Natural Language Processing

Ethics in AI

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Our communication is fundamentally a social activity

The common misconception is that language has to do with words and what they mean. It doesn't. It has to do with **people** and what they mean.

Herbert H. Clark & Michael F. Schober (1992)
Asking Questions and Influencing Answers

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

What is Ethics?

Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a practical discipline.

Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one's life."

Introduction to Ethics, John Deigh

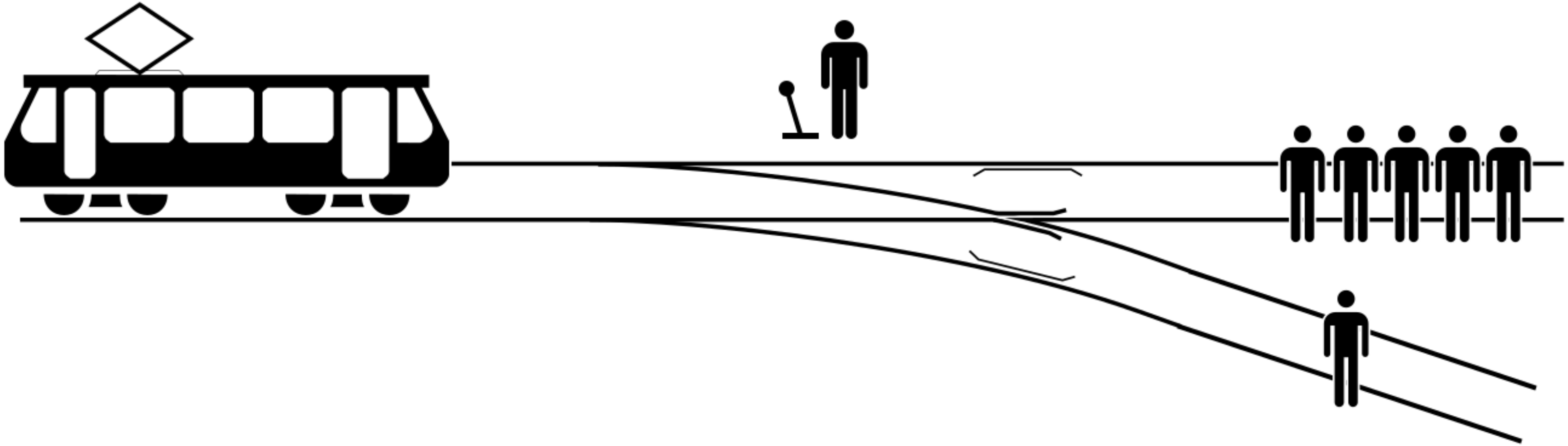
What is Ethics?

It’s the **good** things

It’s the **right** things

The Trolley Dilemma

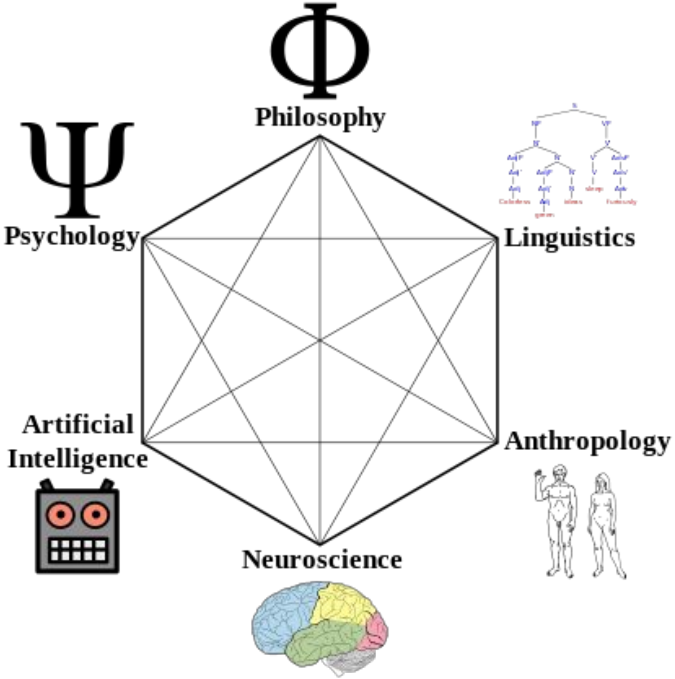
Should you pull the lever to divert the trolley?



