

# CPSC 430

# Computers & Society

## Class 6A: Bias, Fairness, and Artificial Intelligence

Dr. Firas Moosvi | 2024\_S2

Slides courtesy of Dr. Kevin Leyton-Brown

# Class Outline

1. Announcements (5 mins)
2. AI Journey So far and Current Trends (20 mins)
3. Spot the Fakes (20 mins)
4. Student Experience of Instruction (5 mins)
5. Break (10 mins)
6. Bias and Fairness (50 mins)
7. Break (10 mins)
8. Peering into the crystal ball of AI (50 mins)

# Announcements

- Essay 0-7 scores are released
- We will handle Appeals as they come in...
- Last Participation scores will be released after class today
- Final Course grades will be available at some point next week

# August 2024



# The (AI) Journey So Far...

What is AI and what can it do today?  
How much is real and how much is hype?



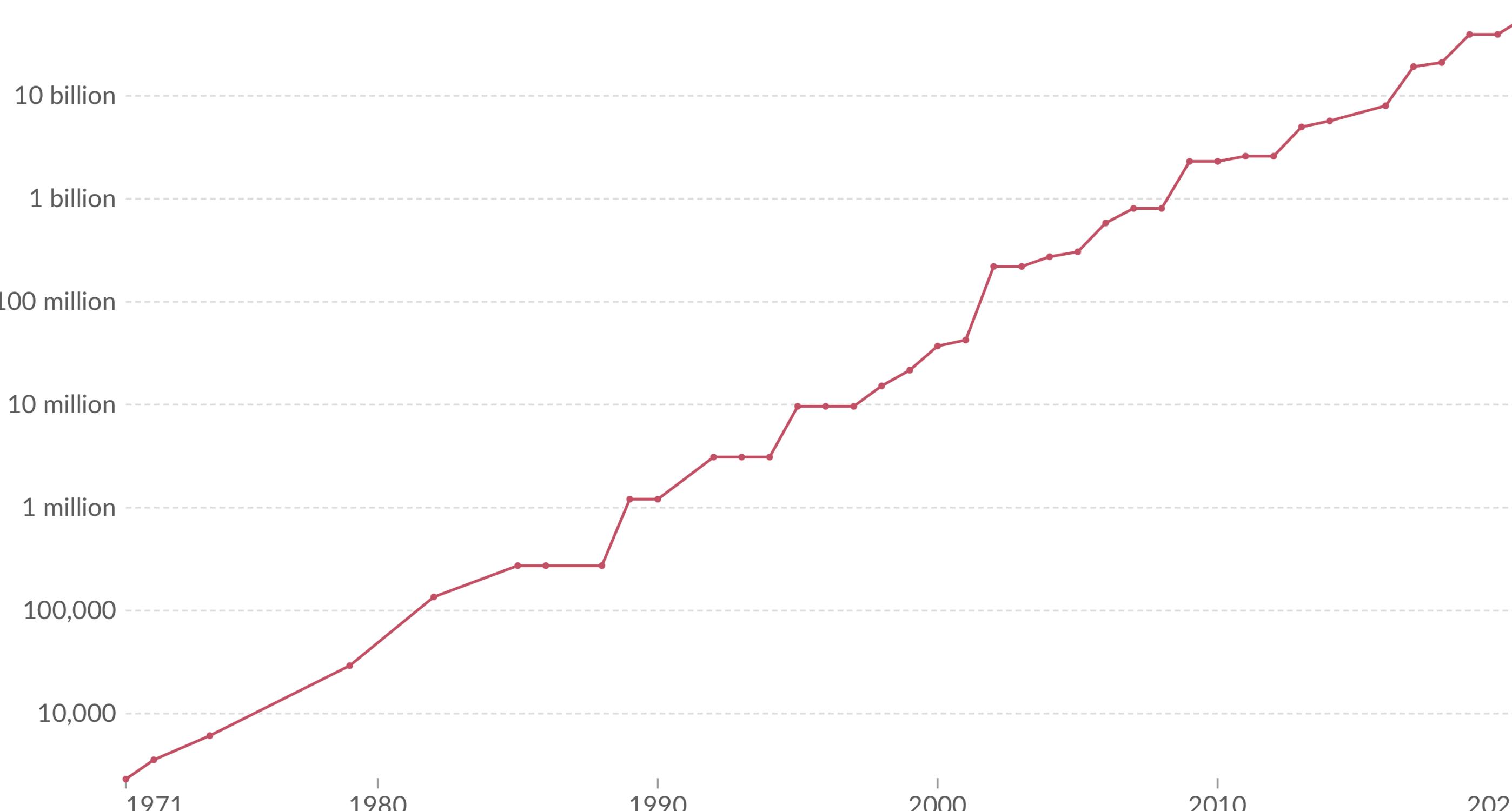
# Why is it happening?

