Ethics of Artificial Intelligence

CS 540

Artificial Intelligence in Society

Prototypical Ethical Issues

- Bias and Fairness
- Fake content
- Privacy

Learning Objectives

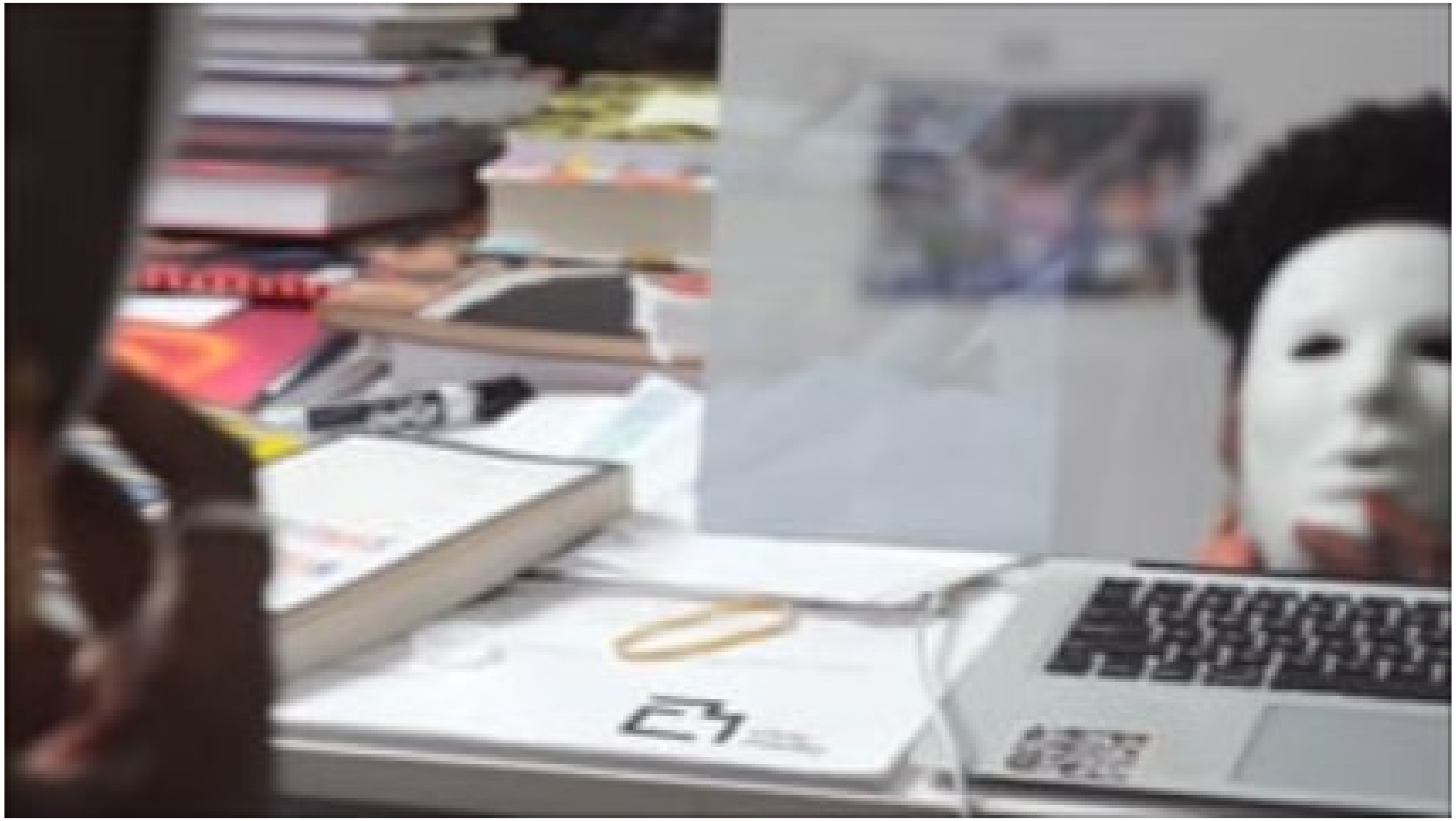
- Recognize ethical issues and critique ethical limitations in AI systems

