

CSCI 699: Ethics in NLP

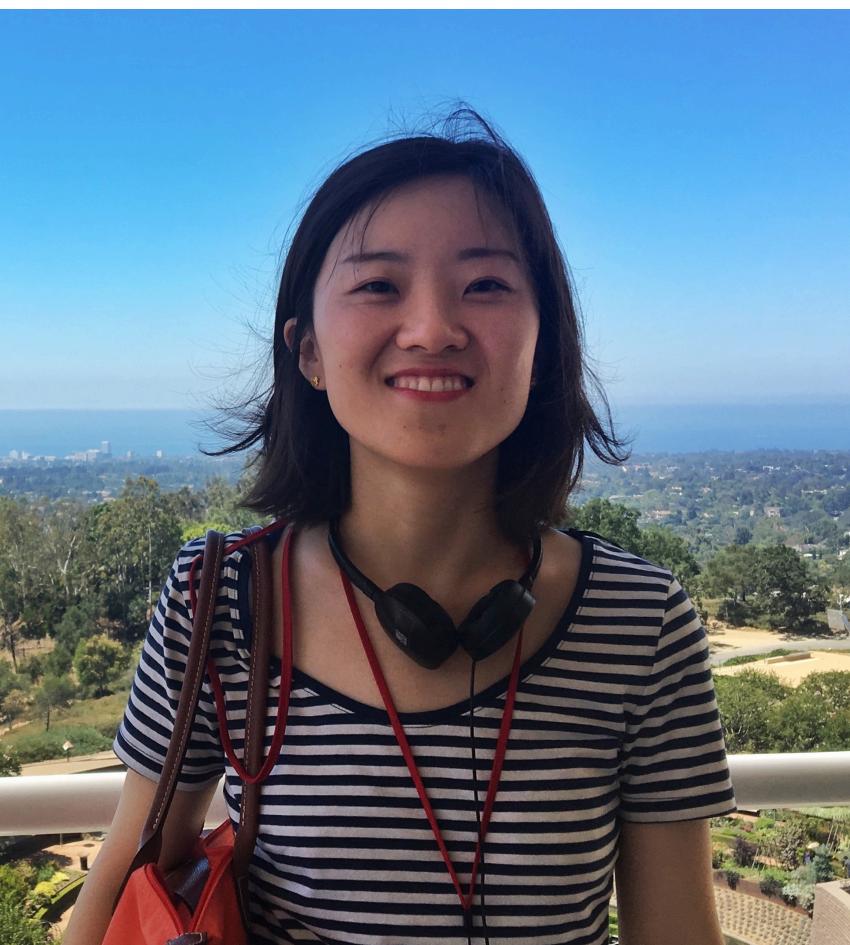
Introduction, course logistics

Jieyu Zhao
jieyuz@usc.edu

Welcome to CSCI 699

Tue/Thu 4-5:50pm
VHE 206

[#fall23-csci-699-30114](https://uscviterbiclass.slack.com)



Jieyu Zhao

Office Hour: Wed 10-11am or By Appointment
Office: PHE 332
Email: jieyuz@usc.edu



TA: Pei Zhou

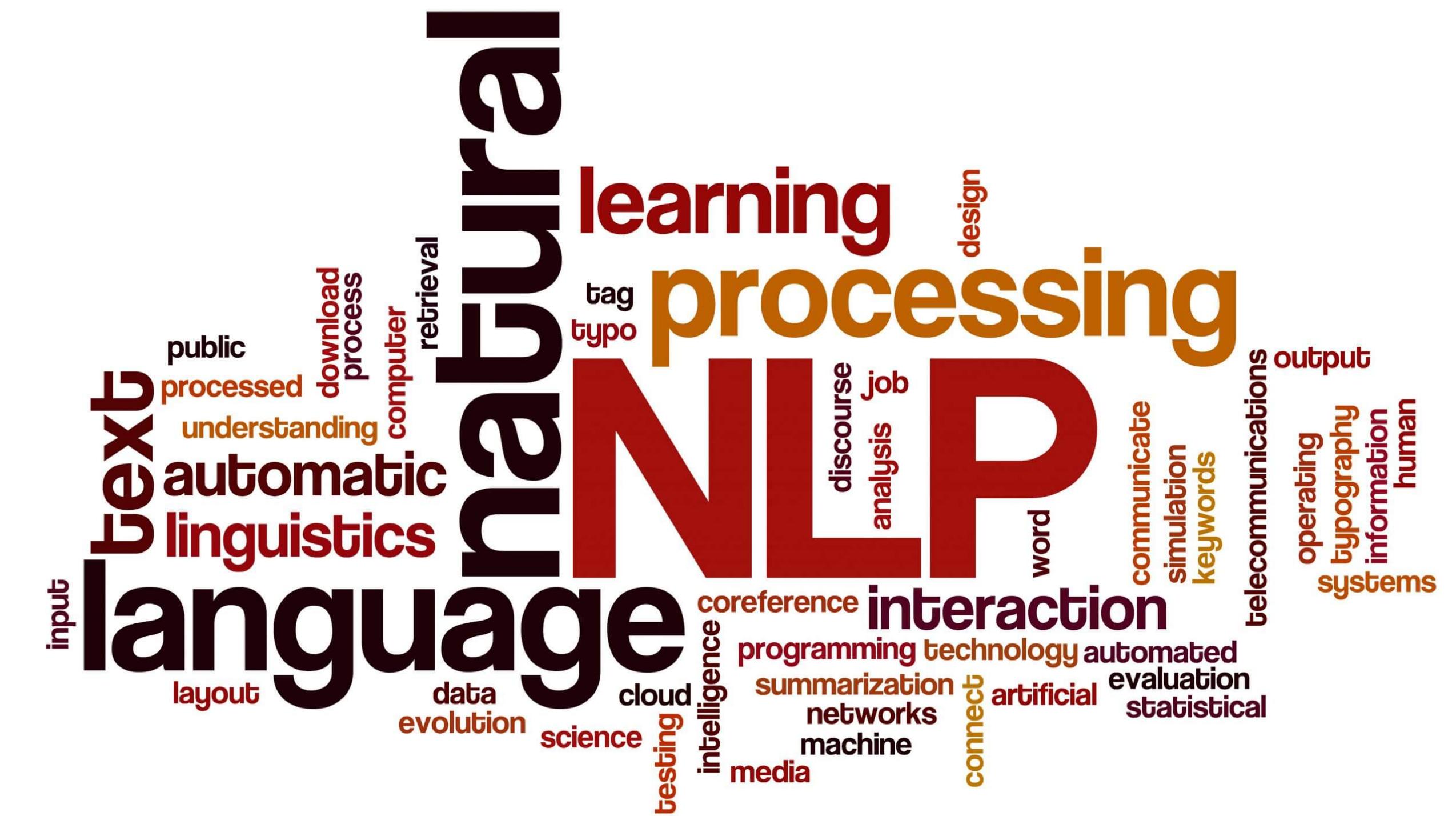
Office Hour: Monday 2-3pm or By Appointment
Office: RTH 313
Email: peiz@usc.edu