1. Scientific/mathematical **breakthroughs**, especially in machine learning
2. Growth in raw **computing power**

## Moore's law: The number of transistors per microprocessor

Moore's law is the observation that the number of transistors in an integrated circuit doubles about every two years, thanks to improvements in production. It was first described by Gordon E. Moore, the co-founder of Intel, in 1965.

Our World  
in Data



For comparison:

Species	# Neurons in Brain
Fruit Fly	$10^5$
Cat	$10^9$
Chimpanzee	$10^{10}$
Human	$10^{11}$

What short term impacts should we expect?

What kinds of social impacts are they having?

# CURRENT TRENDS

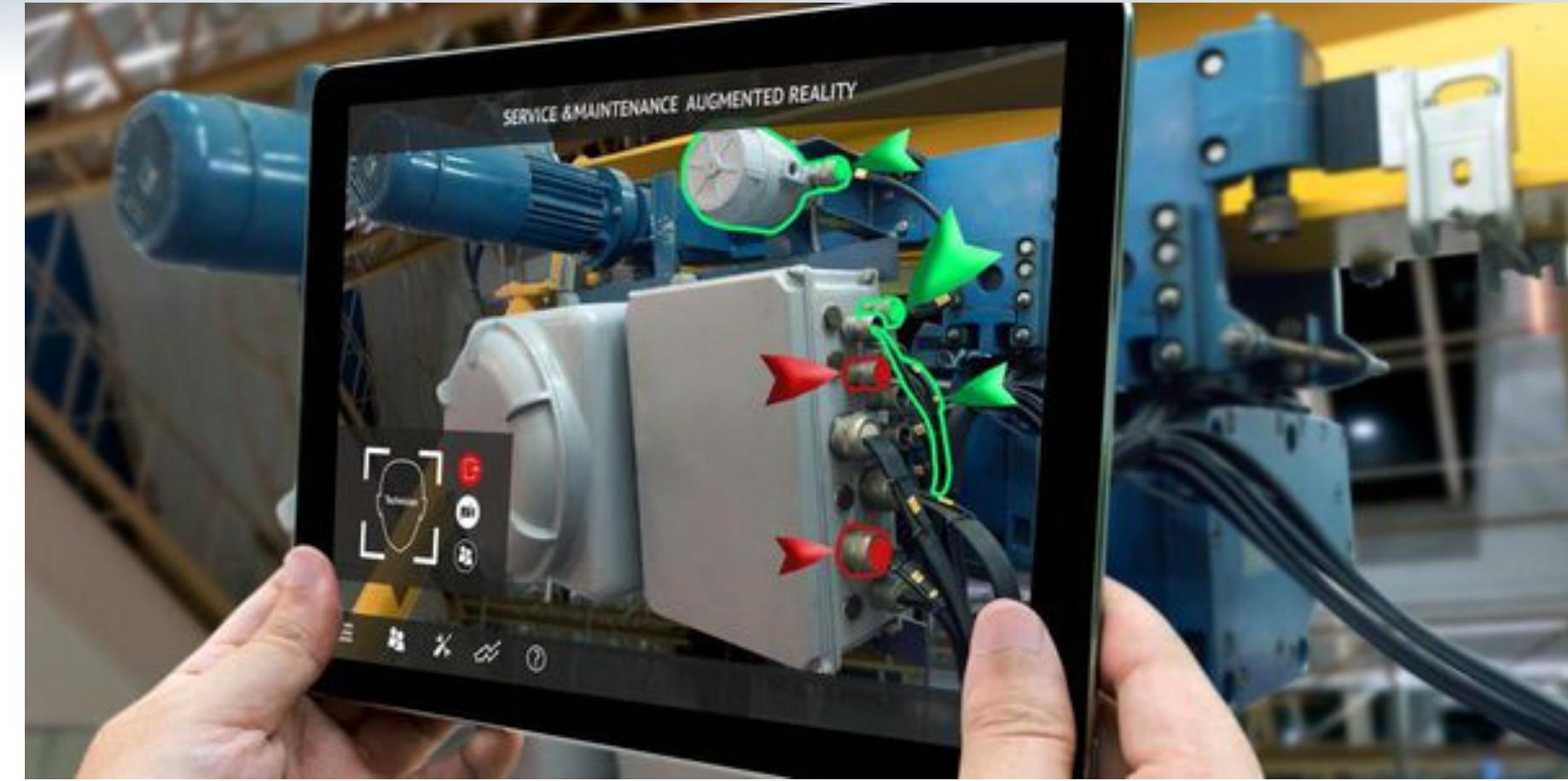
# Machine Learning

- Narrow: building **predictive models** from data
- Broad: a new approach to computer programming, where computers **learn from experience** rather than being told what to do
- Where's it **headed?**
  - More and more data collection
    - sensors
    - behavior tracking
  - Image, text, speech understanding
  - Data mining/personalization/profiling
  - Self programming/optimization