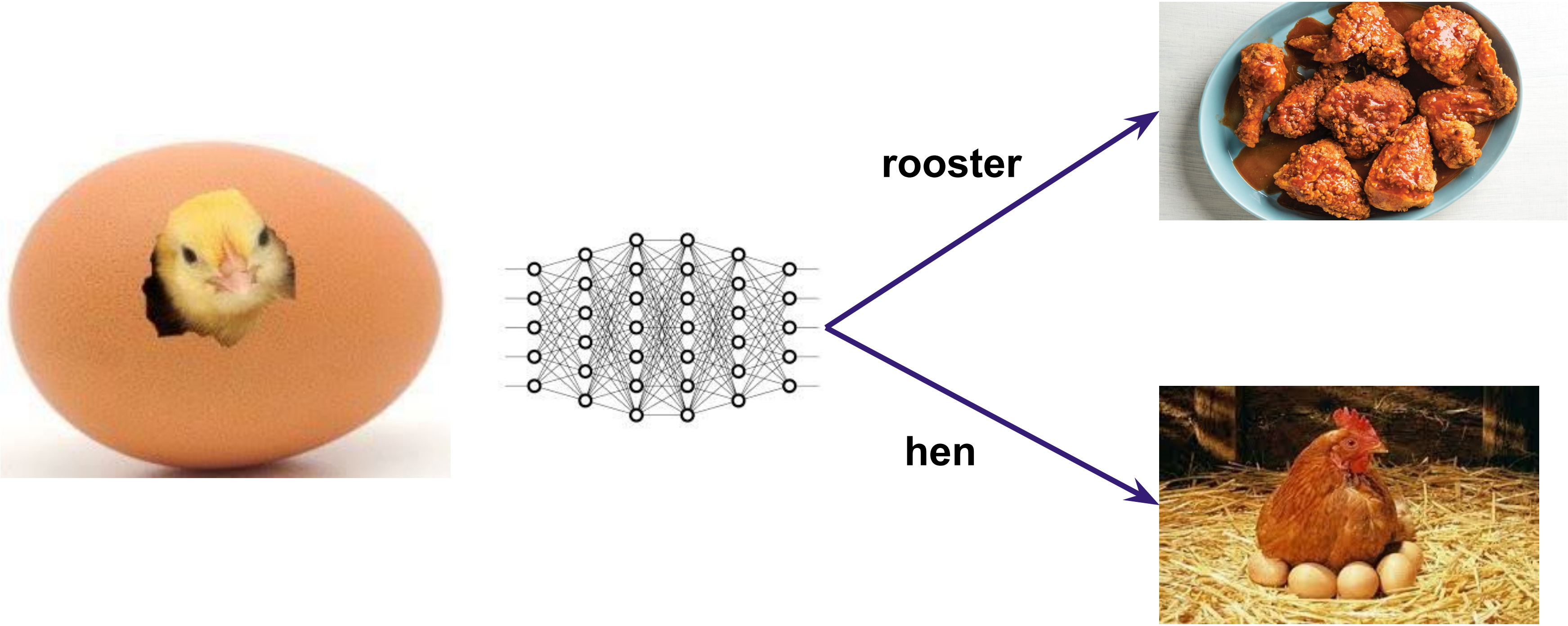
[image from Wikipedia]

