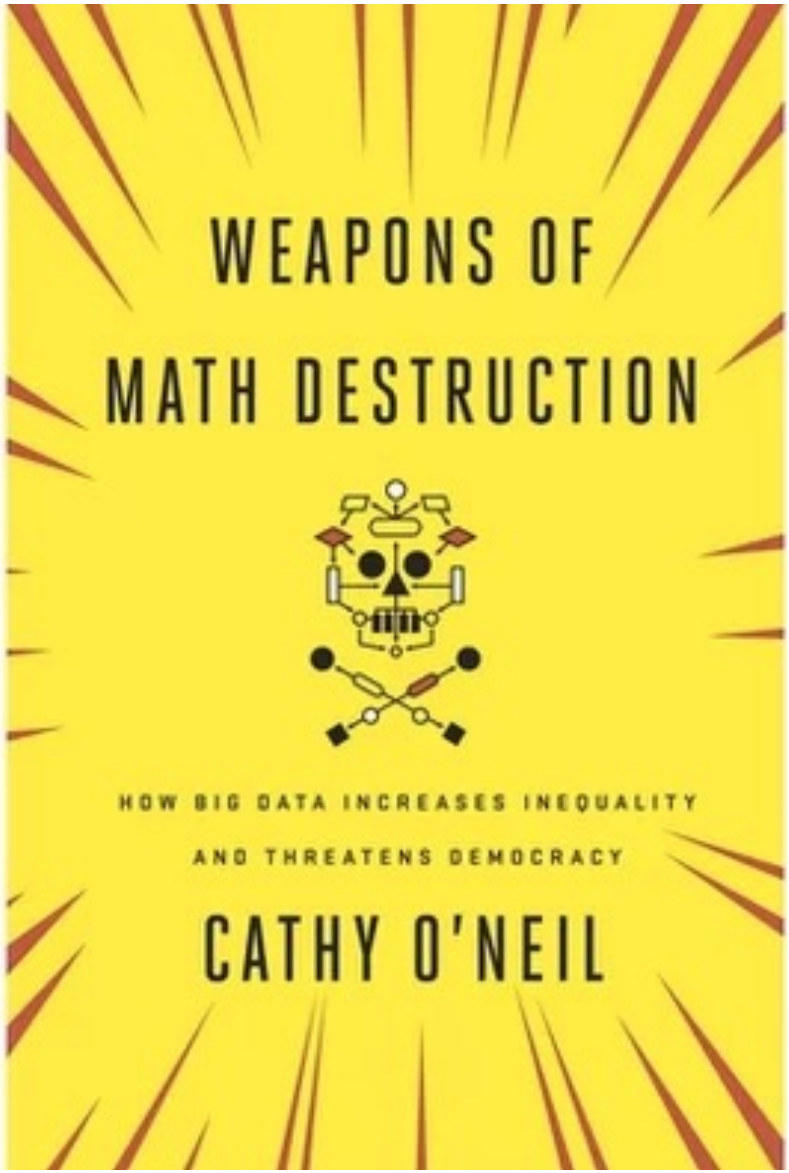
Modified by Mark Hasegawa-Johnson, 4/2021

Including slides by Svetlana Lazebnik, 10/2017

Image source: https://www.britac.ac.uk/audio/machines-morality-and-future-medical-care

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence


WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL

Examples of the problem

- **Opacity**: The “Level of Service Inventory-Revised” (LSI-R) was used to decide who gets parole in at least two states, and many counties/precincts.
  - It did not ask about race.
  - It did ask “when was your first encounter with police” and other questions that are highly correlated with race.
- **Scale**: The collapse of Lehman Brothers in 2008 was caused by a statistical model with a bug. Most large banks used the Gaussian copula model to decide who got home loans; it failed to correctly model the risk of multiple simultaneous defaults.
- **Damage**: Companies can’t use medical tests to determine hiring, but they are allowed to use personality tests. In 2016, a lawsuit found that at least seven companies were using the same personality test, and therefore rejecting the same applicants, for the same frivolous reasons.

Weapons of Math Destruction

Opacity, Scale, and Damage: a WMD is a statistical model afflicted by two of these three.

- Opacity: the relationship between inputs and outputs is hidden.
- Scale: the model is used at a scale much larger than it was ever tested for.
- Damage: negative decisions can damage people's lives.