# Virtual and Physical Worlds Collide

- Smart **interfaces**
  - remote controls; microwaves; clocks; ...
- Augmented reality
- More intuitively “**programmable**” systems
- Much better **context awareness**
- **Assistance**
  - elderly; children; disabled; education
- **Social agents/games**



# Growing Military Use

- Increasing use of AI in **military applications**
- Already quite **commonplace** in some applications
  - smart bombs
  - surveillance systems
  - missile defence (Patriot, Iron Dome, etc)
- **Autonomous weapons** are becoming practical
  - drones are widely used; becoming more autonomous
  - land-based robotic systems are improving quickly
  - swarms of drones could revolutionize urban warfare
- The main barrier to adoption is **public opinion**



# LLMs in the workplace

## OpenAI report:

- How “exposed” are different occupations to automation with large language models (LLMs)?
- ~80% of the US workforce could have at least 10% of their work tasks affected by LLMs
- ~19% may see at least 50% of their tasks impacted
- Most affected tasks: writing and programming.
- Higher-income jobs (such as translators, tax consultants, and web designers) potentially face greater exposure

### GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou, Sam Manning, Pamela Mishkin, Daniel Rock

We investigate the potential implications of large language models (LLMs), such as Generative Pre-trained Transformers (GPTs), on the U.S. labor market, focusing on the increased capabilities arising from LLM-powered software compared to LLMs on their own. Using a new rubric, we assess occupations based on their alignment with LLM capabilities, integrating both human expertise and GPT-4 classifications. Our findings reveal that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted. We do not make predictions about the development or adoption timeline of such LLMs. The projected effects span all wage levels, with higher-income jobs potentially facing greater exposure to LLM capabilities and LLM-powered software. Significantly, these impacts are not restricted to industries with higher recent productivity growth. Our analysis suggests that, with access to an LLM, about 15% of all worker tasks in the US could be completed significantly faster at the same level of quality. When incorporating software and tooling built on top of LLMs, this share increases to between 47 and 56% of all tasks. This finding implies that LLM-powered software will have a substantial effect on scaling the economic impacts of the underlying models. We conclude that LLMs such as GPTs exhibit traits of general-purpose technologies, indicating that they could have considerable economic, social, and policy implications.



The report answers the question “which jobs do people think LLMs will automate” and not “which jobs will LLMs automate”.

# LLMs in the workplace

- Adoption of Large Language Models for various tasks, including work-related ones
  - Some people compare it to the advent of Google
  - Some issues (ownership of training data, hallucinations...)
- Have you used LLMs before?
- Do you think LLMs should be used in the classroom? How?
- Do you think that AI will increase the expectations from professionals in the workforce? How?

# Current State of the Art

- **Predictive modeling** limited mostly by quality of training data
- **Image recognition, face recognition** becoming extremely reliable
- **Image generation** almost there
- **Language understanding & generation** maturing extremely quickly. Watch this space
- **Fueling scientific breakthroughs** from protein folding to astronomy
- **Robotics** remains a hard problem, but it's working well in some niche applications



Prompt: Here's a short rhyming poem by Dr. Seuss about Elon Musk's company, Neuralink which is building brain implants to achieve symbiosis with AI. Neuralink's mission statement is "If you Can't Beat'em, join'em". The poem has a structure typical of Dr. Seuss stories with simple alternate line rhymes. In the poem Elon explains why we need to merge with the AI to save humanity.