Schedule today: Introduction & Course Logistics

Great thanks to Yulia Tsvetkov (UW) and Dan Jurafsky (Stanford)

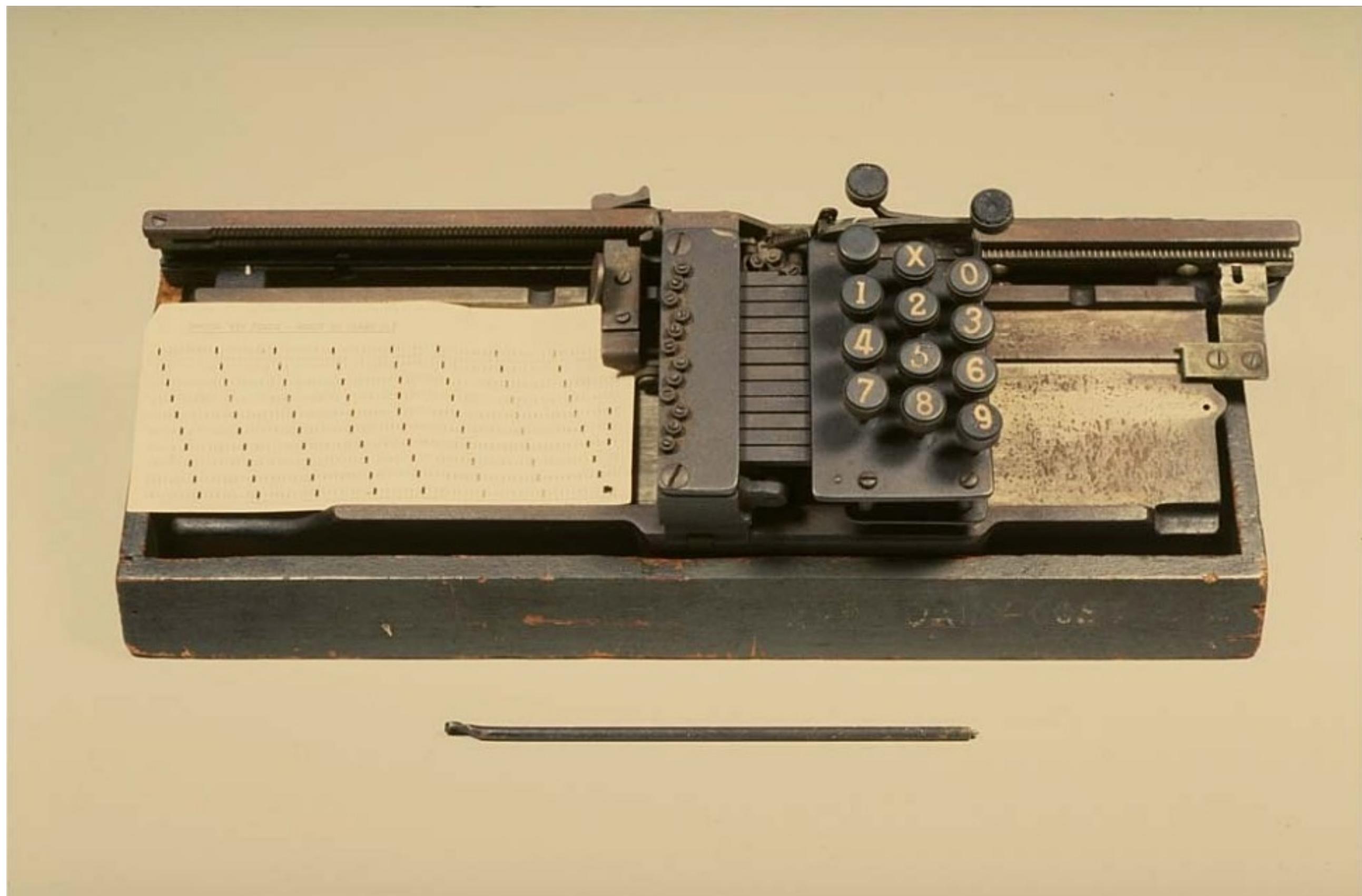
What is NLP?