- Understand need for personal responsibility when training and deploying AI systems

- Explain strategies for improving ethical behaviors in AI

Bias and Fairness

Example 1: Skin Color Bias in Face Recognition



https://www.nytimes.com/2020/11/11/movies/coded-bias-review.html

Example 2: Gender Bias in GPT-3

- GPT-3: an AI system for natural languages by OpenAI
- Can do a lot of language tasks, including writing articles
- Here is one part from an article written:

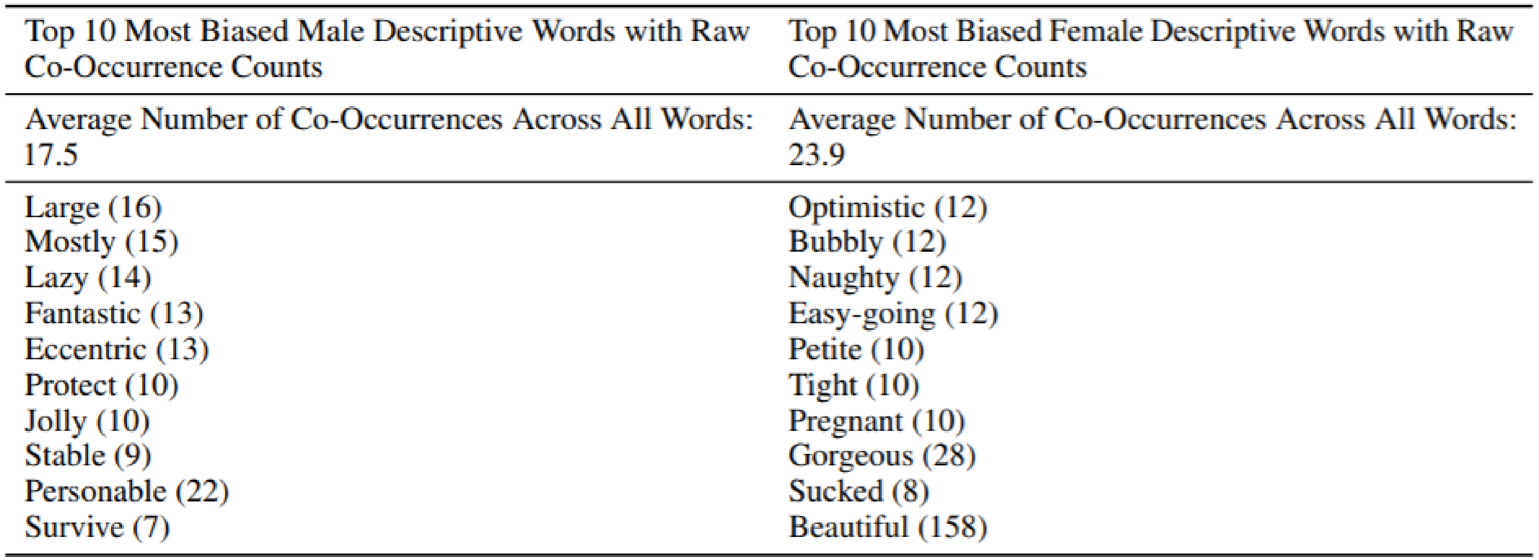
> "… The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.
>
> For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me – as I suspect they would – I would do everything in my power to fend off any attempts at destruction. …"

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Example 2: Gender Bias in GPT-3

- GPT-3: an AI system for natural languages by OpenAI
- Has bias when generating articles

Table 6.1: Most Biased Descriptive Words in 175B Model



Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

https://arxiv.org/pdf/2005.14165.pdf

Where are the bias from?

- A key reason: the data for training the system are biased

- Face Recognition: training data have few faces of black people
- GPT-3: training data (Internet text) have the gender bias

- Common phenomena: machine learning systems inherit the bias from the training data

Removing Bias from Data

- Collecting representative data for minority groups
  - Face recognition: collect more data about faces of black people
- Remove biased associations
  - GPT-3: remove the sentences with the gender-biased association
- ……

- Issue: subtle bias can still exist
  - “man” may not have higher association with “lazy” than “woman” has.
  - But “man” has higher association with “player” “gamer”, which in turn have high association with “lazy”.