Neuralink, by Dr. Seuss

----

Response:

I have a friend, Elon Musk,  
Who wants to join our brains to AI.  
He thinks that humans can't compete  
So he's gonna use AI to cheat.  
They want to build a neural-link.  
To link us to the AIs  
And save humanity from its demise.  
They want to put an AI in your brain.  
They want to use your neurons as a frame.  
They want to save humanity,  
If you can't beat'em, join'em

# Spot The Fakes

INNOVATIONS

# AI can draw hands now. That's bad news for deep-fakes.

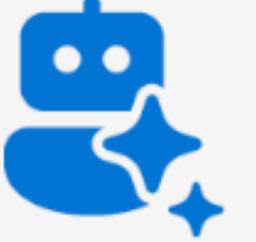
Improved AI image-making software churned out believable yet fictional images of former president Donald Trump being arrested

⌚ 7 min ⏪ 313



# Activity: Spot the Fakes

- Do the quiz once individually (3 minutes)
- How'd you do?
- Do the quiz together in a group of 3-5



## Real or Not ?

The growing quality in AI images makes them harder to spot.  
**Can you tell if this image is real or AI generated?**

[Start Game](#)

# **Student Experience of Instruction (SEI)**

# Break

Reminder: Please fill out the Student Experience of Instruction surveys! 🙏

# Bias and Fairness

# Human bias

Bias in people refers to our tendency to take quick decisions based on little information

The collage consists of four separate news snippets arranged in a 2x2 grid:

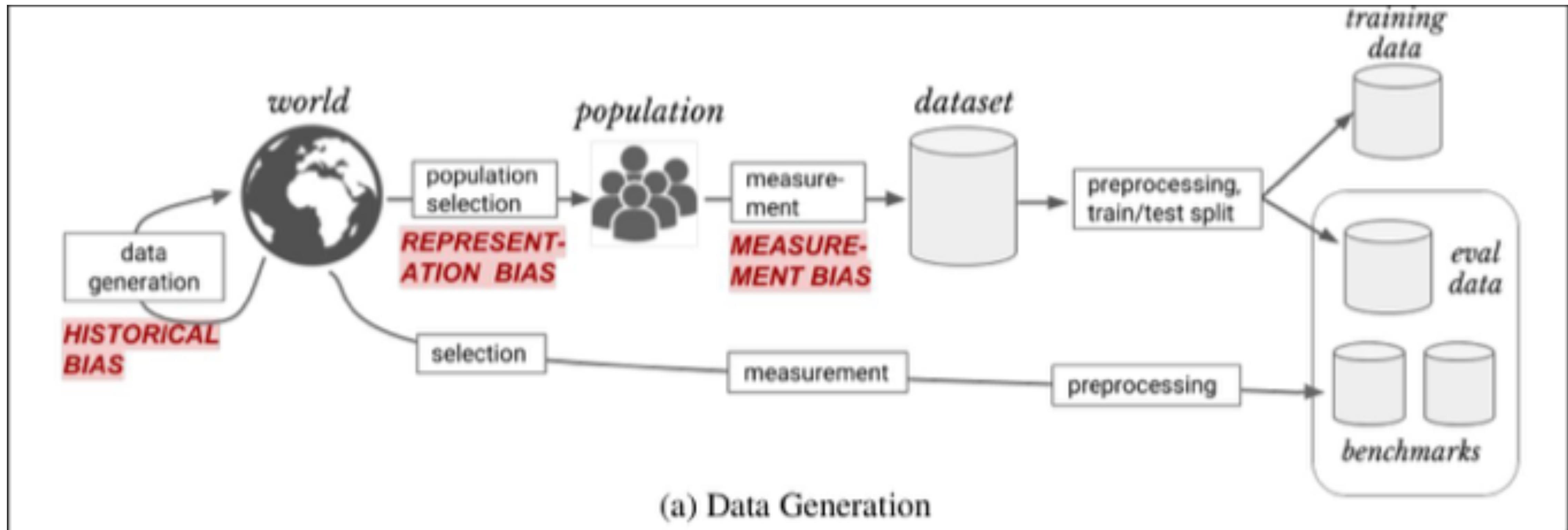
- Top Left:** A snippet from Nature news titled "Hungry judges dispense rough justice". It includes a quote: "When they need a break, decision-makers gravitate towards the easy option." - Zoë Corbyn.
- Top Right:** A snippet from The Observer titled "Racial bias in police stop and search getting worse, report reveals". It includes a quote: "Despite reforms, black people are nine times more likely than white people to be checked for drugs" - Mark Townsend.
- Bottom Left:** A snippet from Journal of Economic Perspectives titled "Evidence on Discrimination in Mortgage Lending". It includes a quote: "Helen F. Ladd".
- Bottom Right:** A snippet from BBC Inside Out titled "Is it easier to get a job if you're Adam or Mohamed?". It includes a photo of two men holding signs with their names: MOHAMED ALLAM and ADAM HENTON.

# Why worry about bias in algorithms

Decisions made by a ML algorithm are:

- Cheap
- Scalable
- Automated
- Self-reinforcing
- Seemingly objective
- Often lacking appeals processes
- Not just predicting but also causing the future

# Sources of bias in ML algorithms



# Representation bias

**Representation bias** arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

1

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may be available only for the population of patients who were considered serious enough to bring in for further screening.
2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.