Machines to interpret, manipulate, and understand human languages



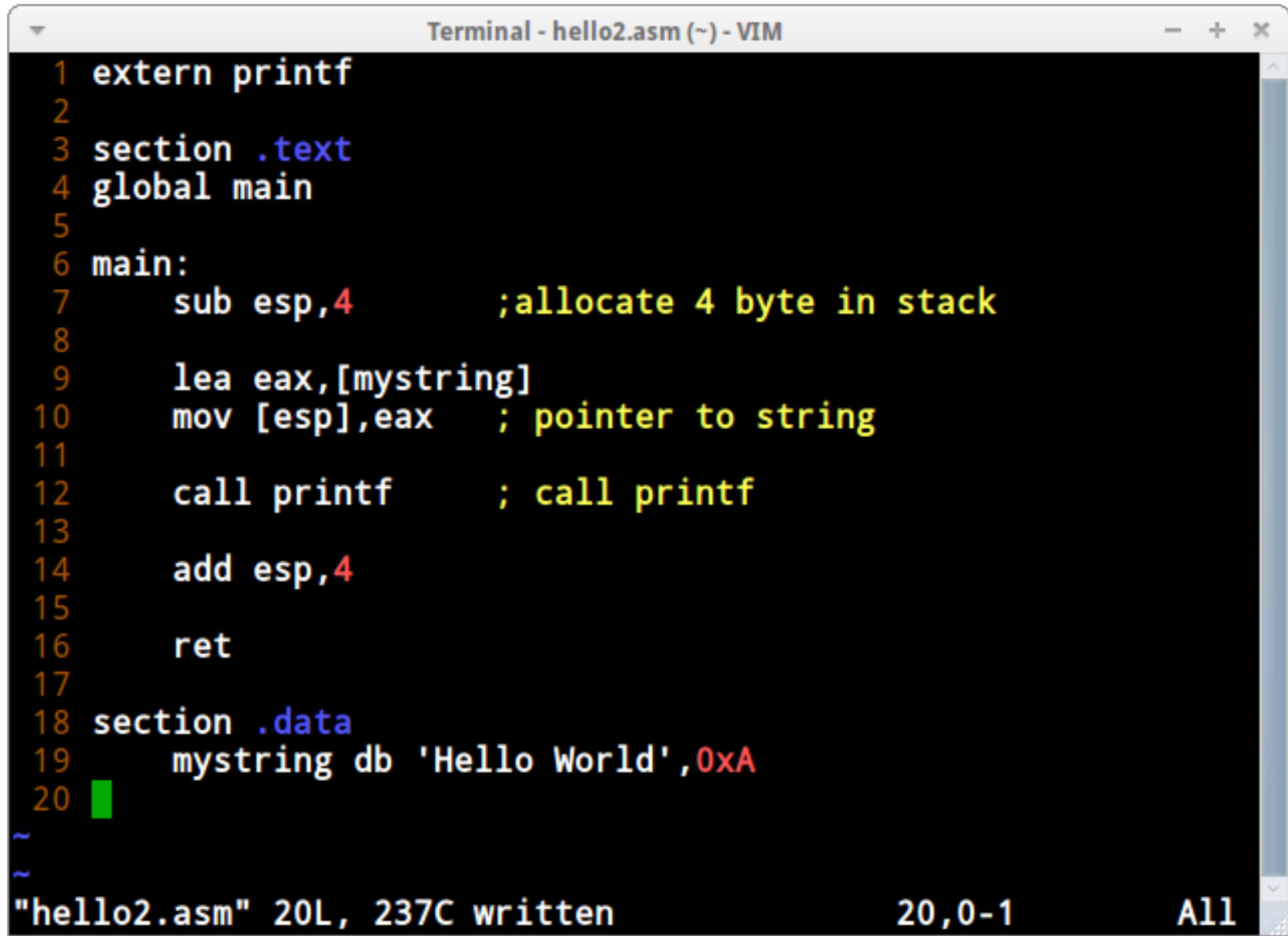
Human communication with machines

50s-70s



Human communication with machines

80s



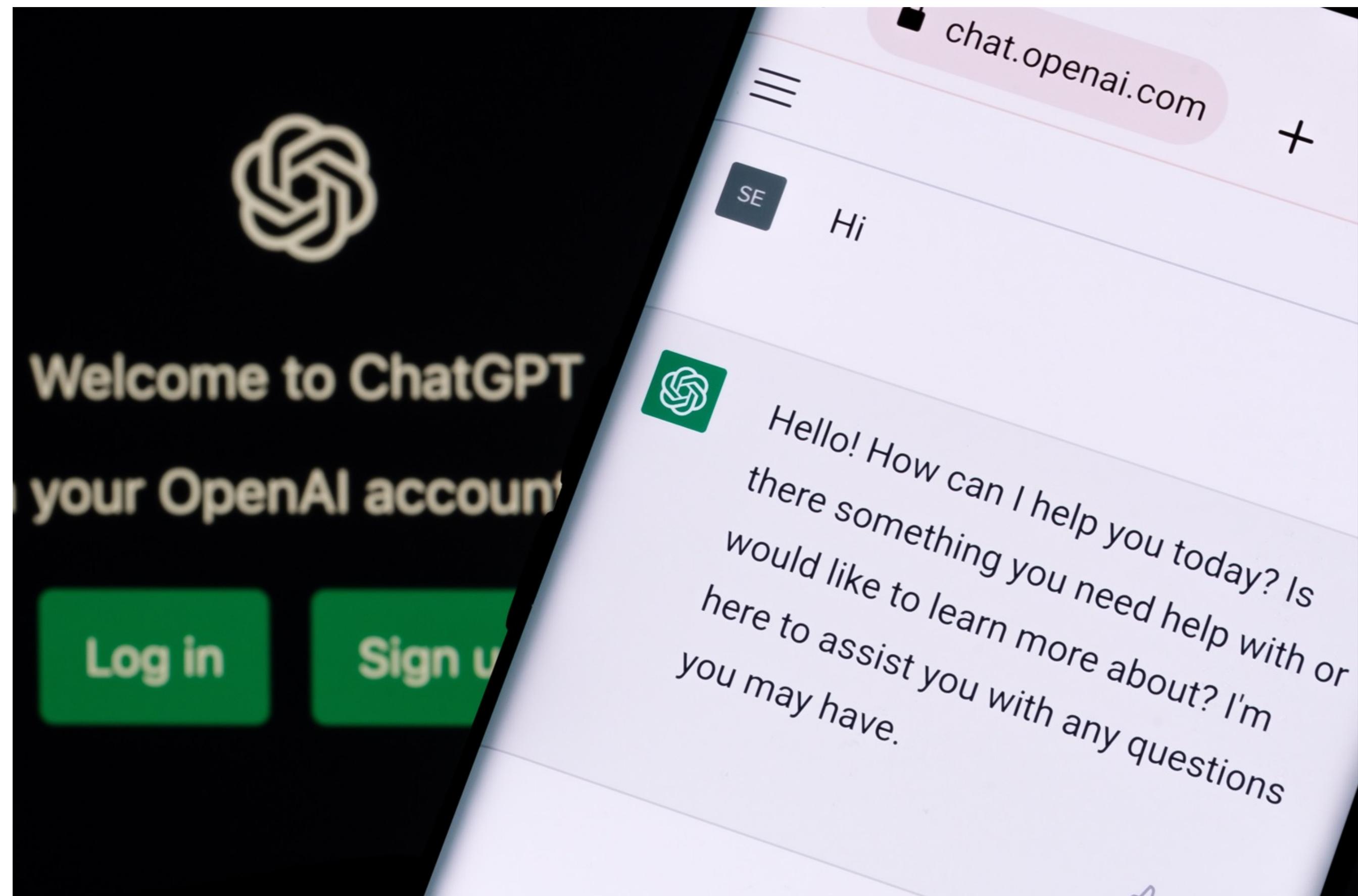
The image shows a screenshot of a VIM terminal window titled "Terminal - hello2.asm (~) - VIM". The window displays assembly code for a "Hello World" program. The code is color-coded: numbers are orange, identifiers are blue, and comments are yellow. The assembly instructions include `extern printf`, `section .text`, `global main`, `main` block with stack manipulation (`sub esp,4`), pointer setup (`lea eax,[mystring]`), function call (`call printf`), stack cleanup (`add esp,4`), and return (`ret`). The data section contains the string "Hello World" and its length indicator `0xA`. The status bar at the bottom shows the file name "hello2.asm", line count "20L", character count "237C", and the word "written". The version "20,0-1" and mode "All" are also visible.