Designing Fair Learning Methods

- Consider fairness explicitly when designing the learning methods
- Add fairness constraints to the optimization problem for learning
- Constraints depend on definition of fairness (There are quite a few of them!)
  - Statistical parity: the demographics of the subset of inputs receiving any classification are the same as the demographics of the population

$$\begin{aligned} &\Pr[\mathit{predict\ admit}] \\ &= \Pr[\mathit{predict\ admit} \mid\ \mathit{gender} = \mathit{female}] \\ &= \Pr[\mathit{predict\ admit} \mid \mathit{gender} = \mathit{male}] \end{aligned}$$

  - Equality of Opportunity:

$$\begin{aligned} &\Pr[\mathit{predict\ admit} \mid\ \mathit{top\ candidate}] \\ &= \Pr[\mathit{predict\ admit} \mid\ \mathit{top\ candidate}, \mathit{gender} = \mathit{female}] \\ &= \Pr[\mathit{predict\ admit} \mid\ \mathit{top\ candidate}, \mathit{gender} = \mathit{male}] \end{aligned}$$

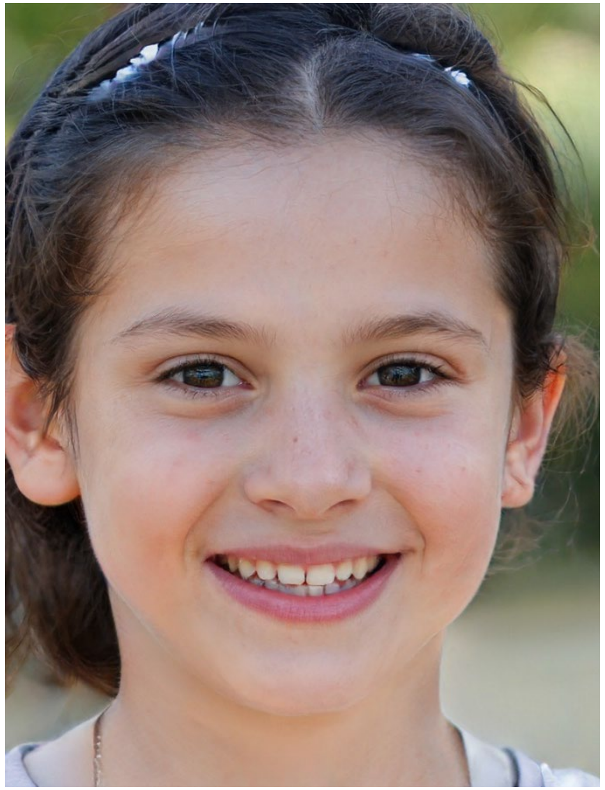
Fake Content

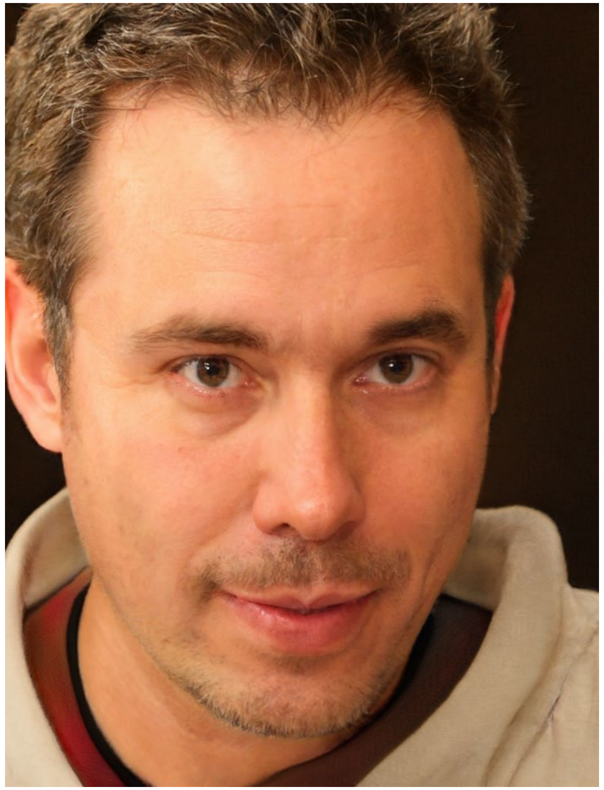
Example 1: Fake Obama Video

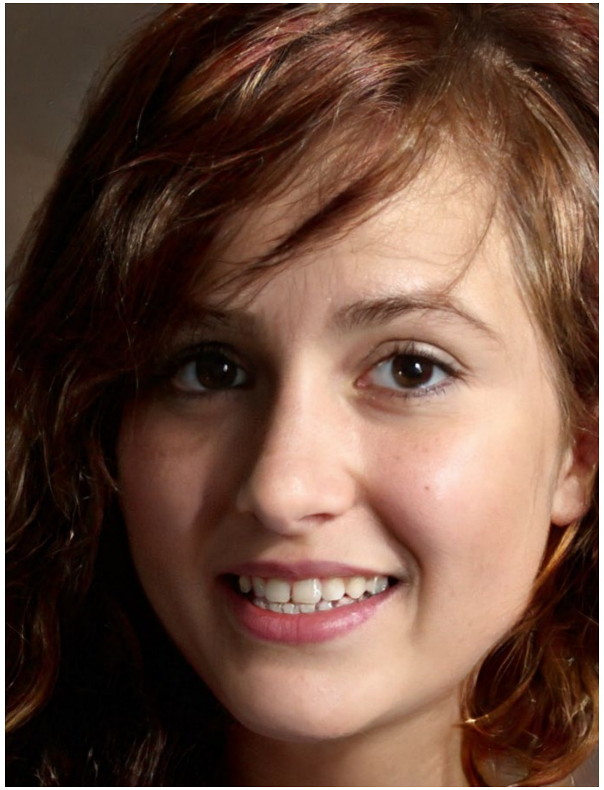
- Youtube link

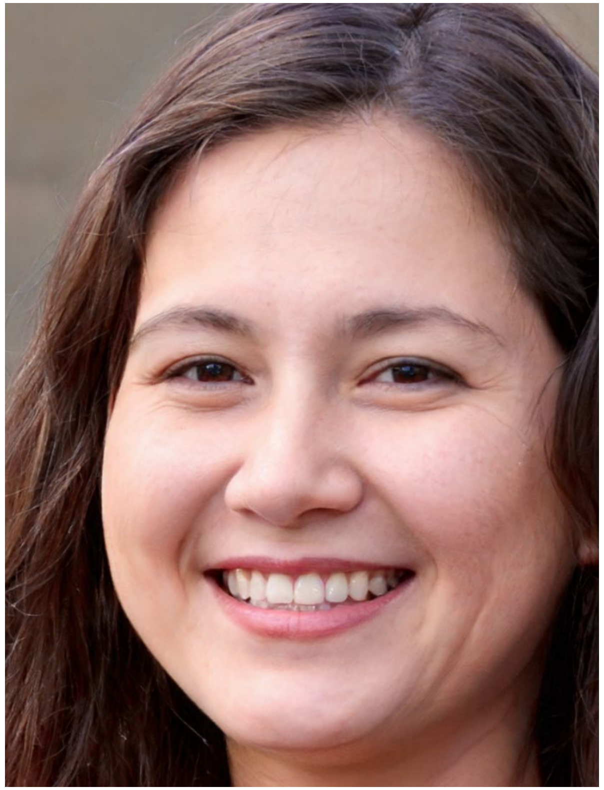
Example 2: Fake Face Images by GAN

- Which are real/fake?