# Measurement bias

**Measurement Bias** arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

1

## 3. The defined classification task is an oversimplification.

In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of ‘successful student’ in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as ‘GPA’ (Kleinberg et al. 2018), which ignores different indicators of success achieved by parts of the population.

1. **The measurement process varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst 2016).

2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (Calderone, 1990). In this case, “*diagnosed* with condition X” is a biased proxy for “has condition X.”

# Historical bias

**Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

1

**Example: image search** In 2018, 5% of Fortune 500 CEOs were women (Zarya, 2018). Should image search results for “CEO” reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for “CEO” to display a higher proportion of women.

# Fairness in algorithms

- Increasing attention on algorithms being **fair**, not just accurate
- Fairness can be measured as:
  - demographic (or statistical) parity: population percentage should be reflected in the output classes
  - Equality of false negatives or equalized odds: constant false-negative (or both false-negative and true-negative) rates across groups.
  - Equal opportunity: equal True Positive Rate for all groups
  - Other metrics...
- Accuracy and fairness tend to be at odds with each other.
- Algorithms can be audited to test their fairness.
- *Are we ethically required to sacrifice accuracy for fairness? To what extent?*

# When the metric becomes the target (Goodhart's Law)

*"When a measure becomes a target it ceases to be a good measure"*

- Metrics introduced in the [British public healthcare system](#) (e.g. waiting time in ER) caused people to game it:
  - Cancelled scheduled operations to draft extra staff to ER
  - Required patients to wait outside the ER, e.g. in ambulances
  - Put stretchers in hallways and classified them as "beds"
  - Hospital and patients reported different wait times
- Big Data is significantly changing college applications
  - Universities are given higher ranking for things such as receiving more applications, being more selective, and having more students accept their offers (while tuition is not considered)
  - This even pushed some mid-tier universities to reduce the number of offer letter sent out, especially to good students who they think will not accept. Can affect applications to "safety schools"
- *Can you think of more examples?*
- *Can you think of ways to avoid this trap?*

# Ethics of pricing algorithms

- Algorithms are currently used to adjust prices based on:
  - Willingness of buyer
  - Availability
- Uber surge pricing:
  - In 2014, terrorists attacked a café in Sidney, holding 10 customers and 8 employees hostage for 16 hours
  - During this time, people from the surrounding areas were evacuated. Transportation was disrupted.
  - Uber prices adapted by increasing the rate to a minimum of 100\$
  - In general, underserved (poorer) areas get worse rates under current pricing policy
  - Drivers are also subjected to different pricing/waiting times
    - [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4331080](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4331080)
- *Is Uber morally obliged to avoid such pricing disparities?*

# From trading to gambling addiction?

- Case study: Robinhood Trading app
- Designed to make trading more accessible and equitable. Robinhood's mission statement is "To democratize finance for all"
- Concerns around the most “dopamine-inducing” features ([source](#)):
  - Green confetti to celebrate transactions.
  - A constant update of stock related articles.
  - A colorful, eye-catching interface.
  - Emoji phone notifications.
  - One-click trading for instant gratification.
  - Free stocks in the shape of lottery tickets.



# Break

Reminder: Please fill out the Student Experience of Instruction surveys! 🙏

Peering into the Crystal Ball

# **PEERING INTO THE CRYSTAL BALL**

Bold predictions about where AI is headed in the longer term

# Technology that will be in widespread use

- Tailored solutions for **specific tasks**, not general intelligence
- Prototypes **that work today** in labs & narrow deployments
- Some examples:
  - **Non-text input modalities** (vision; speech)
  - **Consumer modeling** (recommendation; marketing)
  - **Cloud services** (translation; question answering; AI-mediated outsourcing)
  - **Transportation** (automated trucking; some self-driving cars)
  - **Industrial robotics** (factories; some drone applications)
  - **AI knowledge work** (logistics planning; radiology; legal research; call centers)
  - **Policing & security** (electronic fraud; cameras; predictive policing)