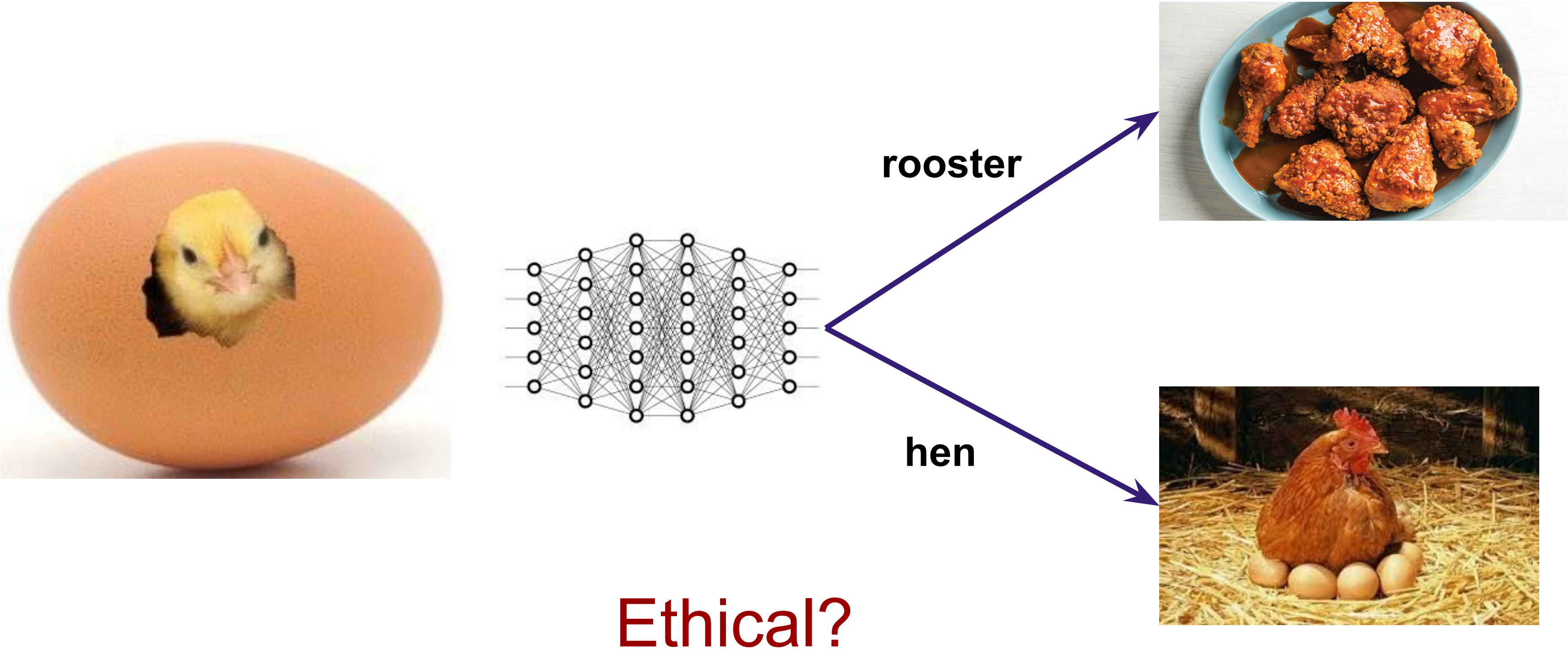
Ethics in AI

- Data origin and ownership
- Social bias in AI
- Algorithmic fairness
- Privacy risks and protection
- Mis-use of AI:
  - disinformation, opinion manipulation
- AI for good
  - content moderation
  - assistive technologies, disaster response
- Societal impacts of AI
  - environmental
  - economic, educational, policy impacts
- ….

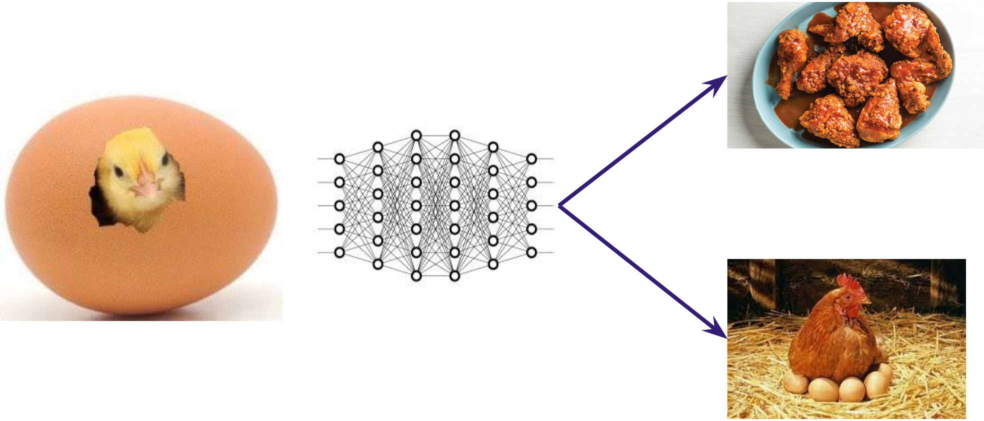


The Chicken dilemma



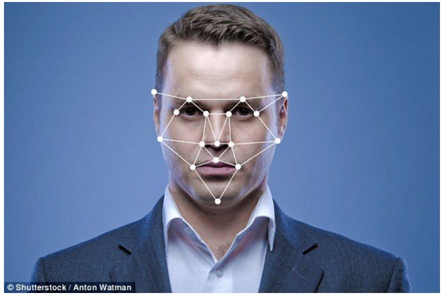


Ethical?