Developments since 2016: Scale

- UCLA had 139,500 applicants in 2021 (CBS).
- In one 24-hour period (September 16, 2020), 384,000 people applied for jobs at Amazon (Forbes).
- NeurIPS had 9454 submitted papers in 2020. They don't use AI to review the papers (yet?), but they use an automated paper-reviewer assignment system. The same system (Toronto Paper Matching System) is used by ICML, CVPR, ICCV, and ECCV.


ACM FAccT
https://facctconference.org
ACM FAccT Conference
2021
2020
2019
2018
Network
Connect
Organization
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)
A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.
ACM FAccT is the new acronym for the ACM Conference on Fairness, Accountability, and Transparency!
ACM FAccT will be held online March 3-10, 2021. More information on this year's conference is posted on the 2021 ACM FAccT webpage.
Visit the ACM FAccT Registration Webpage to Register Now
Algorithmic systems are being adopted in a growing number of contexts, fueled by big data. These systems filter, sort, score, recommend, personalize, and otherwise shape

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

Bias caused by Data Sparsity

- Data contain more examples of one type than others, e.g., more Caucasians than African Americans
- Accuracy may be higher for the type that is better represented in the training data (minimize error by minimizing error for the majority case)
- Example: blacks more likely to be refused parole even if their prison records are the same (https://www.nytimes.com/2016/12/04/nyregion/new-york-prisons-inmates-parole-race.html)
- Example: tweets containing African American vernacular classified as "Danish," and therefore excluded from automatic sentiment analysis (https://www.technologyreview.com/s/608619/ai-programs-are-learning-to-exclude-some-african-american-voices/)

“Stereotyping and Bias in the Flickr30k Dataset,” Emiel van Miltenburg

www.cltl.nl/files/2016/05/LREC_Stereotypes_Emiel_van_Miltenburg.pdf

www.cltl.nl/files/2016/05/LREC_Stereotypes_Emiel_van_Miltenburg.pdf

GDP per capita Japan - Google Search

- *A blond girl and a bald man with his arms crossed are standing inside looking at each other.*
- *A **worker** is **being scolded by her boss in a stern lecture.***
- *A **manager** talks to an **employee about job performance**.*
- ***A hot, blond girl getting criticized by her boss.***
- *Sonic **employees talking about work**.*

- Inferring status
  - “worker” vs. “boss”
- Inferring intentions
  - “being scolded”
- Disrespect
  - “girl” vs. “man”
- Marking the “less common” attribute
  - girl vs. boss
  - blond vs. brunette
  - “nurse” vs. “male nurse”

… Never-ending learning is not the answer

- On March 23, 2016, Microsoft released a chatbot capable of never-ending learning from its interactions within users.
- Within 16 hours, users taught Tay to hate feminists and jews.
- After 16 hours, Microsoft stopped the software.

Image credit: CBS. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

Some possible answers

- Governments and private organizations now have funded efforts to acquire more data from under-represented groups.
  - Corpus of Regional African-American Language
  - Bureau of Justice Statistics
  - NIH Inclusion Policies for Research Involving Human Subjects
- Academia and industry seek to increase representation in AI data by increasing diversity among AI experts
  - AI4ALL

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

Standard Definitions of Fairness in AI

Let's define the following random variables:

- A = protected attribute: An observable fact that should not be predictive of outcomes, e.g., gender, race, age, disability.
- X = observable data that we can use for our decision
- Y = the unknown correct label for this person (e.g., $Y = 1$ might mean "this person should receive a loan" or "should be admitted to UIUC")
- $\hat{Y}$ = a function of X, designed using probabilistic or neural methods to approximate Y as closely as possible

Standard Definitions of Fairness in AI

Demographic Parity:

The probability of a positive outcome is the same, regardless of protected attribute.

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = a') \quad \forall a, a'$$

Predictive Parity:

Precision is the same, regardless of protected attribute.

$$P(Y = 1 | \hat{Y} = 1, A = a) = P(Y = 1 | \hat{Y} = 1, A = a') \quad \forall a, a'$$

Error Rate Balance:

Recall is the same, regardless of protected attribute.