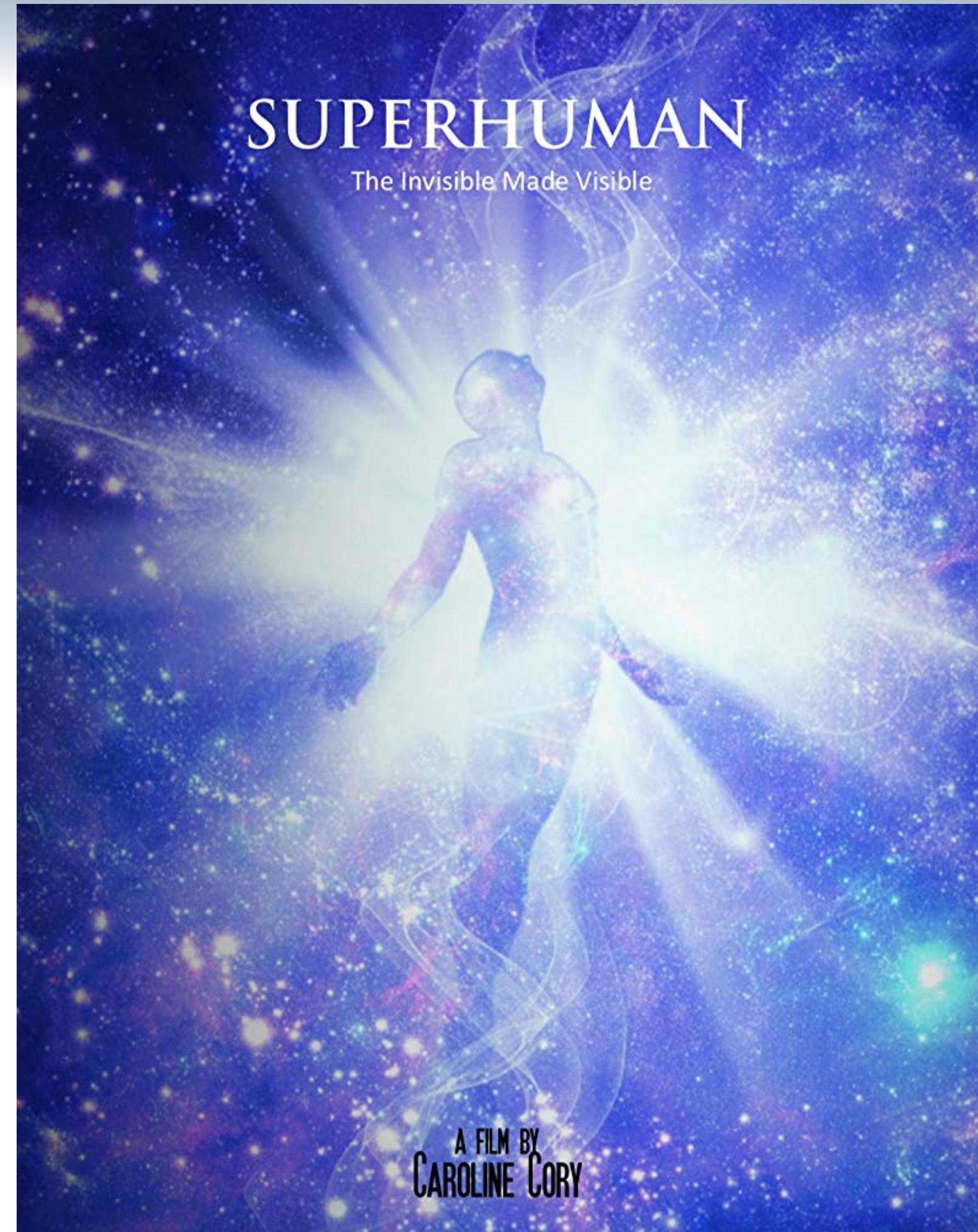
# Technologies that won't take off as quickly

- Overall, areas in which
  - major entrenched **regulatory regimes** need to be navigated
  - there exist **social/cultural barriers** to adoption
  - the **human touch** is crucial
  - substantial **new hardware** would need to be developed
- Some **examples**:
  - childcare, healthcare, eldercare
  - education
  - coaching, counselling
  - consumer robots beyond niche applications
  - semantically rich language understanding



