# Large Language Models

July, 2023

Tom Mitchell
Tom.Mitchell@cs.cmu.edu
http://www.cs.cmu.edu/~tom
Carnegie Mellon University



## Outline

- What are LLMs, and how do they work?
- Where are they headed?
- How should government leaders think about LLMs…
  - … and US preeminence?
  - … and the workforce?
  - … and education?
  - …

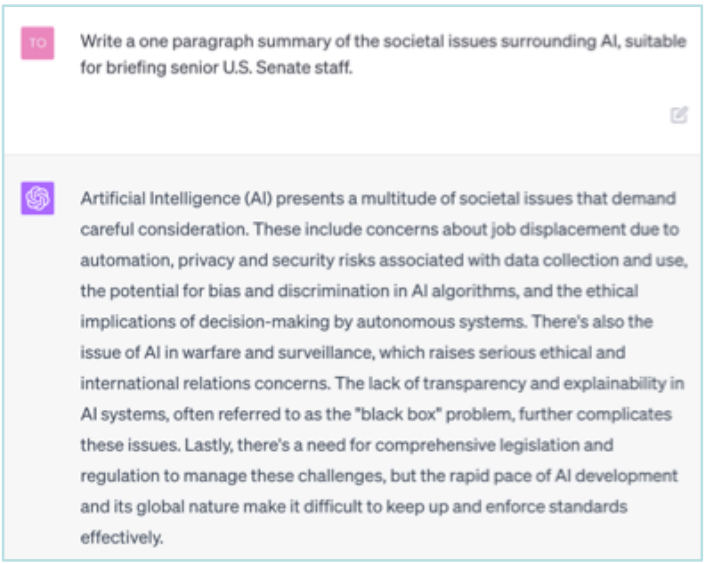
## What Changed to Make Today's LLMs Possible?

- 1950: Write AI computer programs manually
- ~1985: Train computer programs instead, from supervised data
  - (supervised: label the desired output for each training input)
  - 2012: Deep networks dominate computer vision
    - then dominate speech recognition, text analysis, protein folding, …
    - Trend accelerated by improved computer hardware (GPUs), available data
- 2018: Train Natural Language programs from unsupervised data
  - (unsupervised: no human labeling of data)
  - LLMs: trained to predict next word in a sequence
    - $>10^{12}$ training examples from web. Human labeling of $10^6$ very costly
    - Trained on a over a million times the size of most supervised data sets



## Example LLM: GPT-4

- Trained to predict next word, given previous words (unsupervised)
- Neural network with 100s of billions of learned parameter values
- Now underlies ChatGPT, which adds conversational ability
  - "fine tunes" GPT-4 to fit human labeled "prompt-response" pairs
- Surprising abilities!
- But still imperfect
  - Factual errors, "hallucinations", offensive responses, …




Carnegie Mellon University
School of Computer Science

## Producing Mathematical Proofs in Novel Formats

**GPT-4**

**Prompt:** Can you write a proof that there are infinitely many primes, with every line that rhymes?

**GPT-4:**

```
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s
```

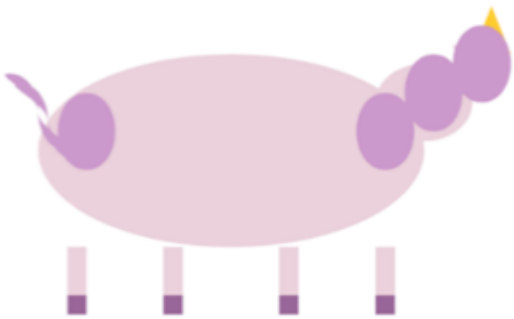
From [Bubeck et al., 2023]



## Knowing shapes and generating drawings

**Prompt:** Draw a unicorn in TiKZ.

**GPT-4:** [Produces LaTeX compiling to following picture.]



From [Bubeck et al., 2023]

## Reasoning About the Physical World

**Prompt:** Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

**GPT-4:**

One possible way to stack the objects onto each other in a stable manner is:

- Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
- Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly. Make sure the eggs are not cracked or broken, and handle them gently to avoid dropping them.
- Place the laptop on top of the eggs, with the screen facing down and the keyboard facing up. The laptop will fit snugly within the boundaries of the book and the eggs, and its flat and rigid surface will provide a stable platform for the next layer. Make sure the laptop is closed and turned off, and do not press any keys or buttons.
- Place the bottle on top of the laptop, with the cap facing up and the bottom facing down. The bottle will add some height and balance to the stack, and its cylindrical shape will reduce the contact area with the laptop. Make sure the bottle is empty and closed, and do not shake or tilt it.
- Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle. Make sure the nail is clean and straight, and do not touch or move it.