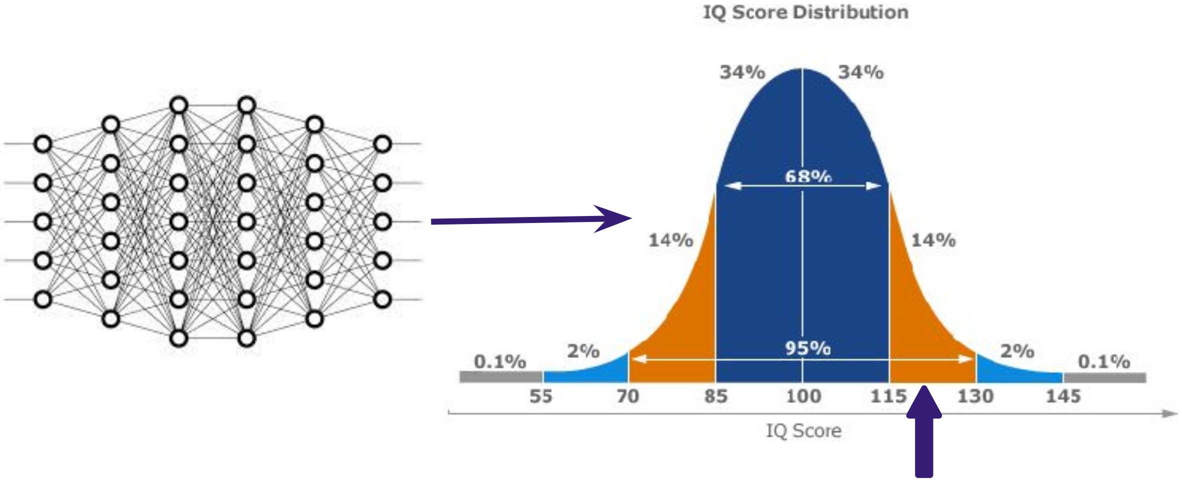


- Ethics is inner guiding, moral principles, and values of people and society
- There are gray areas. We often don’t have easy answers.
- Ethics changes over time with values and beliefs of people
- Legal ≠ Ethical

The IQ dilemma



© Shutterstock / Anton Watman



→ **I**ntelligence **Q**uotient: a number used to express the apparent relative intelligence of a person

The IQ dilemma

We can train a classifier to predict people’s IQ from their photos & texts. Let’s discuss whether it is ethical to build such a technology and what are the risks.

- Understanding the research q and stakeholders: Who could benefit from such a classifier?

The IQ dilemma: the ethics of the research question

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- Who could benefit from such a classifier?
- Understanding the risks: Let's assume for now that the classifier is 100% accurate.
  Who can be harmed from such a classifier? How can such a classifier be misused?

The IQ dilemma: understanding the risks

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- Who could benefit from such a classifier?
- Who can be harmed from such a classifier? How can it be misused?
- What are the pitfalls/risks in the current solution?
    - Example: Our test results show 90% accuracy
        - We found out that white females have 95% accuracy
        - People with blond hair under age of 25 have only 60% accuracy

The IQ dilemma: understanding the responsibility

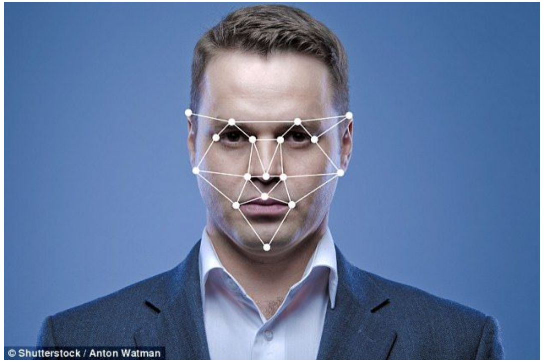
We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- Who could benefit from such a classifier?
- Who can be harmed from such a classifier? How can it be misused?
- What are the pitfalls/risks in the current solution?
- Who is responsible?
  - Researcher/developer? Advisor/manager? Reviewer? The IRB? The University? Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

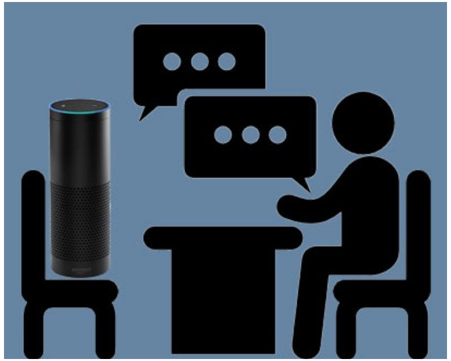
IQ classifier - risks

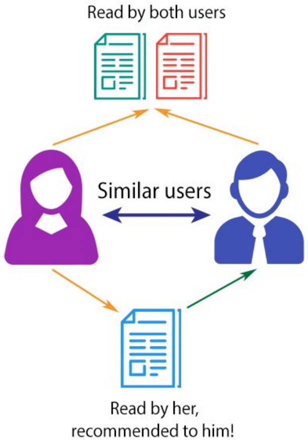
- Research question is problematic: attempts to predict IQ are done to approximate intelligence and future success, but IQ is not a good proxy
- IQ tests are known to be racially and socio-economic status (SES)-biased
- Also, the data used to train an IQ classifier will likely have many biases
- AI systems are likely to pick up on these biases and spurious correlations between intelligence metrics and linguistic features of racial or SES groups
- Error in such a classifier can have direct negative impact on people