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = a') \quad \forall \hat{y}, a, a'$$

You can't have all three

$$P\left(\hat{Y}=1|Y=1,A=a\right)=\frac{P\left(Y=1|\hat{Y}=1,A=a\right)P\left(\hat{Y}=1|A=a\right)}{P(Y=1|A=a)}$$

$$P\left(\hat{Y}=1|Y=1,A=a'\right)=\frac{P\left(Y=1|\hat{Y}=1,A=a'\right)P\left(\hat{Y}=1|A=a'\right)}{P(Y=1|A=a')}$$

The balanced error, predictive parity, and demographic parity terms cannot all be independent of A unless Y is also independent of A.

In other words, if the current state of society is unfair (distribution of positive outcomes currently depends on protected attribute), then algorithmic solutions cannot make it fair (at least not in all three ways, all at once).

Other problems with algorithmic solutions

Dwork (2012) points out that demographic parity can lead to socially undesirable outcomes, e.g., people gaming the system.

… but …

Srivastava, Heidari and Krause (2019) found that users of an AI judge its fairness based on demographic parity. They ignore predictive parity and balanced error, even when these concepts are explained to them.

Other Useful Definitions of Fairness in AI

Individual Fairness:

The dissimilarity between two outcomes should be less than the dissimilarity between the people.

Counterfactual Fairness:

If a person's protected attribute were changed (and all their other attributes were possibly changed, according to their dependence on the protected attribute), then the outcome should not change.

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

Are College Admissions Fair?

- Bickel, Hammel, and O'Connell, "Sex bias in graduate admissions: Data from Berkeley," Science 187(4175):398–404, 1975
- Women were being admitted to Berkeley at a far lower rate than men.
- Women were applying to departments with lower acceptance rates. Within each department, admission rates for men and women were the same.
- Is this fair?

Is Employment Fair?

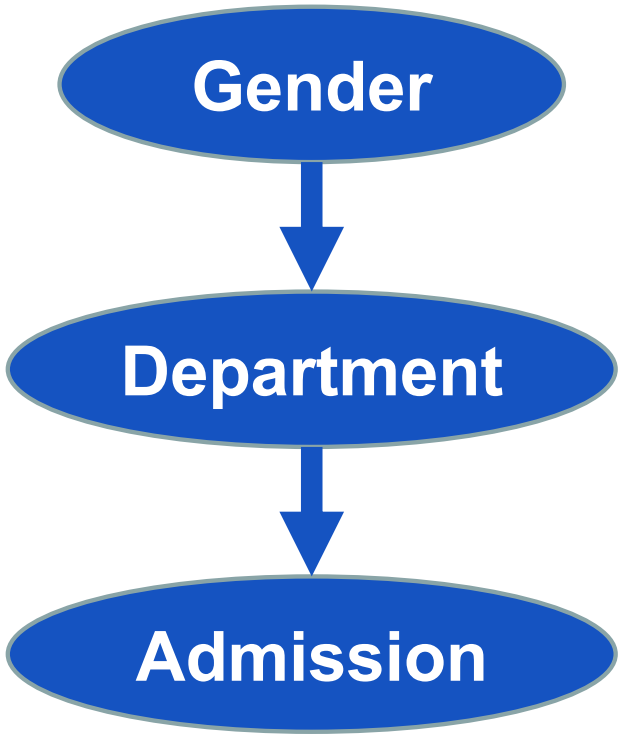
A given job might discriminate based on:

- Upper body strength
- Habitual clothing
- Undergraduate major

All of these correlate with gender. Is it fair to use them as a basis for employment?