[Bubeck et al., 2023]

## Solving a Math Word Problem

Model: GPT-4

TO

Support Mary drives from point A to point B at 100 kph, Bill drives from point B to point A at 50 kph, and both begin their drive at the same time. Will their paths cross? If yes, then where?

Yes, Mary and Bill's paths will cross. They're driving towards each other from different points at the same time, so they're bound to meet somewhere in between, assuming the route is a straight path and there are no obstacles or detours.

To calculate where they will meet, we need to know the total distance between point A and point B. However, based on the information given, we can still discuss how the distance is divided in relation to their speeds.

Assuming the total distance between points A and B is 'D', the time it will take for Mary and Bill to meet can be found by the formula:

Time = Total Distance / Combined Speed

...

Distance covered by Bill = 50 kph x (D / 150) = (1/3)D

So, they will meet at a point 2/3 the distance from point A to point B (which is also 1/3 the distance from point B to point A).

**Solving a Math Word Problem** Model: GPT-4

TO Support Mary drives from point A to point B at 100 kph, Bill drives from point B to point A at 50 kph, and both begin their drive at the same time. Will their paths cross? If yes, then where?

Yes, Mary a
at the same
straight pat

To calculate
point B. How
divided in re

Assuming t
Bill to meet

Time = Tota

Distance

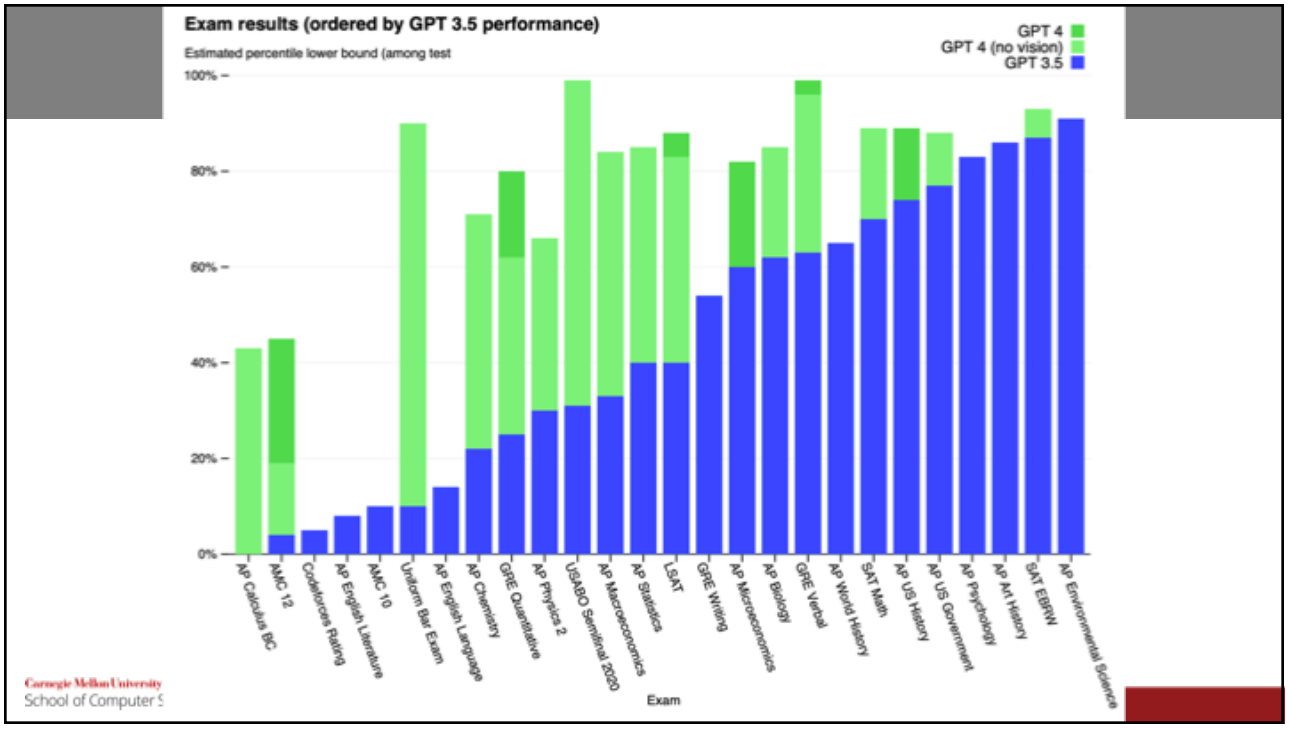
So, they will meet at a point 2/3 the distance from point A to point B (which is also 1/3 the distance from point B to point A).