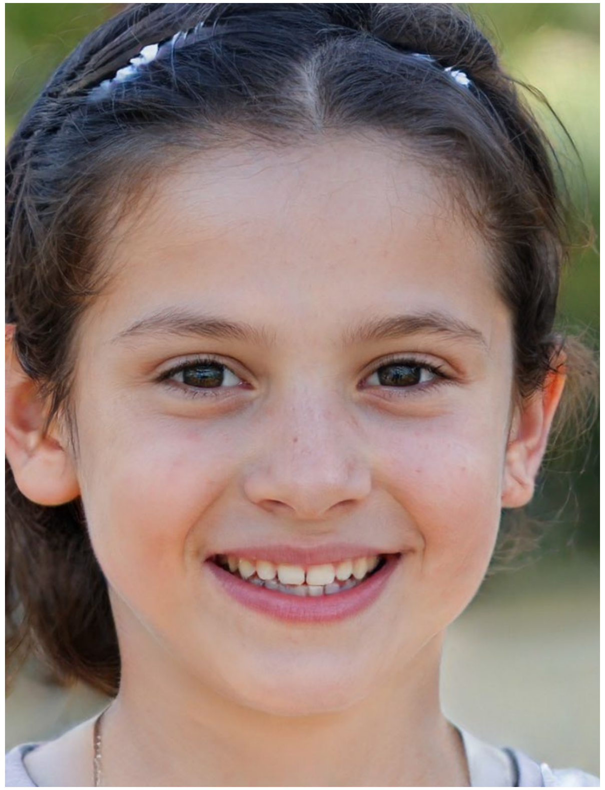


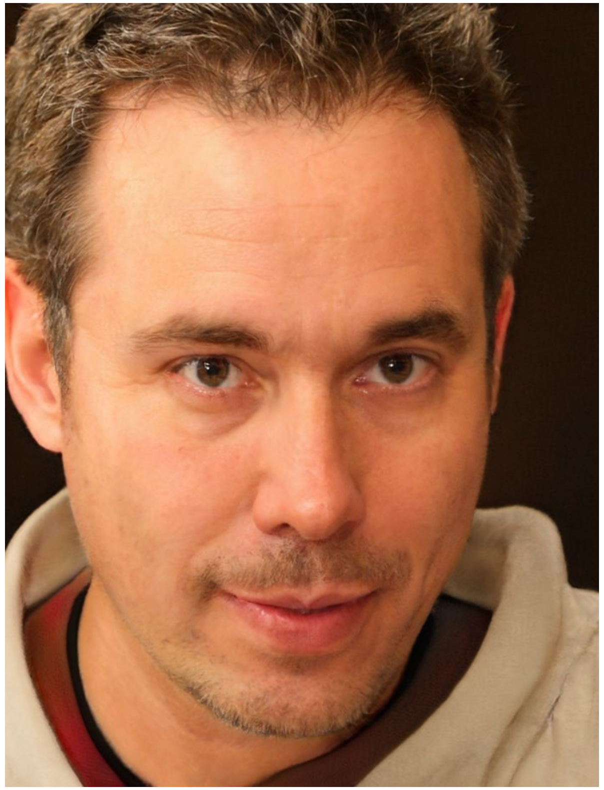


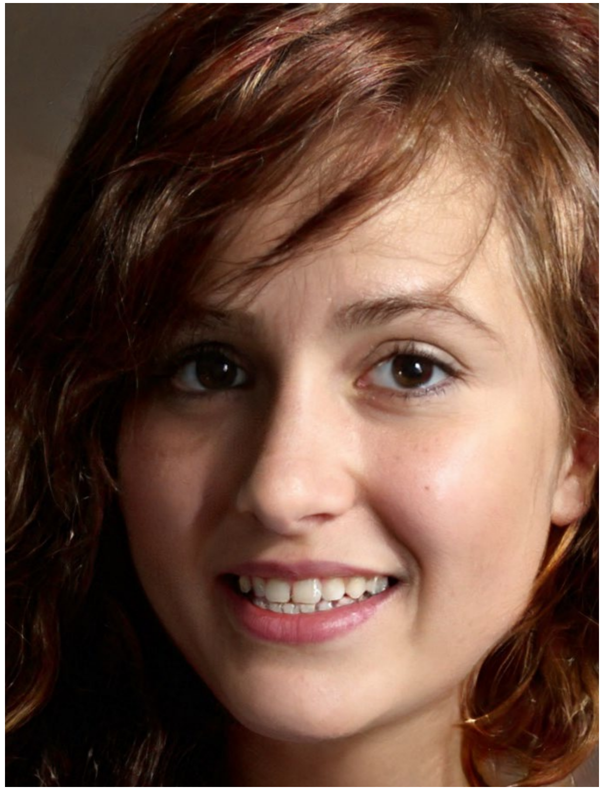


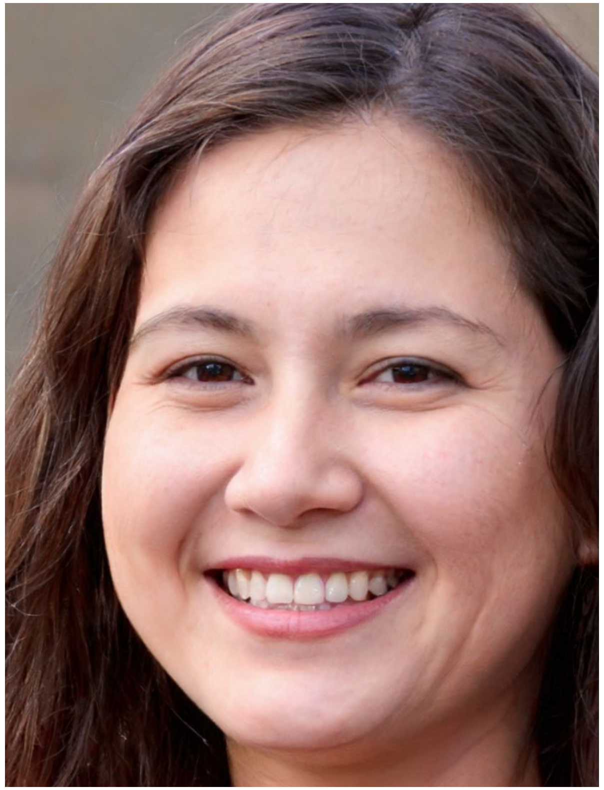
Example 2: Fake Face Images by GAN