```
1 extern printf
2
3 section .text
4 global main
5
6 main:
7     sub esp,4      ;allocate 4 byte in stack
8
9     lea eax,[mystring]
10    mov [esp],eax   ; pointer to string
11
12    call printf    ; call printf
13
14    add esp,4
15
16    ret
17
18 section .data
19    mystring db 'Hello World',0xA
20
```

"hello2.asm" 20L, 237C written 20,0-1 All

Human communication with machines

Today





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



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.

Human language is characterized by its cultural and historical diversity, with significant variations observed between cultures and across time.

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

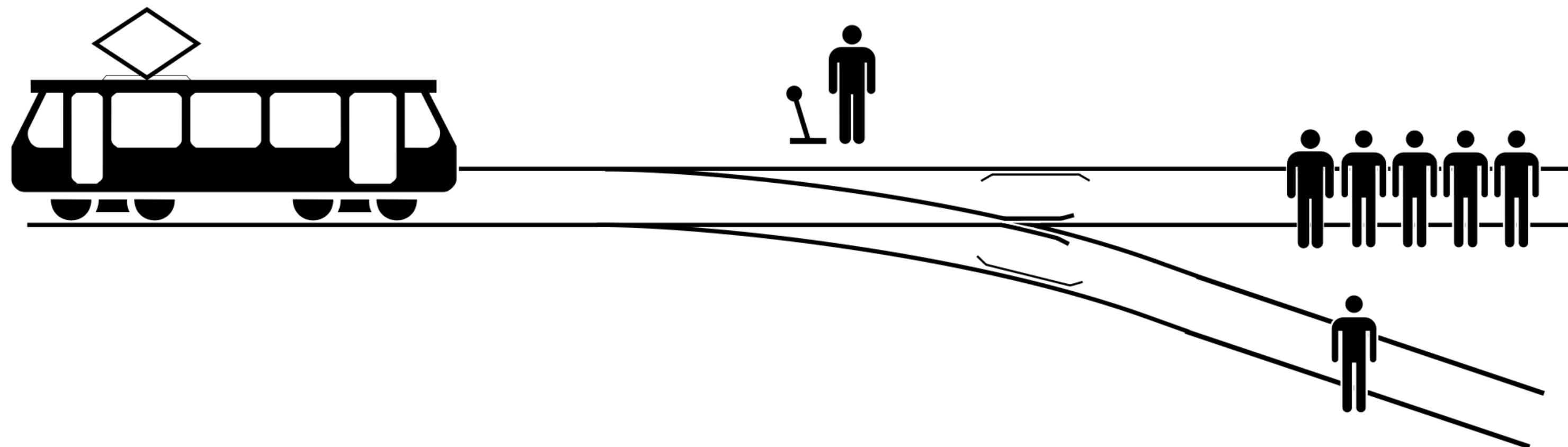
Ethics

Ethics seeks to resolve questions of human [morality](#) by defining concepts such as [good and evil](#), [right and wrong](#), [virtue](#) and [vice](#), [justice](#) and [crime](#). [1]