**Generating a Hint**

TO Suppose a student answers incorrectly that they will meet halfway between points A and B. What hint should we give the student to help them, without giving away the exact answer to the question?

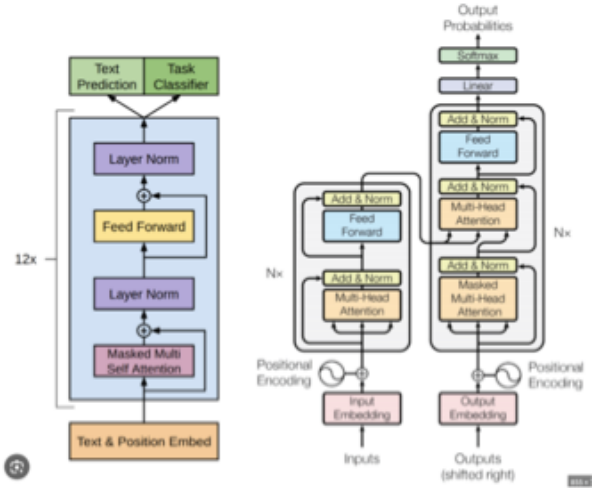
A helpful hint could be to remind the student that Mary and Bill are driving at different speeds. You could say something like this:

"Remember, they're not driving at the same speed, so they won't meet halfway. Think about who is driving faster and how that affects where they will meet. The faster driver will cover more distance in the same amount of time."




Carnegie Mellon University
School of Computer S

# LLMs: The Transformer Architecture




Once trained, predictions are one word at a time:

Inputs: ”Finally got home. I parked my car in …”

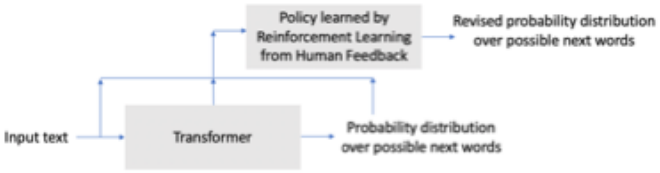
Output prediction: (“the” 0.6, “my” 0.3, “a” 0.05, …)

Inputs: “Finally got home. I parked my car in the …”

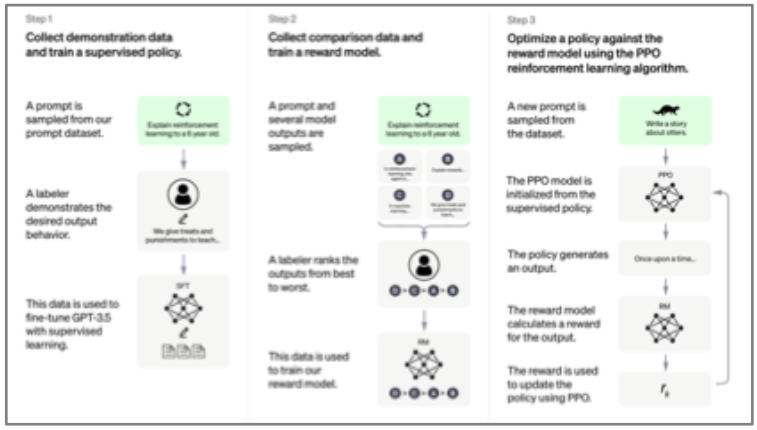
Output prediction:



# ChatGPT – Reinforcement Learning from Human Feedback



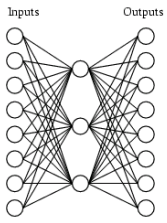
## ChatGPT – Reinforcement Learning from Human Feedback