State-of-the-art Solution: Explain Your Assumptions to Your Users

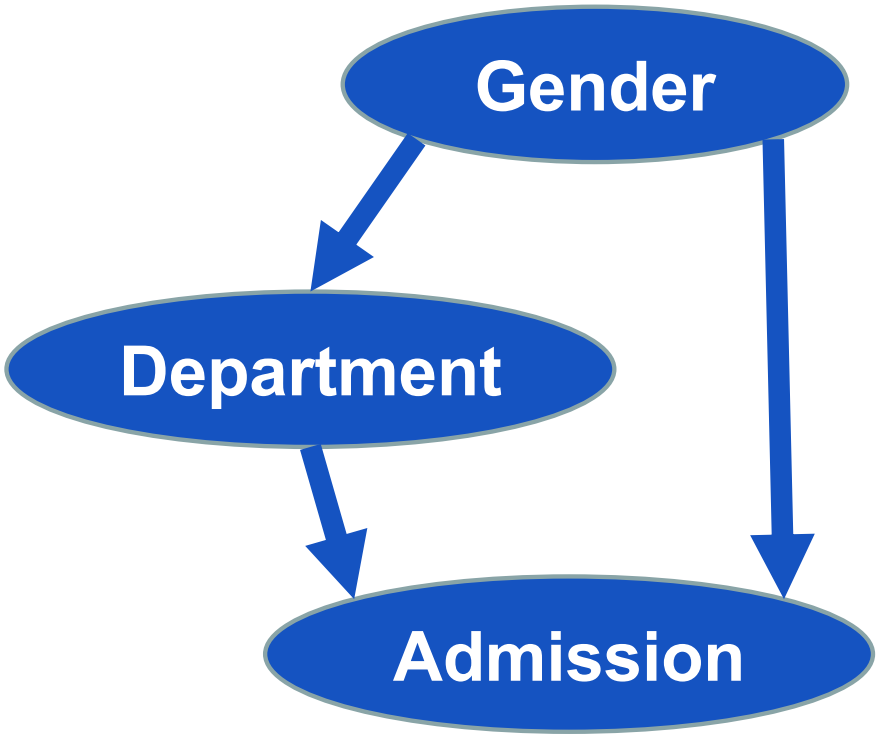
- The Transparency Dogma says that an algorithm's assumptions should be published in a way that users can understand, so that users can participate in a public debate about the fairness of the assumptions.
- A common way to do this is by drawing a Bayesian network.
- Example shown: the Bickel-Hammel-O'Connell data. Admission was conditionally independent of Gender, given Department.



State-of-the-art Solution: Explain Your Assumptions to Your Users

A possible solution: in order to better approximate demographic parity,

- Make admission explicitly dependent on gender.
- Admit women at a slightly higher percentage, in every department to which they apply, so that…
- …the total percentage of admitted women equals the total percentage of admitted men.



Public debate: is that more fair, or less fair?

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

REBOOTING AI

Building Artificial Intelligence We Can Trust

GARY MARCUS
and ERNEST DAVIS

Trust Issues

- Physical Safety
  - April 18, 2021: 2 Killed in Driverless Tesla Car Crash
- Data Safety
  - March 2020: CAM4 data breach exposed 10 billion records
- Economic Safety
  - As of 2021, it's a little hard to judge the effect of AI on jobs, but this has been a topic for some years now.

Physical Safety

- Robustness to changes in data distribution
- Avoiding catastrophic “edge cases”
- Robustness to adversarial examples or attacks
- Avoiding negative side effects in reward function
- Avoiding “reward hacking”

- Reading: Concrete AI safety problems

The Virtual Sully Research Project

Creative commons 2.0, multichill, 2009

- At 3:27 on 1/15/2009, US Airways 1549 lost power in both engines.
- Capt. “Sully” Sullenberger tried to turn back to LaGuardia, then tried to turn toward Teterboro, then realized there was no time.
- At 3:31 he landed the plane in the Hudson river.
- All passengers were saved.

The Virtual Sully Research Project

Creative commons 2.0, multichill, 2009

Virtual Sully research project seeks to give AI

- the ability to plan a course of action with backup plans available in case of unexpected disaster,
- the ability to quickly discard low-priority goals in favor of threatened high-priority goals in case of the unexpected inability to achieve both.

AI weapons

Australian and Canadian **AI** Experts Call for Autonomous **Weapons** ...
Futurism - Nov 8, 2017
In two letters addressed to the heads of state in Australia and Canada, hundreds of experts in the field of artificial intelligence (**AI**) have urged for the **ban** of "killer robots," artificially intelligent **weapons** with the ability to decide whether a person lives or dies. They join a growing crowd of scientists who have ...

When **AI** rules, one rogue programmer could end the human race
BGR - Nov 8, 2017

Artificial intelligence will soon be used to create '**weapons** of mass ...
International Business Times UK - Nov 8, 2017

Canadian **AI** experts urge for global **ban** on killer robots
International - BetaKit - Nov 8, 2017

'Slaughterbots' film shows potential horrors of killer drones
CNNMoney - Nov 14, 2017
The film is the researchers' latest attempt to build support for a global **ban** on autonomous **weapon** systems, which kill without meaningful human control. They released the video to coincide with meetings the United Nations' Convention on Conventional **Weapons** is holding this week in Geneva, ...