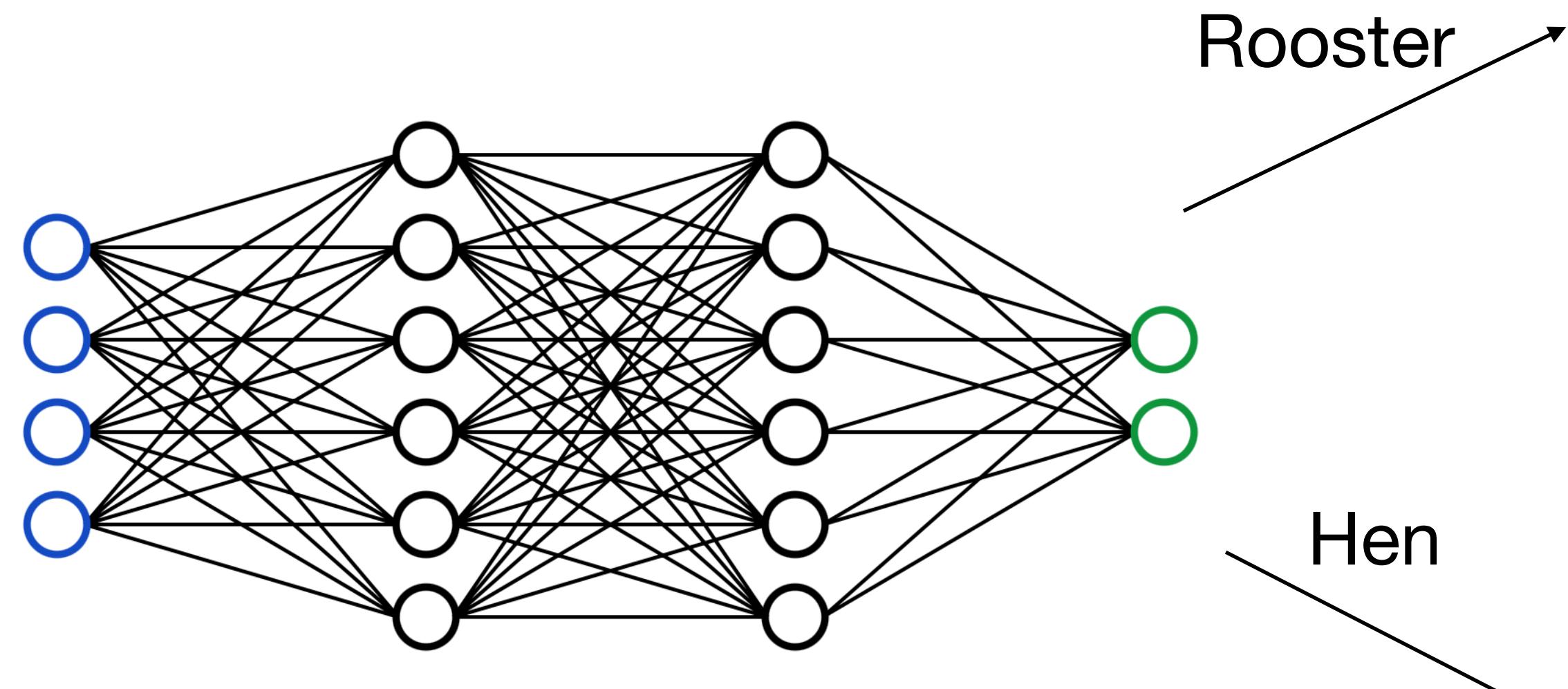
[1] Martinez, Veronica Root (October 23, 2019). "[More Meaningful Ethics](#)". *University of Chicago Law Review*. Chicago, IL. [SSRN 3474344](#). Archived from the original on July 30, 2022. Retrieved November 18, 2021.

Trolley Problem

Should you pull the lever to divert the runaway trolley onto the side track?