## How can training to predict the next word produce such impressive behavior?

- Neural networks learn to re-represent their inputs internally, so they can produce correct output.
  - Example →
- In order to predict the next word in the sequence, one needs to represent the ”meaning” of the sequence

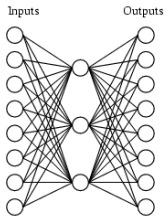



Training Data

| Input | | Output |
|---|---|---|
| 10000000 | → | 10000000 |
| 01000000 | → | 01000000 |
| 00100000 | → | 00100000 |
| 00010000 | → | 00010000 |
| 00001000 | → | 00001000 |
| 00000100 | → | 00000100 |
| 00000010 | → | 00000010 |
| 00000001 | → | 00000001 |

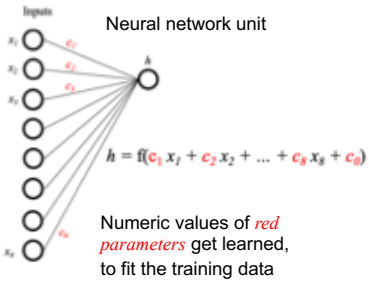
## How can training to predict the next word produce such impressive behavior?

- Neural networks learn to re-represent their inputs internally, so they can produce correct output.
  - Example →
- In order to predict the next word in the sequence, one needs to represent the ”meaning” of the sequence




Training Data

| Input | | Output |
|---|---|---|
| 10000000 | → | 10000000 |
| 01000000 | → | 01000000 |
| 00100000 | → | 00100000 |
| 00010000 | → | 00010000 |
| 00001000 | → | 00001000 |
| 00000100 | → | 00000100 |
| 00000010 | → | 00000010 |
| 00000001 | → | 00000001 |




Neural network unit

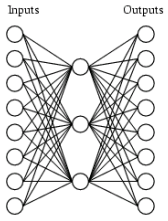
Numeric values of *red parameters* get learned, to fit the training data

Parameters of all network units are learned simultaneously



## How can training to predict the next word produce such impressive behavior?

- Neural networks learn to re-represent their inputs internally, so they can produce correct output.
  - Example →
- In order to predict the next word in the sequence, one needs to represent the ”meaning” of the sequence




Training Data

| Input | | Output |
|---|---|---|
| 10000000 | → | 10000000 |
| 01000000 | → | 01000000 |
| 00100000 | → | 00100000 |
| 00010000 | → | 00010000 |
| 00001000 | → | 00001000 |
| 00000100 | → | 00000100 |
| 00000010 | → | 00000010 |
| 00000001 | → | 00000001 |

Learned Internal Representation of Each Input

| Input | | Hidden Values | | | | Output |
|---|---|---|---|---|---|---|
| 10000000 | → | .89 | .04 | .08 | → | 10000000 |
| 01000000 | → | .01 | .11 | .88 | → | 01000000 |
| 00100000 | → | .01 | .97 | .27 | → | 00100000 |
| 00010000 | → | .99 | .97 | .71 | → | 00010000 |
| 00001000 | → | .03 | .05 | .02 | → | 00001000 |
| 00000100 | → | .22 | .99 | .99 | → | 00000100 |
| 00000010 | → | .80 | .01 | .98 | → | 00000010 |
| 00000001 | → | .60 | .94 | .01 | → | 00000001 |

↑

Learned binary encoding
100, 001, 010, 111, 000, 011,
101, 110
of the eight possible inputs!

Carnegie Mellon University
School of Computer Science

## Image synthesis, voice synthesis, and soon multi-modal models

Image produced from input text "A falcon landing on a camel in an oasis." Produced by Microsoft Image Creator, based on the OpenAI DALL-E system.

Image produced from input text "An art piece depicting the feeling of regret in a winter setting, with a focus on snow and frozen landscapes." Produced by the Midjourney system.

Image produced from input text "An art piece depicting the feeling of regret in a winter setting, with a focus on snow and frozen landscapes." Produced by the Stable Diffusion XL system.