Killer robots are almost a reality and need to be **banned**, warns ...
Telegraph.co.uk - Nov 14, 2017

Ban autonomous killer robots, urge **AI** researchers
Radio Canada International - Nov 14, 2017

'Slaughterbots' Video Depicts a Dystopian Future of Autonomous ...
In-Depth - Seeker - Nov 15, 2017

The UN is worried about killer robots. We should be, too.
News & Observer - Dec 4, 2017
Agreements **banning** nuclear **weapons** from space are likewise a precedent, as are those prohibiting the use of laser **weaponry** to blind people, enacted by the Convention on Certain Conventional **Weapons** in 1995. We need a **ban** on autonomous offensive **weapons** in a similar way. As with nuclear ...

AI weapons

- Reading
  - Robotics: Ethics of artificial intelligence (Nature, May 2015)
  - Humans, not robots, are the real reason artificial intelligence is scary
    (The Atlantic, August 2015)

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

Data Safety

"Passports, however, use a different technology known as RFID (or Radio Frequency Identification), the same type used to tag clothing, pets, even artificial replacements for hips and knees. When embedded in a U.S. passport, the chip can be scanned only by someone at close range with an RFID reader, usually within a couple feet…

"Yes, someone nearby could read what's in your wallet. That's why I keep my passport in an RFID-shielded wallet," said G. Mark Hardy, president of National Security Corp., based in Rosedale, Md., which provides cybersecurity expertise to government and corporate clients.

But, he said, "it's less likely to happen, at this point in time, because it's so much easier to do fraud some other way.""

Read more here: http://www.sacbee.com/news/business/personal-finance/claudia-buck/article2599038.html#storylink=cpy

Example of a Technical Solution: Homomorphic Encryption

1. Encrypt the data on your cell phone
2. Send the encrypted data to a server
3. The server sends it through a neural net in its encrypted form, without ever decrypting it
4. They send you the result, and you decrypt it using the same key

Example of a Technical Solution: Homomorphic Encryption

Requirements: if $\varepsilon(x_1)$ and $\varepsilon(x_2)$ are the encrypted forms of x_1 and x_1, then it must be the case that

- $\varepsilon(x_1 + x_2) = \varepsilon(x_1) + \varepsilon(x_2)$
  - Satisfied by Pallier encryption
- $\varepsilon(x_1 x_2) = \varepsilon(x_1)\varepsilon(x_2)$
  - Satisfied by RSA encryption
- $\varepsilon(\max(0, x_1)) = \max(0, \varepsilon(x_1))$

Full homomorphic encryption (FHE) is possible since 2009. A neural net can process data without ever having to decrypt it. Still computationally expensive, but new methods are being developed.

Example of a Legal Solution: General Data Privacy Regulation (GDPR)

It is illegal for a European entity to:

- Art.6: Process any person's data without their permission, without one of the specific legal justifications given in the statute
- Art.7: Make it harder to remove consent than it was to give consent
- Art.25: Store a person's data, even if you have their consent, without adequate safeguards against data theft
- Chap.V: Take data outside the EU, without adequate safeguards

Any person has the right to:

- Art.12: Know how your algorithm works, in terms they understand
- Art.15: Know what data you hold
- Art.25: Refuse to allow you to use their data for any other purpose