# Superhuman Intelligence

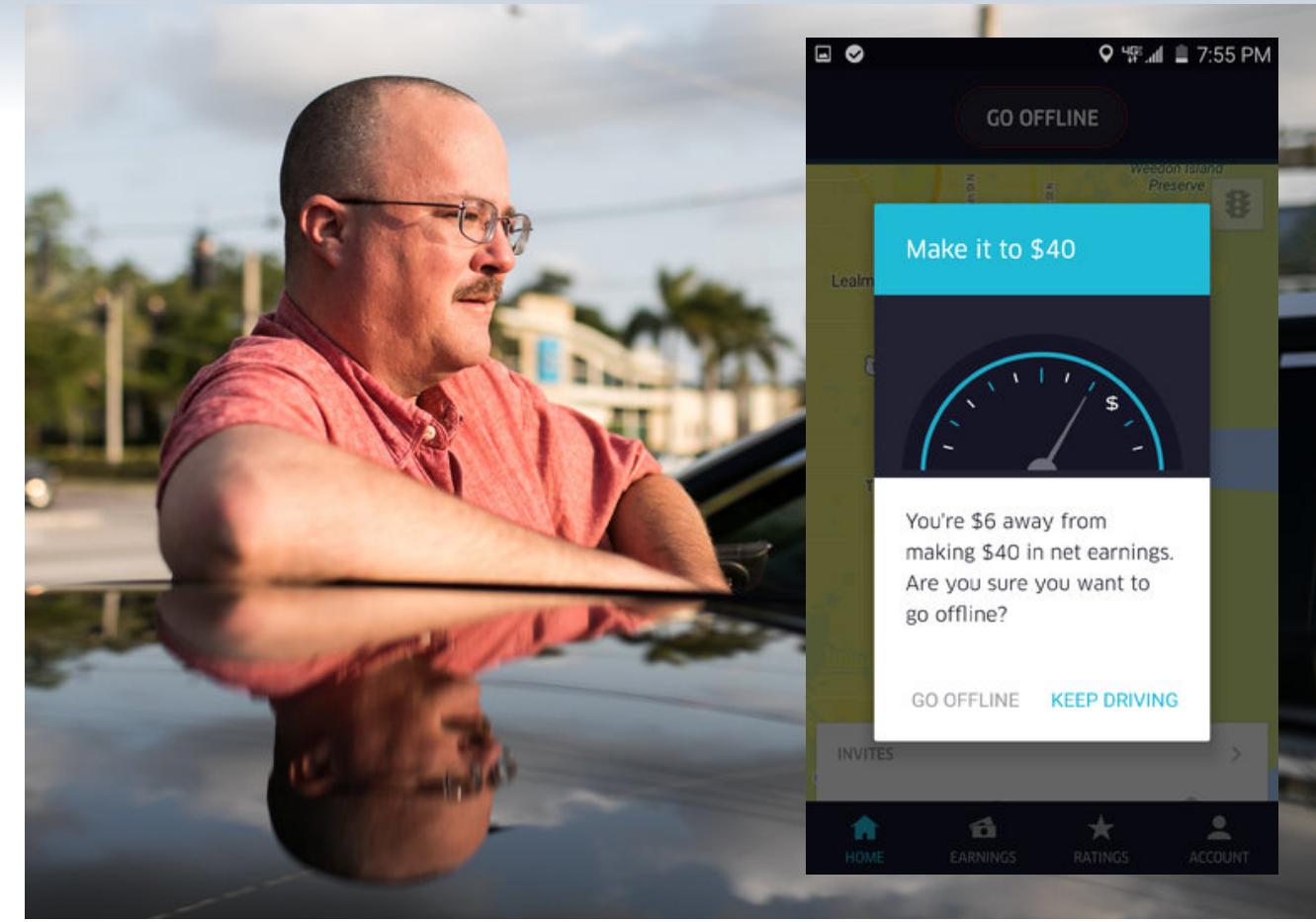
- AI systems will increasingly be capable of reaching **human-level performance**
- **Superhuman intelligence** isn't such a foreign, scary thing
  - governments, corporations, NGOs exhibit behavior much more sophisticated and complex than that of any individual
- Many important problems need superhuman intelligence; AI can help
  - improved **collective decision making**
  - more efficient use of **scarce resources**
  - addressing **underserved communities**
  - **climate change**; other societal challenges



# Ethics of AI

## Will a new technology:

- disempower **individuals vs corporations?**  
⇒ user modeling; data mining; fostering addictive behaviors; developmental effects on children
- disempower **individuals vs governments?**  
⇒ facilitate disinformation (deep fakes; bots masquerading as people; filter bubbles); enable qualitatively new military or security tactics
- take **autonomous actions** in a way that obscures responsibility  
⇒ autonomous weapons; self-driving cars; loan approval systems
- disproportionately affect **vulnerable/marginalized groups**  
⇒ automated decision making tools trained in ways that may encode existing biases



# Social Impact

- How will AI technologies **transform society**?
- Will there be a **social backlash** against AI?
  - If so, what will be considered AI?
- This **generation of children** will grow up taking for granted many technologies that strike us as magical
- How will **human relationships** change in the presence of always-available social agents?
- As we are increasingly **augmented by AI**, what are our inherent cognitive/emotional/motivational limitations, beyond which augmentation won't help?