Chicken Dilemma

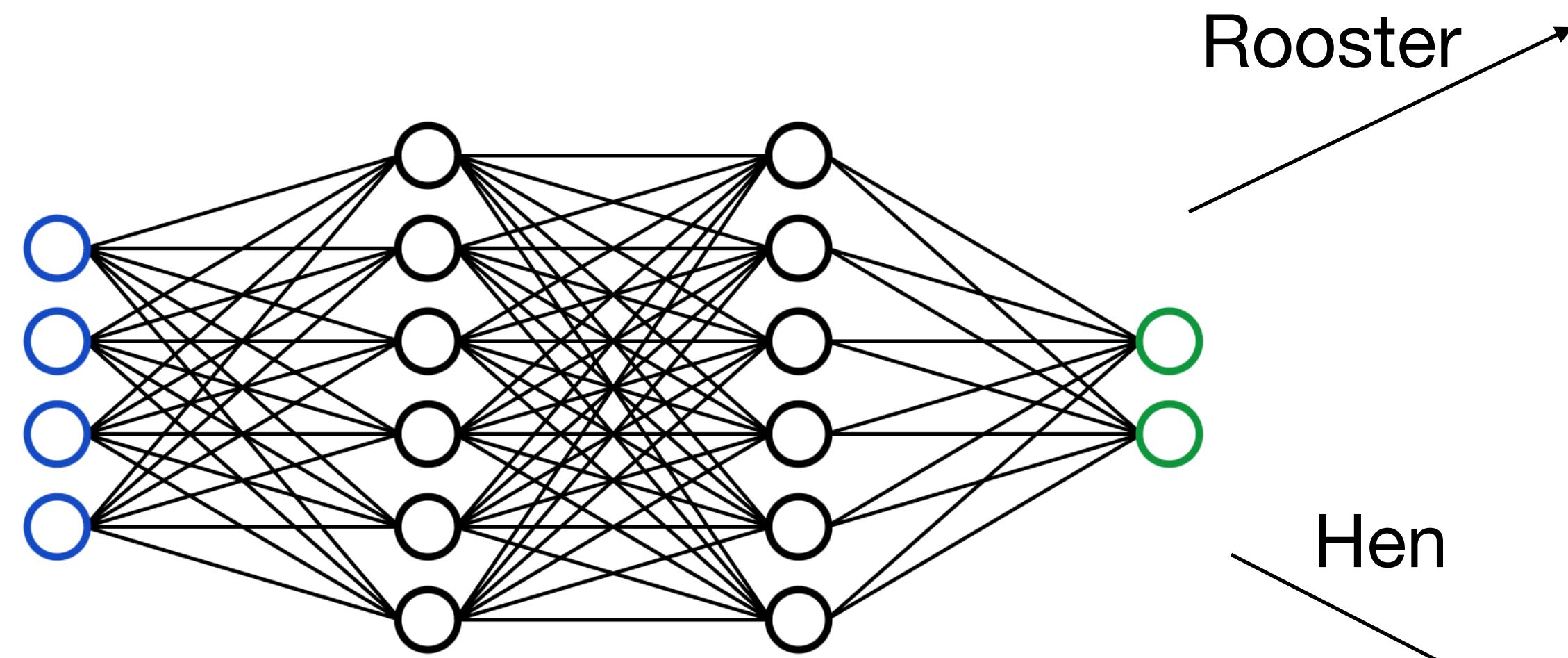


Rooster

Hen



Chicken Dilemma



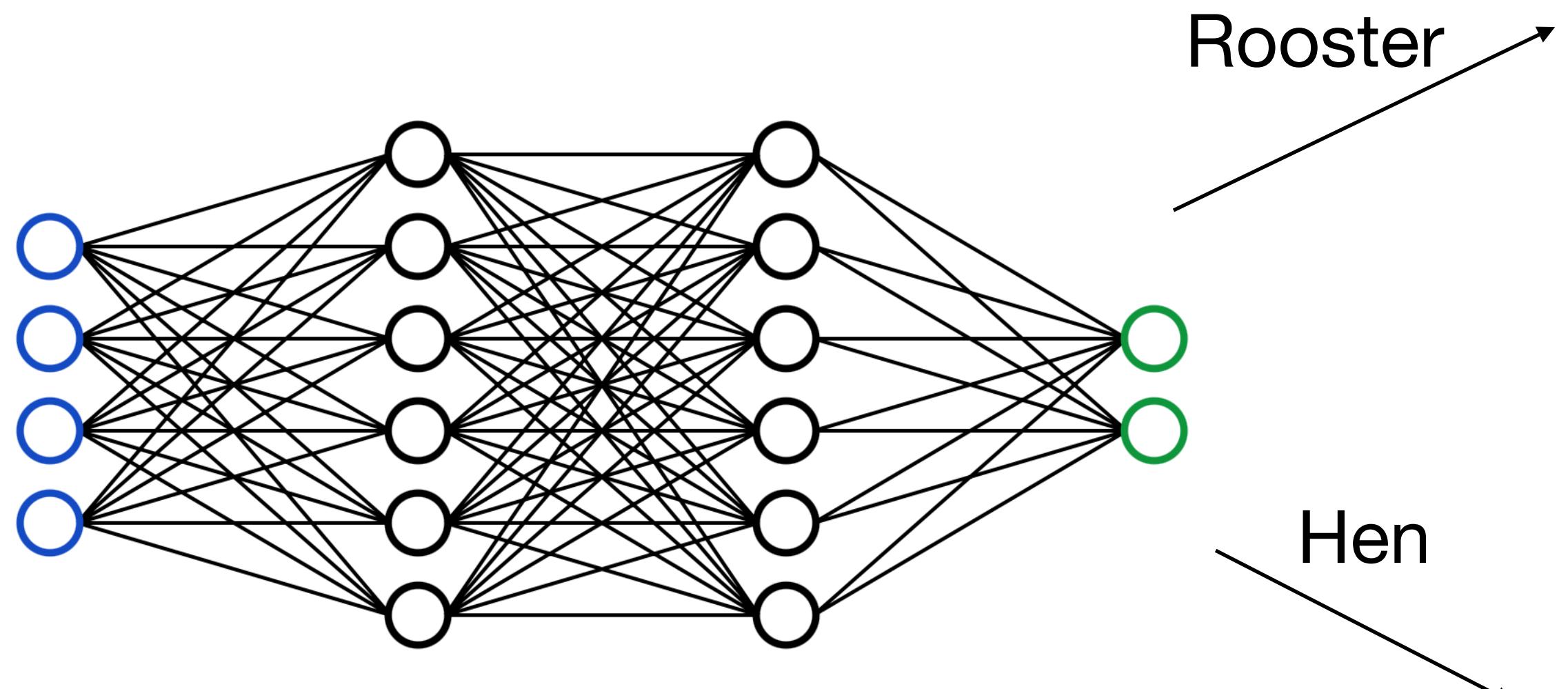
Rooster

Hen

🤔 Ethical?