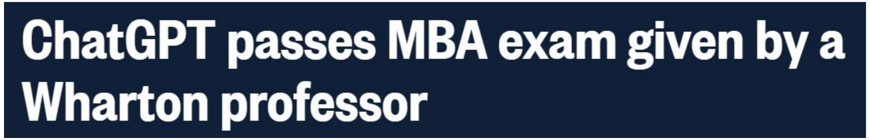
© Shutterstock / Anton Watman

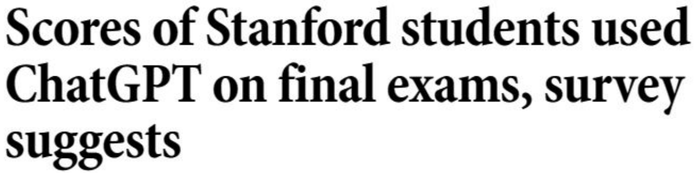
AI and people

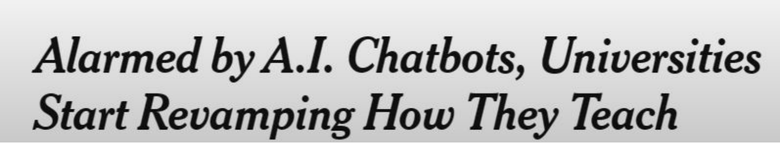


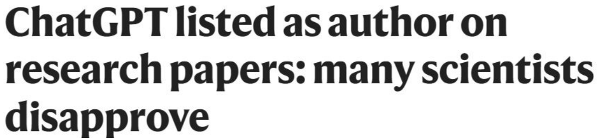


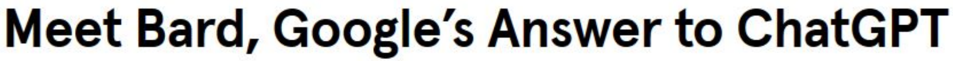
AI and people


ChatGPT passes MBA exam given by a Wharton professor


Scores of Stanford students used ChatGPT on final exams, survey suggests


Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach


ChatGPT listed as author on research papers: many scientists disapprove


Meet Bard, Google's Answer to ChatGPT

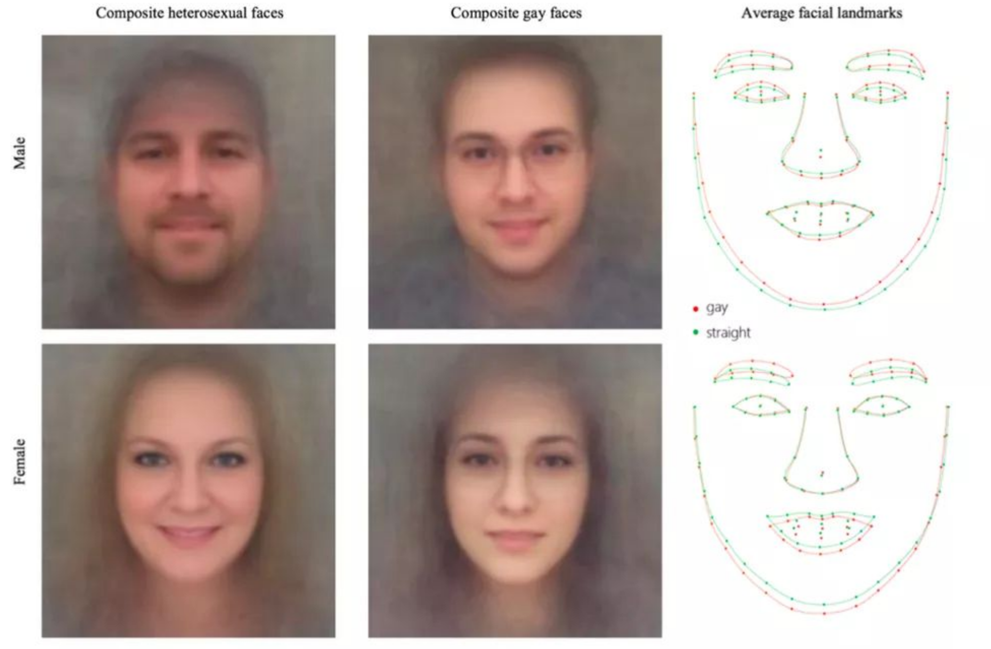
What are important ethical questions to ask in development and deployment of AI systems?

Why do these issues become especially relevant now?

- Data: the exponential growth of user-generated content
- Technological advancements: machine learning tools have become powerful and ubiquitous

What are important ethical questions to ask in development and deployment of AI systems?

A recent study: the “AI Gaydar”, 2017



A recent study: the “AI Gaydar”

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women
- Motivation for the study: expose a threat to the privacy and safety of gay men and women

Let’s discuss…

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

What went wrong?

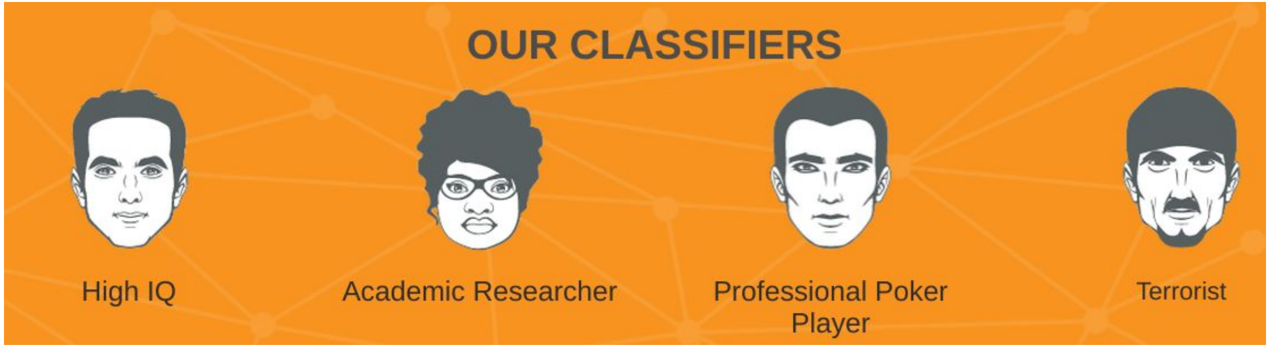
Questioning the ethics of the research question