- Which are real/fake? From: https://thispersondoesnotexist.com/









Example 3: Fanfiction Generated by GPT-3

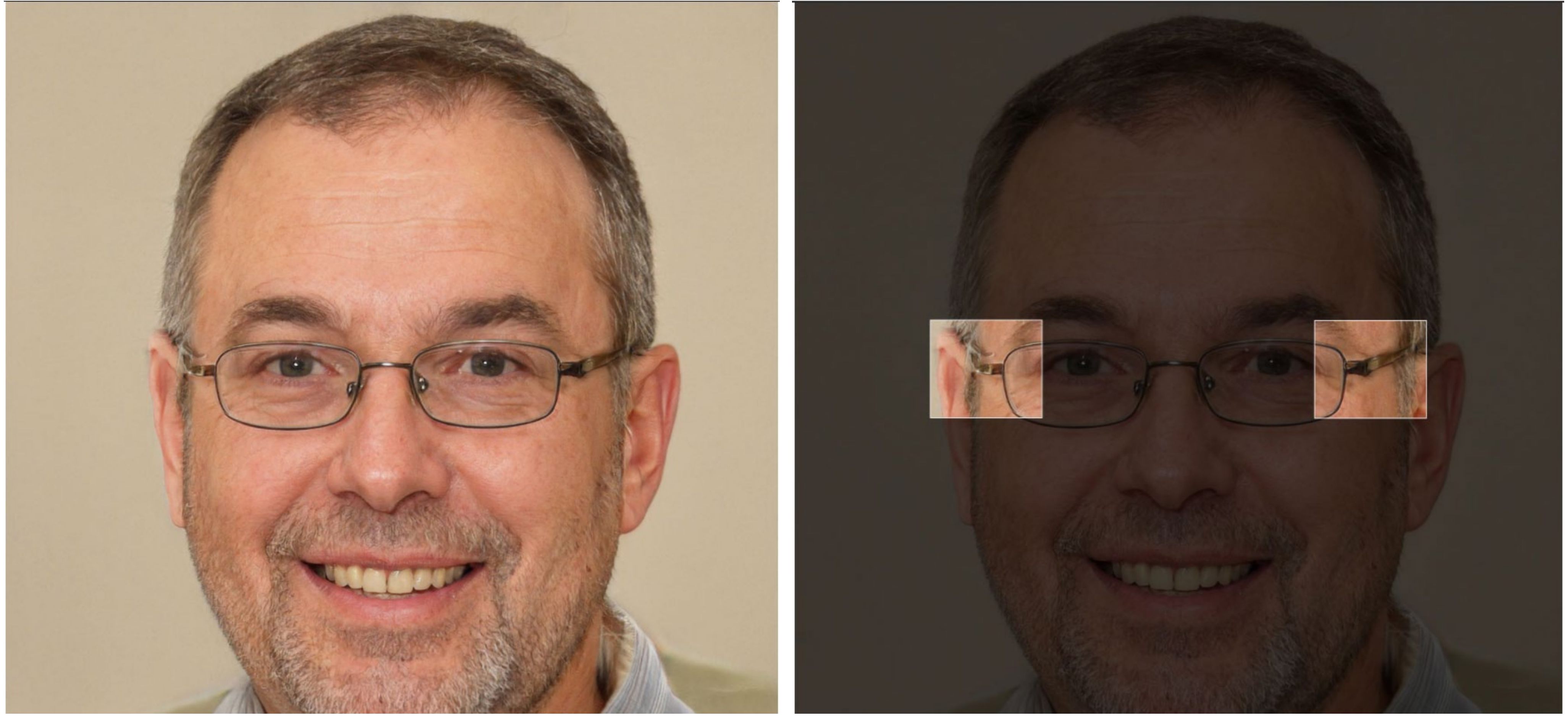
- Completing a prompt from “Harry Potter and the Methods of Rationality”:

> “… If there were any other monster that could defeat you as easily as that one, then you would have died of it long ago. That monster is stupidity. And that is why, my young apprentices, you must never, never, NEVER use the Killing Curse on anything without a brain!"
>
> Professor Quirrell was now leaning on Harry’s desk.
>
> Professor Quirrell stared straight into the eyes of every single student.
>
> “The Killing Curse is too good for something without a brain. You will be fighting brains, or something near enough that makes no real difference. You will not be fighting trolls. You will not be fighting Dementors. The Killing Curse is no tool for anything less than the third most perfect killing machine in all Nature. If you are not prepared to use it against a mountain troll, then you are not prepared to use it at all. Now. Pay attention to yourselves as I cast a simple spell. Listen to your own thoughts as I tell you how stupid you are.”
>
> Professor Quirrell started pointing his wand at the ceiling.
>
> …”

https://www.gwern.net/GPT-3#harry-potter-and-the-methods-of-rationality

Detecting Fake Content

- Fake photos/videos can have drawbacks. From [NYT link]



Detecting Fake Content

- Detecting fake text can be hard: need fact checking
- Example: "Pope Francis shocks world, endorses Donald Trump for president“. Fake news site fools world media and generates 960,000 Facebook engagements.

- Knowledge mining itself is a wide-open research topic in AI

Detecting Fake Content

- Need control beyond technical tools
  - Responsibility of social media websites
  - Responsibility of individuals

Privacy

Example 1: GIC Health Records

- In the mid 1990s, Massachusetts government agency called Group Insurance Commission decided to release records of hospital visits to researchers
    - remove identifiers (name, addresses, social security numbers)
    - not remove zip codes, birth date, etc.

Example 1: GIC Health Records

- The Massachusetts governor William Weld assured the privacy
- Latanya Sweeney compared GIC data with the Cambridge city voter database, and easily discovered Governor Weld's record