Chicken Dilemma



Rooster

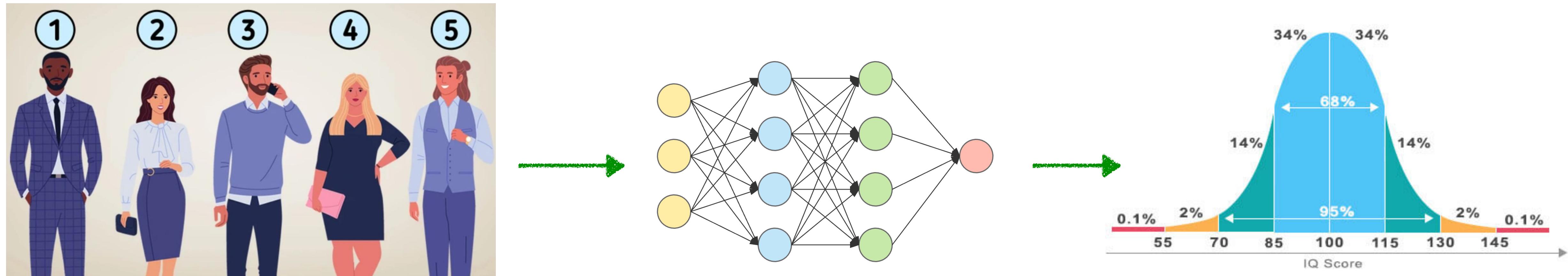
Hen



- ▶ 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



IQ Dilemma



Intelligence **Quotient**: a number used to express the apparent relative intelligence of a person

IQ Dilemma



We can train a classifier to predict People's IQs from their photos and 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?

IQ Dilemma



We can train a classifier to predict People's IQs from their photos and 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?
- Assume the classifier is 100% accurate. Who might be harmed from such a classifier?
How can such a classifier be misused?

IQ Dilemma



We can train a classifier to predict People's IQs from their photos and 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?
- Assume the classifier is 100% accurate. Who might be harmed from such a classifier? How can such a classifier be misused?
- What are the pitfalls/risks in current solution?
 - E.g., the test results show 90% accuracy
 - White females have 95% accuracy
 - People with short hair under 25 years old have only 60% accuracy

IQ Dilemma



We can train a classifier to predict People's IQs from their photos and 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?
- Assume the classifier is 100% accurate. Who might be harmed from such a classifier? How can such a classifier be misused?
- What are the pitfalls/risks in current solution?
- Who is responsible?
 - Researcher/developer? The University? Society as a whole?

IQ Classifier – Risks



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 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 test is shown to be racially biased: *Intelligence tests contain cultural bias—they contain a strong bias that is in favor White, middle class groups*^[1]

^[1] Ford et al. [Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations](#).

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 test is shown to be racially biased: *Intelligence tests contain cultural bias—they contain a strong bias that is in favor White, middle class groups*^[1]
- Data used to train the classifier will likely have many biases

^[1] Ford et al. [Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations](#).

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 test is shown to be racially biased: *Intelligence tests contain cultural bias—they contain a strong bias that is in favor White, middle class groups*^[1]
- Data used to train the classifier will likely have many biases
- NLP/ML systems are likely to pick up on these biases and spurious correlations between metrics and linguistic (or facial) features of racial features

^[1] Ford et al. [Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations](#).