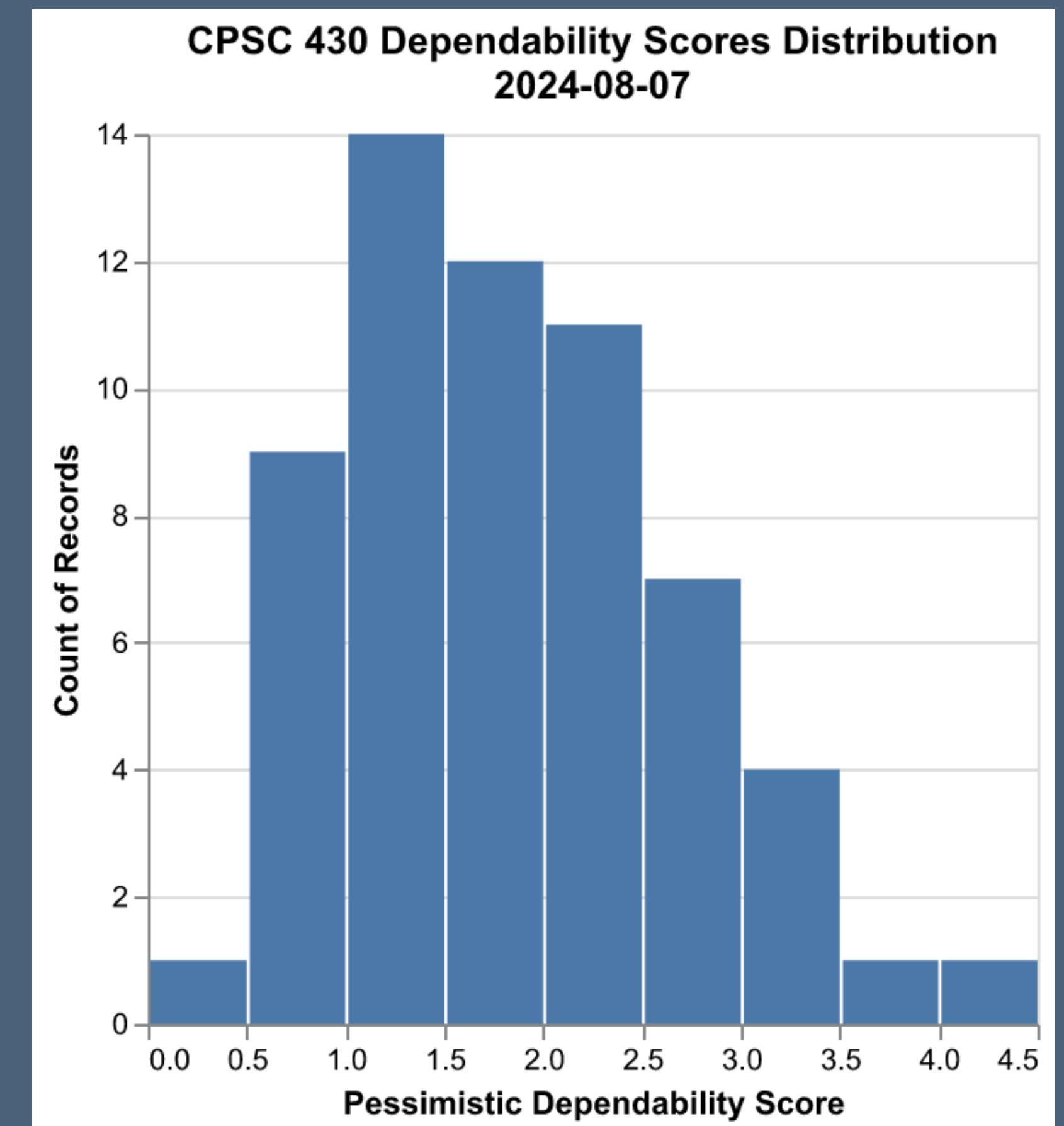
# Ethical consideration of advances in AI

- *Is it wrong to create machines capable of making human labor obsolete? Will humans become demoralized by the presence of vastly more intelligent robots?*
- *How can we ensure that intelligent robots will not be put to an evil purpose by a malevolent human? How can we ensure they do not adopt malevolent purposes themselves?*
- *Is it morally acceptable to create “personal” (self-conscious) AI?*



# Retrospective

- You've written a LOT of essays this term, and made a lot of arguments for a huge range of topics and subjects!
  - Congratulations !
  - Over 80% of you made it over the 1.0 dependability score threshold !



# Retrospective

- Last prompt of the term!
- In this course...
  - Green: “What was the most surprising thing you learned?”
  - Blue: “What was the most upsetting thing you heard/read?”
  - Yellow: “What was the most inspiring/optimistic thing you heard/read?”



Photo by Giftpundits.com from Pexels