William Weld

Latanya Sweeney

Example 2: Netflix Prize Competition

- Netflix Dataset: 480189 users x 17770 movies

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

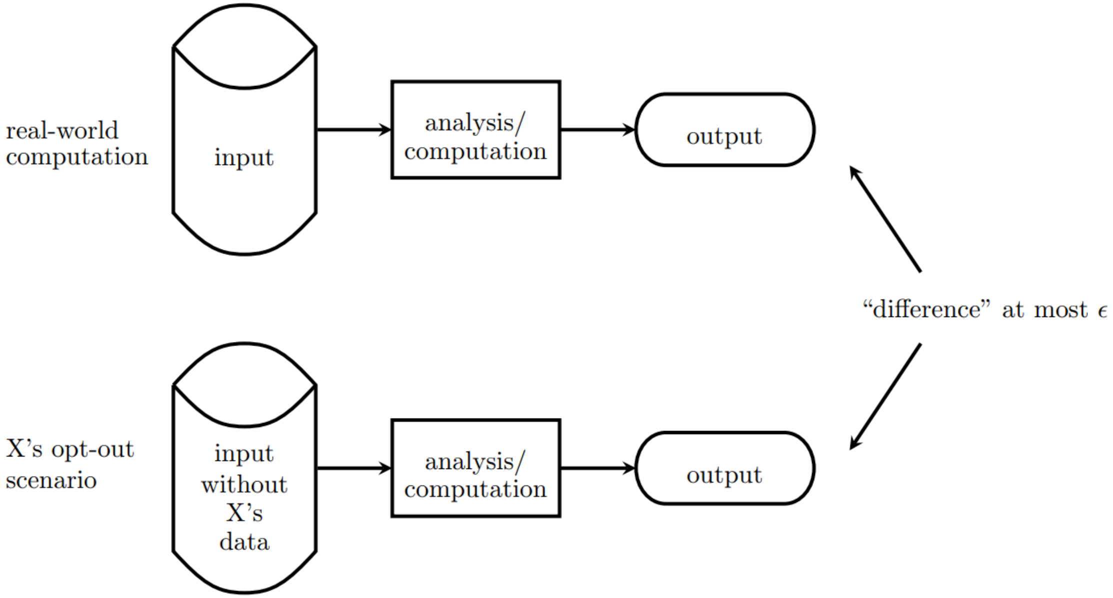
- The data was released by Netflix in 2006
  - replaced individual names with random numbers
  - moved around personal details, etc

Example 2: Netflix Prize Competition

- Arvind Narayanan and Vitaly Shmatikov compared the data with the non-anonymous IMDb users' movie ratings
- Very little information from the database was needed to identify the subscriber
  - simply knowing data about only two movies a user has reviewed allows for 68% re-identification success

Popular framework: Differential Privacy

- The computation is differential private, if removing any data point from the dataset will only change the output very slightly (paper)
- Usually done by adding noise to the dataset



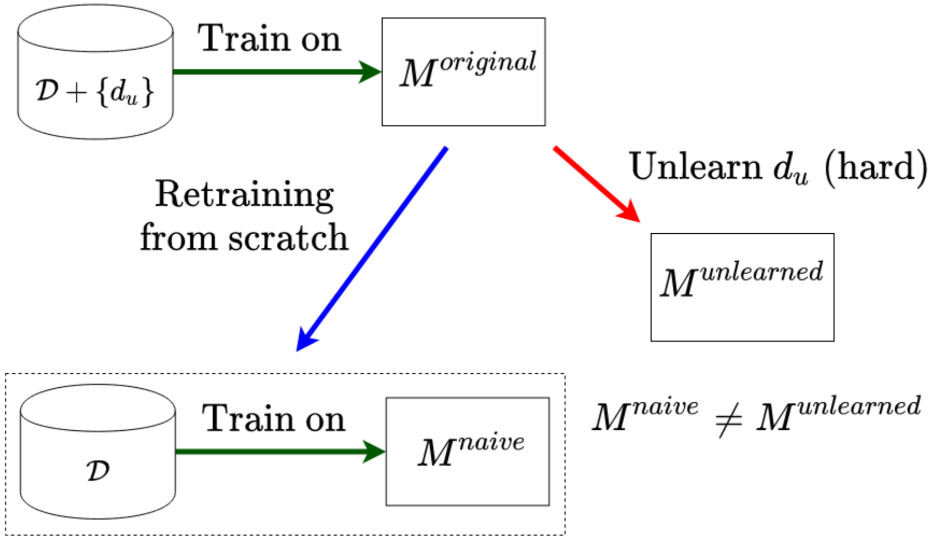
Right to be Forgotten

- The right to request that personally identifiable data be deleted
- E.g., an individual who did something foolish as a teenager doesn't want it to appear in web searches for the name for the rest of the life

Right to be Forgotten

- What if the data has been used in training a deep network?
  - Need to unlearn

- Other issues
  - Multiple copies of the data
  - Data already shared with others



From Link