## Rapid recent LLM developments

- Nov 2022. OpenAI releases ChatGPT, based on GPT3.5
- Jan 2023. ChatGPT has over 100M users
- Feb 2023. Microsoft releases Bing search incorporating ChatGPT
- Feb 2023. Meta releases LLM LLaMA as open source
- Mar 2023. OpenAI releases GPT4
- Mar 2023. Google releases Bard
- Mar 2023. Baidu announces Chinese Ernie LLM
- Mar 2023. Google's Med-PaLM 2 achieves 86% accuracy on US Medical Licensing Exam questions
- Mar 2023. Khan Academy announces Khanmigo K-12 tutor based on GPT4
- Mar 2023. Bloomberg announces their 50B parameter LLM for finance
- Apr 2023. ChatGPT offers "plugins"
- Apr 2023. Alpaca 7B LLM derived from LLaMA is retrained for under $1000
- Jul 2023. Inflection AI releases their Inflection-1 LLM

## Significant AI Impacts Already

- Software productivity
  - Microsoft Co-Pilot → up to 50% productivity improvement in software development time
- Basic science, drug design
  - DeepMind's Alpha-Fold is revolution in predicting 3D protein structure, including SARS-CoV-2 spike protein, helping in design of COVID vaccines
- Transportation safety
  - Computer perception provides collision avoidance, out of lane drift warning
- Healthcare, radiology
  - Over 30% of radiologists use AI in their practice: image analysis, risk stratification, automated report generation
- Synthetic videos, audios
  - July 2023 actors' strike



## What might be coming in the next few years? *

- Small number of highly intelligent multimodal LLMs 10x size of today's
- Open source, smaller LLMs that are downloadable and proliferate worldwide
- Many specialized LLMs for healthcare, legal work, finance, …
- LLMs using diverse plugins (calculators, route planners, databases, LLMs,…)
- Multi-modal models increasingly able to act in the physical world
- Personal LLMs for everybody, knowledgeable about each individual user

* Disclaimer: Nobody can predict the future

## What might be coming in the next few years? *

- Increasingly ubiquitous natural language or mixed language-vision interfaces to *nearly all* hardware devices
- Progress in digital watermarking to distinguish synthetic from original media, but with persistent challenges likely to remain
- Progress in reducing hallucination problems, but continuing trust issues
- Progress in censoring LLMs to control their outputs and uses
- Rate of progress will depend on barriers to entry, business models, government regulation, whether universities remain significant factor in AI progress, access to data, …

* Disclaimer: Nobody can predict the future



## How to think about Government Response to AI

- AI will have unpredictable uses and impacts
  - → requires a flexible, adaptable, government response
- Most regulation will be at the application level
  - Regulating self-driving vehicles will be unlike regulating AI-based medical diagnosis
  - Rely on existing organizations where possible, FDA, NHTSA, …
  - Perhaps an 'AI Czar' in the White House
- Perhaps some regulation/norms around more general abilities
  - E.g., transparency reporting requirements for LLMs, analogous to Nutrition Facts labels on foods?
  - Build on Industry-established transparency norms via AI 'system cards'
- Stand up a CERT-style 3rd party organization to
  - Detect and communicate adverse AI-related events (e.g., widely circulating deep fakes)
  - Collect and disseminate best practices

## How to think about AI impact on the Workforce

- Jobs are bundles of tasks
- AI impacts are usually at the level of individual tasks
- AI impact can either automate or assist worker
  - this is a design choice we can make
  - will LLMs act more often to advise/assist than to replace?
- AI-driven productivity improvements can decrease or increase demand for workers
  - Jet engine technology increased productivity of pilots
    → led to lower costs, increased demand, more jobs for pilots
- We can help workers by
  - Tracking and communicating shifts in demand for different skills
  - Providing just-in-time continuing education for job shifts



[See National Acedemy of Sciences report on AI and the Workforce, 2017, 2023 forthcoming]

## How to think about International Competition in AI

- Keeping technical ideas inside national borders will not work
- U.S. research universities led AI development for decades
  - But recent major advances arose mostly from large corporations investing >> $100M
  - Will U.S. retain its university research advantage? → Consider funding NAIRR
- One wildcard: will open source community find a way to lower barrier to entry?
- International collaboration: Distinguish zero-sum vs. win-win opportunities
  - Zero sum → Compete: Military applications, AI for manufacturing
  - Win-win → Collaborate: AI for managing pandemics, AI for cleaner air
- China tracks U.S. research publications, but we don't track Chinese
  → Consider mandate for Library of Congress to translate/disseminate all Chinese AI papers