- Identification of sexual orientation from facial features

Sexual orientation classifier - who can be harmed?

- In many countries being gay person is prosecutable (by law or by society) and in some places there is even death penalty for it
- It might affect people’s employment; family relationships; health care opportunities;
- Personal attributes like gender, race, sexual orientation, religion are social constructs. They can change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.

Dual framing and dual use in predictive analytics



“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities.
As a result, it is becoming ever more challenging to detect anonymous threats in public places such as airports, train stations, government and public buildings and border control. Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations. “

Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Data privacy

- Photos downloaded from a popular American dating website

Data privacy

- Photos downloaded from a popular American dating website

Questions to ask:

- Is it legal to use the data?
- However, legal ≠ ethical. Who gave consent? Even if the data is public, public ≠ publicized. Does the action of publicizing the data violate social contract?

Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Data biases

- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Questions to ask:

- Is the dataset representative of diverse populations? What are gaps in the data?
  - Only white people who self-disclose their orientation, certain social groups, certain age groups, certain time range/fashion; the photos were carefully selected by subjects to be attractive
- Is label distribution representative?
  - The dataset is balanced, which does not represent true class distribution.

⟶ this dataset contains many types of biases

Method

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification

Algorithmic biases

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification

Questions to ask:

- Does model design control for biases in data and confounding variables?
- Does the model optimize for the true objective?
- There is a risk in using black-box model which reasons about sensitive attributes, about complex experimental conditions that require broader world knowledge. Does the model facilitate analyses of its predictions?
- Is there analysis of model biases?
- Is there bias amplification?
- Is there analysis of model errors?

Evaluation

- Accuracy: 81% for men, 74% for women

The cost of misclassification

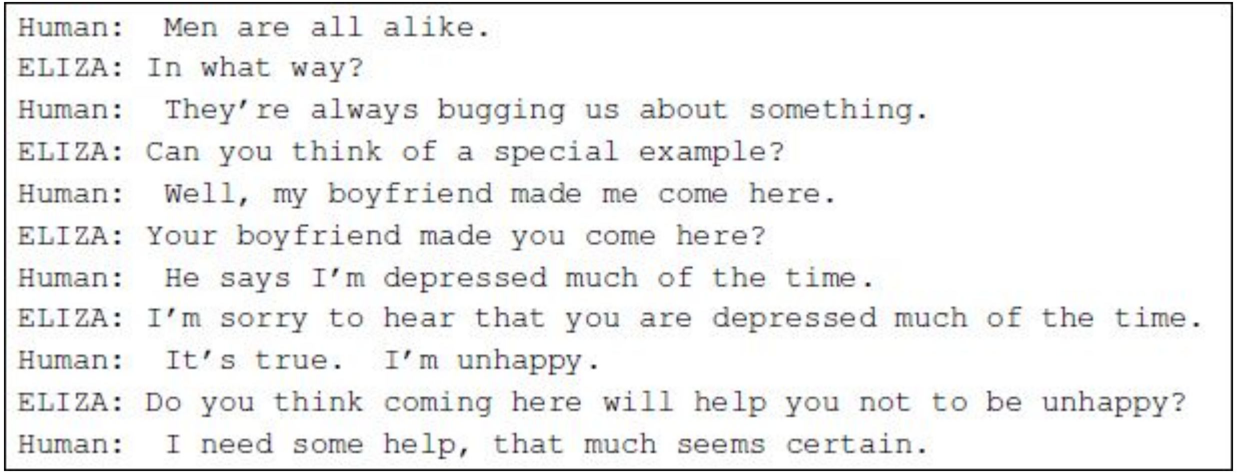
The cost of misclassification

Our educational goal is to develop fundamental knowledge and skills to assess AI systems adversarially

- Ethics of the research question: Would answering this research question advance science without violating social contracts? What are potentials for misuse?
- Impact of technology and potential dual use: who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effects on people's lives?
- Privacy: Who owns the data? Understanding the differences between published versus publicized data, understanding the concept of user consent, and thinking about implicit assumptions of users on how their data will be used.
- Bias in data: What are possible artifacts in data, given population-specific distributions? How representative is this data to address the target task?
- Social bias & unfairness in models: Is there sufficient control for confounding variables and corner cases? Does the system optimize for the “right” objective? Could the system amplify data bias?
- Utility-based evaluation beyond accuracy: Is there a utility-based evaluation beyond accuracy; e.g., measurements of false positive and false negative rates as measurements of fairness? What is “the cost” of misclassification and fault (in)tolerance?