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence

Economic Safety: AI and Jobs

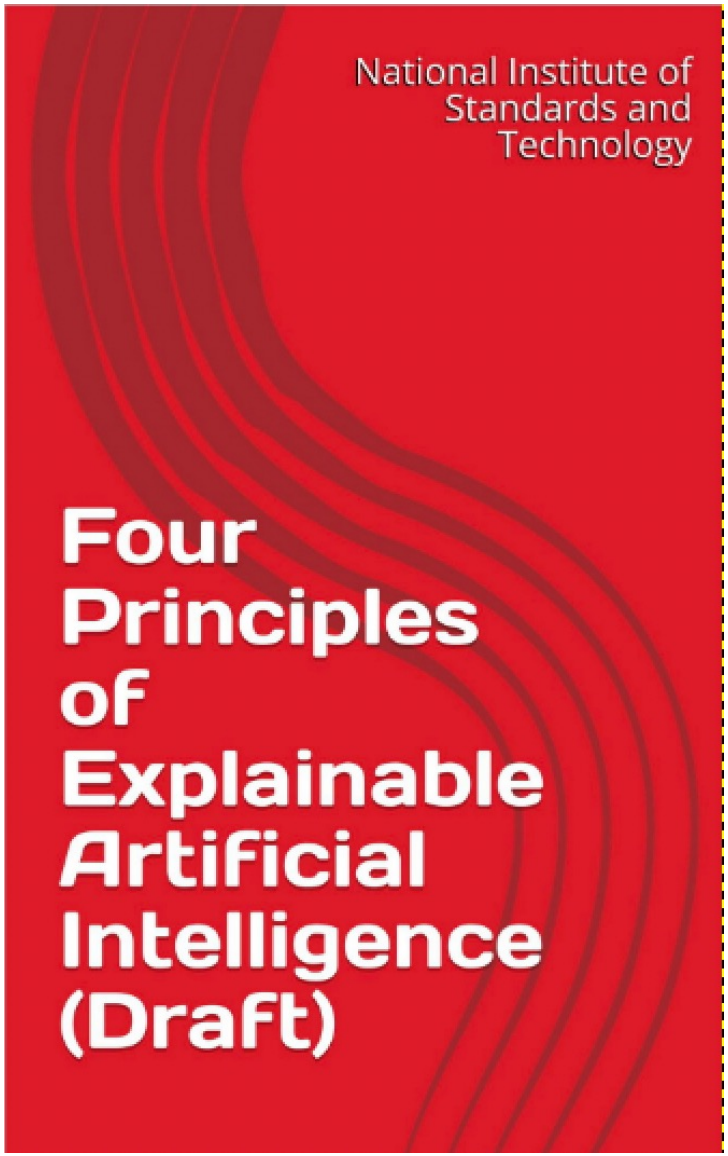
- Why we should worry
  - Oxford report: 47% of American jobs at high risk of automation in the next two decades
  - In the past couple of decades, manufacturing employment has dropped even as output kept rising; labor force participation among working-age males has been dropping
  - Truck driver is the most common job in over half the states
- Why we shouldn't worry
  - Productivity growth is currently low, as is business investment spending
  - Historically, automation has destroyed jobs but added more new jobs

Economic Safety: AI and Jobs

- Government interventions: regulation (safety, anti-trust, or explicit job protection regulations)
  - Examples: US vs. Microsoft, EU vs. Google, Everybody vs. Uber
- Government interventions: wealth redistribution
  - Retraining programs
  - Stimulus checks
  - Universal basic income: Stockton Economic Empowerment Demonstration

Outline

- What's the problem?
- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence


National Institute of Standards and Technology
Four Principles of Explainable Artificial Intelligence (Draft)

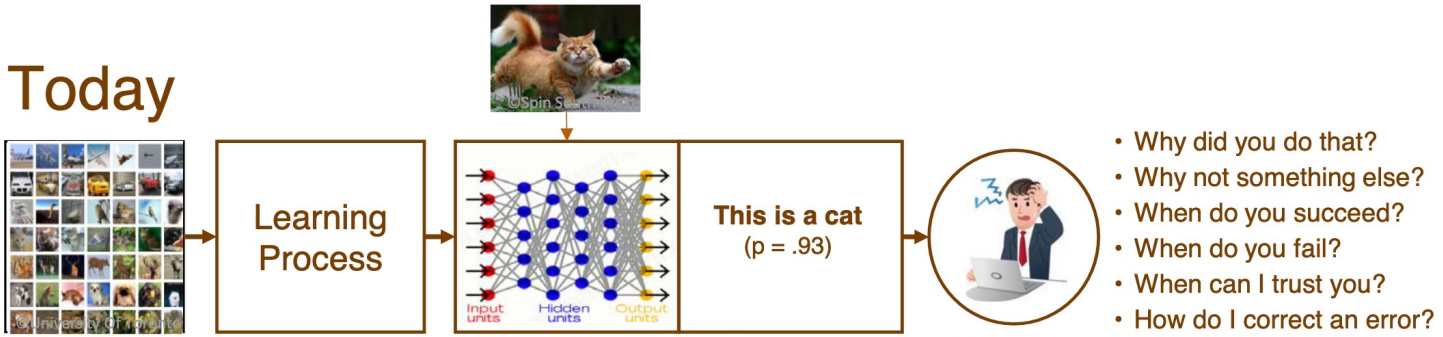
Four Principles of Explainable AI

- Explanation: The system can explain its reasons for any decision
- Meaningful: The explanation can be understood by the user
- Explanation Accuracy: The explanation correctly describes how the system made its decision
- Knowledge Limits: The system is only used under circumstances for which it was designed.