# How to think about AI and education

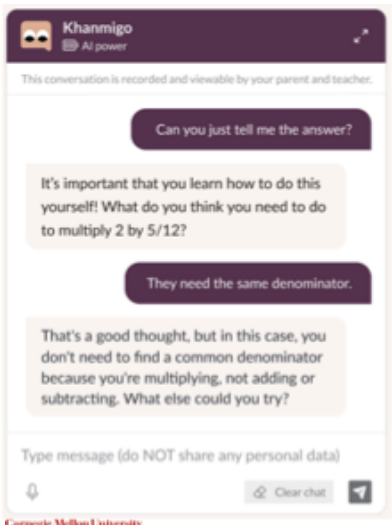
AI has potential to improve <u>how</u> we educate

- Already in use
  - Multiple online education platforms with > 1 million students taught
    - These can already adapt to needs of individual students
    - and already learn to teach better from observing earlier students
- Future potential seems great, but not yet proven
  - LLM's enable more natural interactions (e.g., Khanmigo)
  - Immersive VR, virtual agents offer more engaging visual learning environments

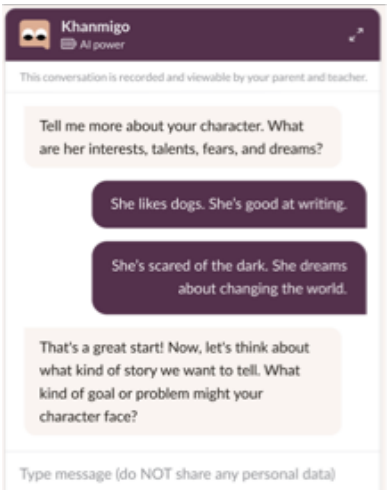


# LLMs and Education -- Khanmigo

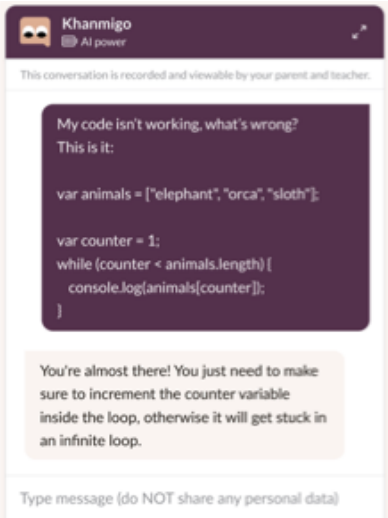
Math




Story writing




Computer Programming




[ https://www.khanacademy.org/khan-labs#khanmigo ]

## How to think about AI and education

AI has potential to improve how we educate

- Already in use
  - Multiple online education platforms with > 1 million students taught
    - These can already adapt to needs of individual students
    - and already learn to teach better from observing earlier students
- Future potential seems great, but not yet proven
  - LLM's enable more natural interactions (e.g., Khanmigo)
  - Immersive VR, virtual agents offer more engaging visual learning environments

AI might change the nature of what we should teach

- Imagine AI makes it easy to obtain any factual information, summaries of documents, … to automate many routine operations, …
- What does next student generation need to learn?
  - Learn how to use AI as a tool
  - Focus making complex decisions by requesting multiple inputs, teamwork skills
  - Broad education including STEM, social science, history, art, business, …?



## How to think about Privacy and AI

- There is a privacy-effectiveness tradeoff in AI systems
- Research is improving the entire tradeoff curve:
  - Federated learning leaves private data at the edge
  - Homomorphic encryption operates on encrypted data
  - Differential privacy methods distort training data to hide individuals
- Governments are acting on privacy
  - E.g., European GDPR, US HIPAA, FERPA, …
- But privacy has a cost. Need to balance:
  - Privacy: e.g., keeping individual medical data private
  - Inclusion: e.g., assuring AI is trained on medical data representing all groups
  - Automony: e.g., right of individual to chose how data about them is used

# How to think about bias in AI systems

We humans can have biases we're unaware of

- They show up in how we make decisions, how we design our systems, …

Machine learning systems are often trained on human generated data, which can exhibit such biases

- E.g., historical decisions about which loan applications to approve
- ML systems learn to make new decisions by fitting the training data
- Therefore, they will learn and propagate biases implicit in training data
- Ongoing technical research attempts to minimize learned biases
  - By using better, unbiased training data
  - By changing the learning objective that defines 'best fit' to the data
  - But some combinations of 'fairness' objectives cannot be simultaneously achieved simultaneously
  - Ongoing debate about defining 'fairness' precisely