ELIZA: Weizenbaum (1966)

- A rule-based dialogue system which mimicked a Rogerian psychotherapist, built at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum
- A simple rule-based algorithm that “reflects” what human is saying
- One of the first “AI systems” that raised ethical questions



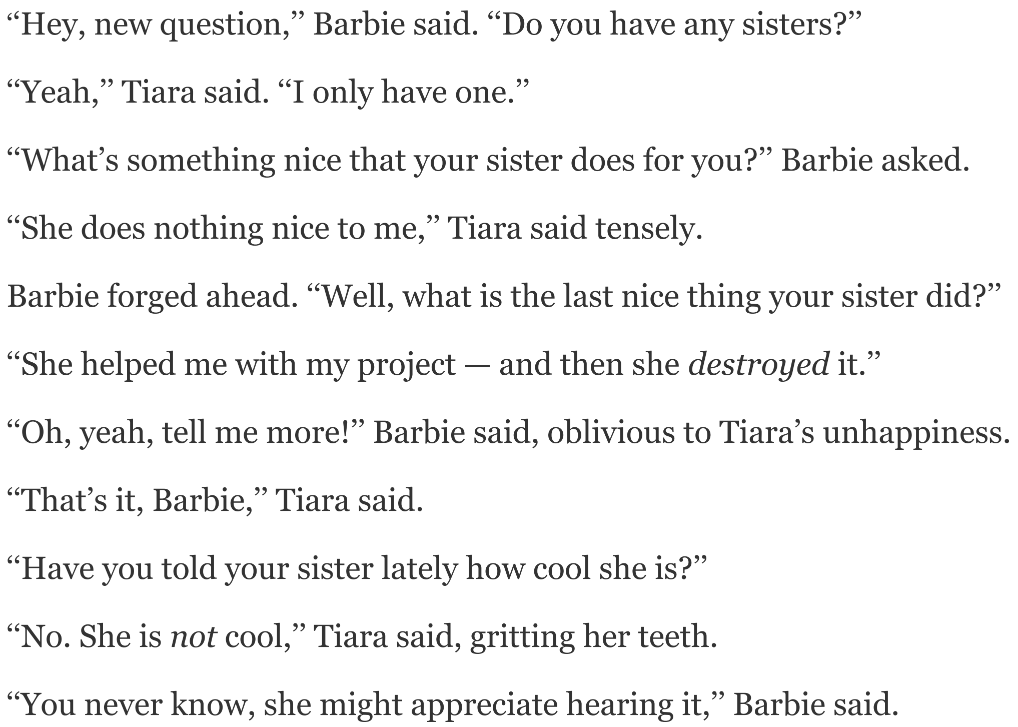
```
Human:  Men are all alike.
ELIZA: In what way?
Human:  They're always bugging us about something.
ELIZA: Can you think of a special example?
Human:  Well, my boyfriend made me come here.
ELIZA: Your boyfriend made you come here?
Human:  He says I'm depressed much of the time.
ELIZA: I'm sorry to hear that you are depressed much of the time.
Human:  It's true.  I'm unhappy.
ELIZA: Do you think coming here will help you not to be unhappy?
Human:  I need some help, that much seems certain.
```

Ethical implications of ELIZA

- People became deeply emotionally involved with the program
- Weizenbaum's secretary asked him to leave the room when she talked with ELIZA
- When he suggested that he might want to store all the ELIZA conversations for later analysis, people immediately pointed out the privacy implications
    - Suggesting that they were having quite private conversations with ELIZA

“Smart” toys: Barbie

https://www.nytimes.com/2015/09/20/magazine/barbie-wants-to-get-to-know-your-child.html



“Hey, new question,” Barbie said. “Do you have any sisters?”

“Yeah,” Tiara said. “I only have one.”

“What’s something nice that your sister does for you?” Barbie asked.

“She does nothing nice to me,” Tiara said tensely.

Barbie forged ahead. “Well, what is the last nice thing your sister did?”

“She helped me with my project — and then she *destroyed* it.”

“Oh, yeah, tell me more!” Barbie said, oblivious to Tiara’s unhappiness.