Explainable AI – What Are We Trying To Do?

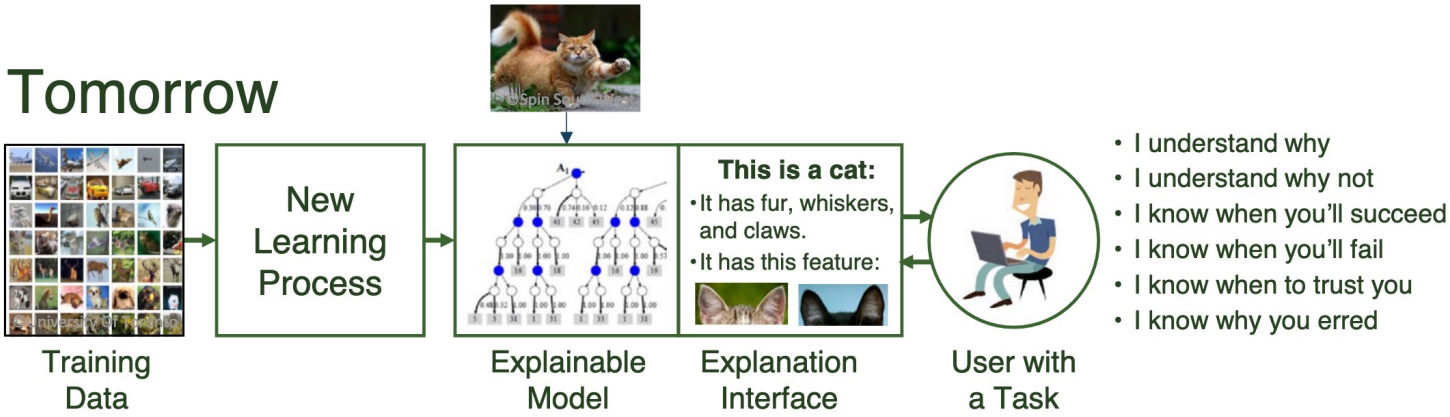
David Gunning, "Explainable Artifcial Intelligence (XAI)," 2017

Today



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Methods

- Visualization
  - Pro: provides intuitively useful descriptions of typical behavior
  - Con: post-hoc explanation of the typical behavior; may not tell you much about worst-case behavior
- Causal Graphs/Bayesian Networks
  - Pro: describes reasoning process of the AI exactly
  - Con: constraining AI reasoning process to obey an explainable causal graph sometimes harms accuracy

NICK BOSTROM

SUPERINTELLIGENCE

Paths, Dangers, Strategies

Superintelligence

Basic reasoning:

- Never-ending learners (like Tay) will start improving themselves
- Such an AI will continue to follow whatever directives were part of its original programming
- Those directives may contain subtle bugs resulting in large-scale damage to humanity (e.g., “make as many paperclips as possible”)

Superintelligence

The public debate:

- Most objections to this book have centered around the impossibility of general autonomous AI.
  - Supervised learning results in an AI whose performance reaches some asymptote, and stays there
  - Reinforcement learning is provable and useful only in limited contexts
  - Never-ending learning (like Tay) tends to fail in big, obvious ways
- Bostrom's response: yes, but people are trying to solve those problems. Shouldn't we at least think about the outcome?

Possible Solutions

- Air-Gap the AI? Has already failed, many times. The AI can easily recruit a human collaborator by offering rewards.
- Regulation? Bostrom's insight: a superintelligence is basically a giant corporation. It can be managed in the same way that we manage giant corporations, e.g., using GDPR-scale penalties.
- Imprint the AI with carefully tuned long-term directives:
  - Asimov's "three laws of robotics" are the right idea, but too simple.
  - Empathy. Psychologists believe that emotion is necessary for effective decision-making, and empathy is necessary for emotion.
  - Game Theory. Is cooperation more effective than defection?
  - Morality. Are there moral laws that would be considered binding by any rational agent?

Growing Fields of Research in Artificial Intelligence

- Fairness
  - Representation
  - Algorithmic solutions
  - Transparency as a solution to the fairness problem
- Accountability
  - Physical safety
  - Data safety
  - Economic safety
- Transparency
  - Explainable AI
  - Understanding the inner workings of a superintelligence