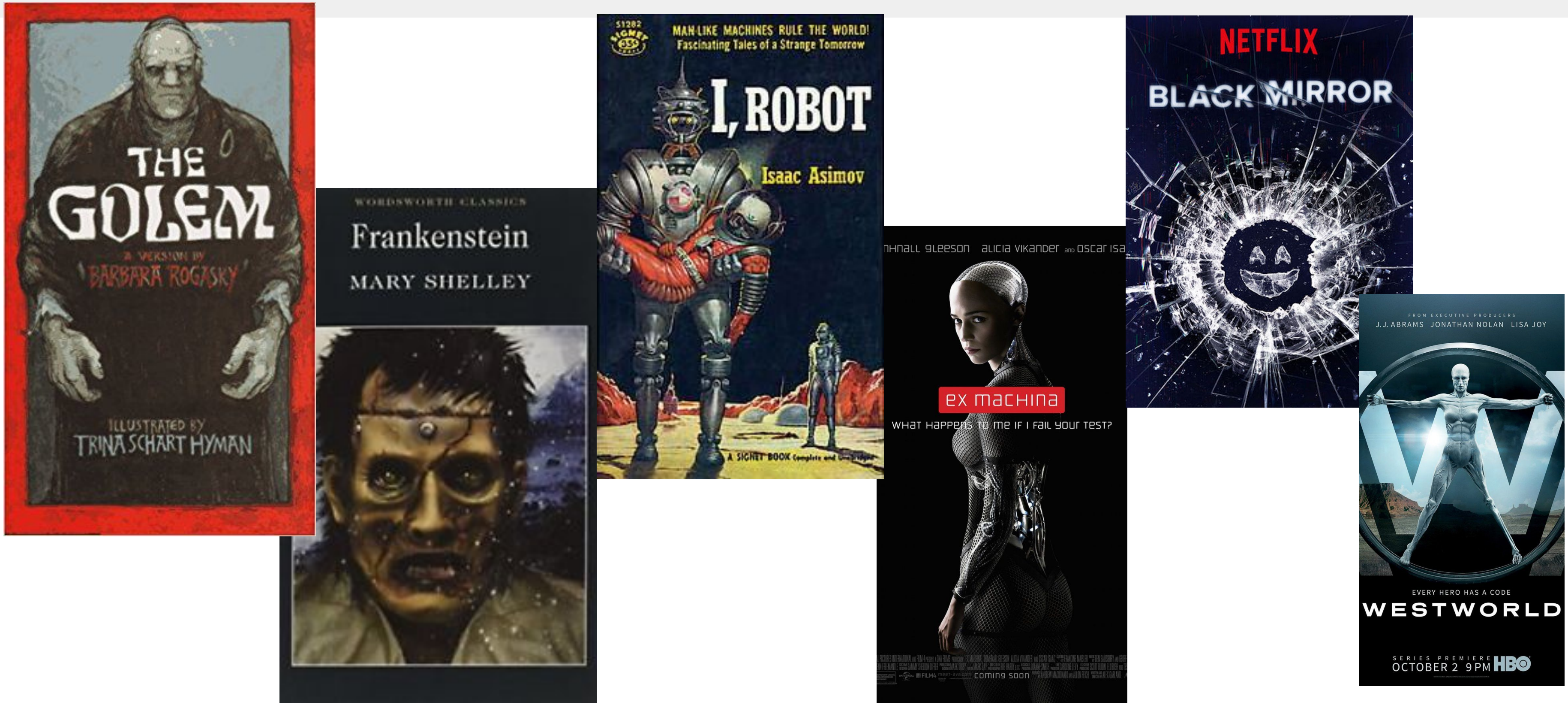
“That’s it, Barbie,” Tiara said.

“Have you told your sister lately how cool she is?”

“No. She is *not* cool,” Tiara said, gritting her teeth.

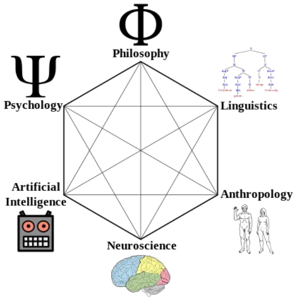
“You never know, she might appreciate hearing it,” Barbie said.

The Long History of Ethics and AI



Topics on ethical and social issues in AI

- Social bias and algorithmic fairness: social bias in data & AI models
- Privacy and safety: Privacy violation & language-based profiling
- Incivility: Hate-speech, toxicity, incivility, microaggressions online
- Misinformation: Fake news, information manipulation, opinion manipulation
- Technological divide: Unfair technologies underperforming for speakers of minority dialects, for languages from developing countries, and for disadvantaged populations
- Environmental impacts of AI models



Recommended introductory readings and talks

- Barocas & Selbst (2016) Big Data's Disparate Impact California Law Review
- Barbara Grosz talk (2017) Intelligent Systems: Design & Ethical Challenges
- Kate Crawford NeurIPS keynote (2017) The Trouble with Bias
- Yonatan Zunger blog post (2017) Asking the Right Questions About AI
- Weidinger et al. (2022) Taxonomy of Risks Posed by Language Models FAccT
-

Please refer to the reading list and additional resources on the course website, we'll constantly updating and expanding the list

http://tiny.cc/CSE582-sp24

Thank you!