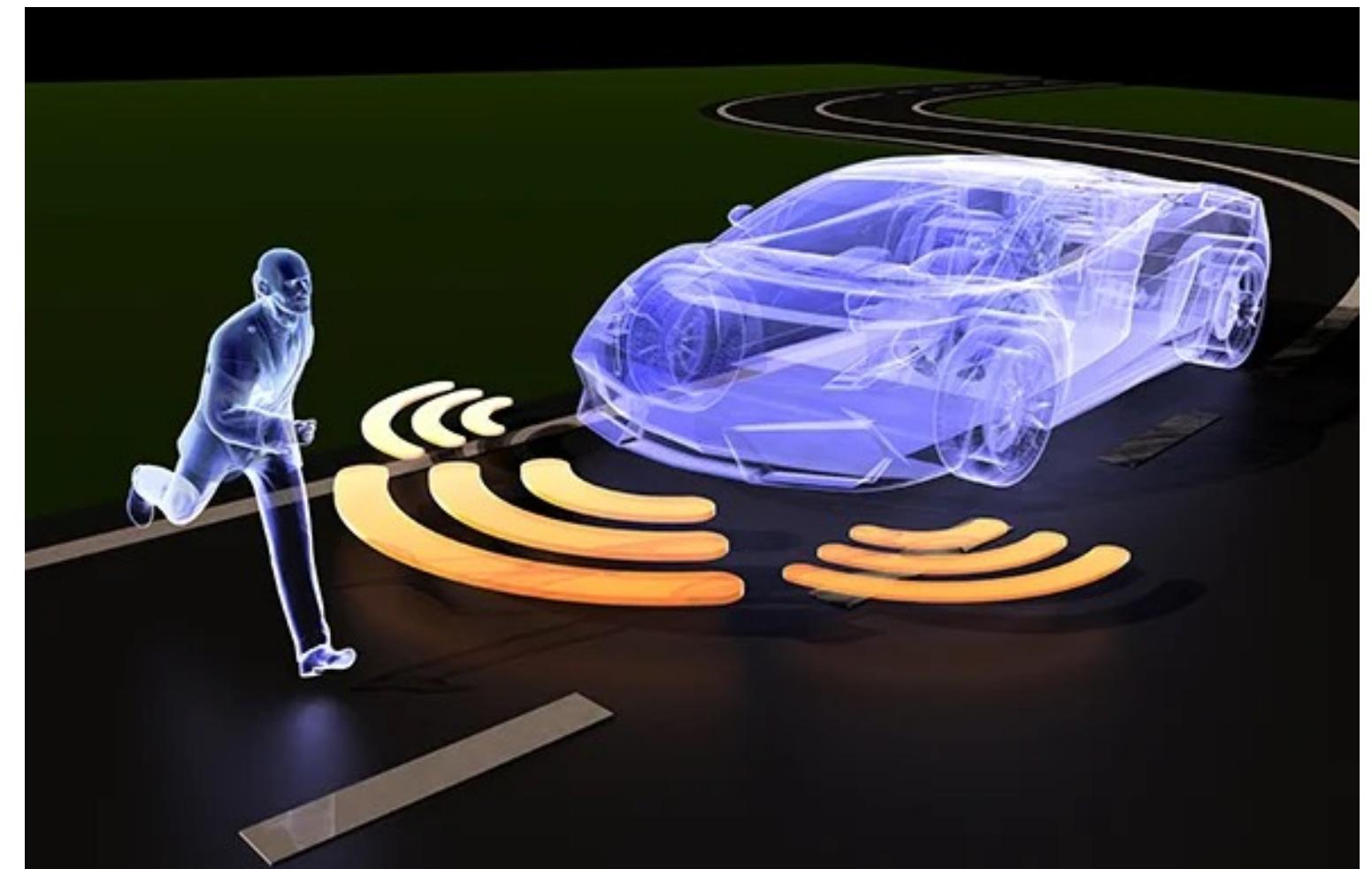
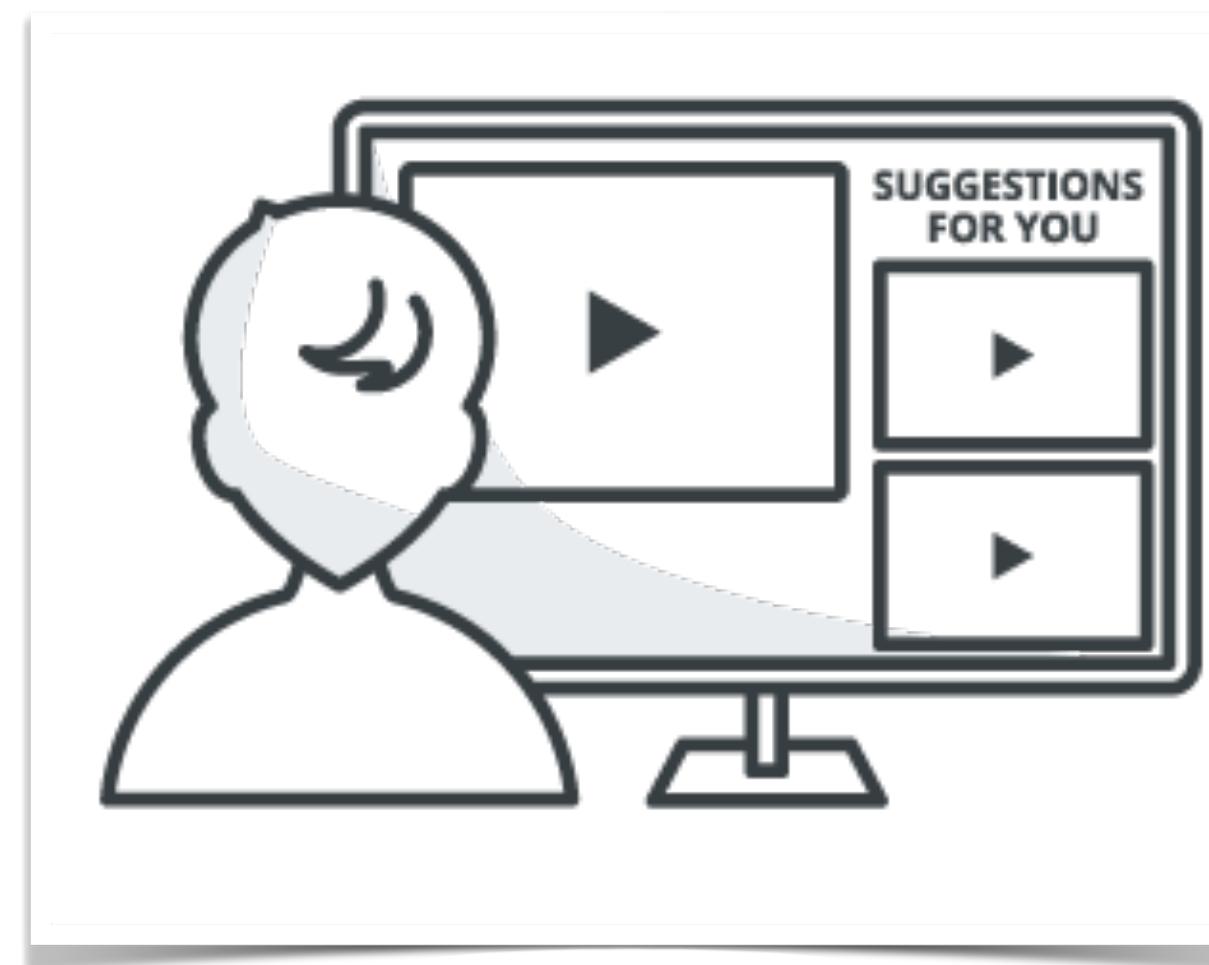
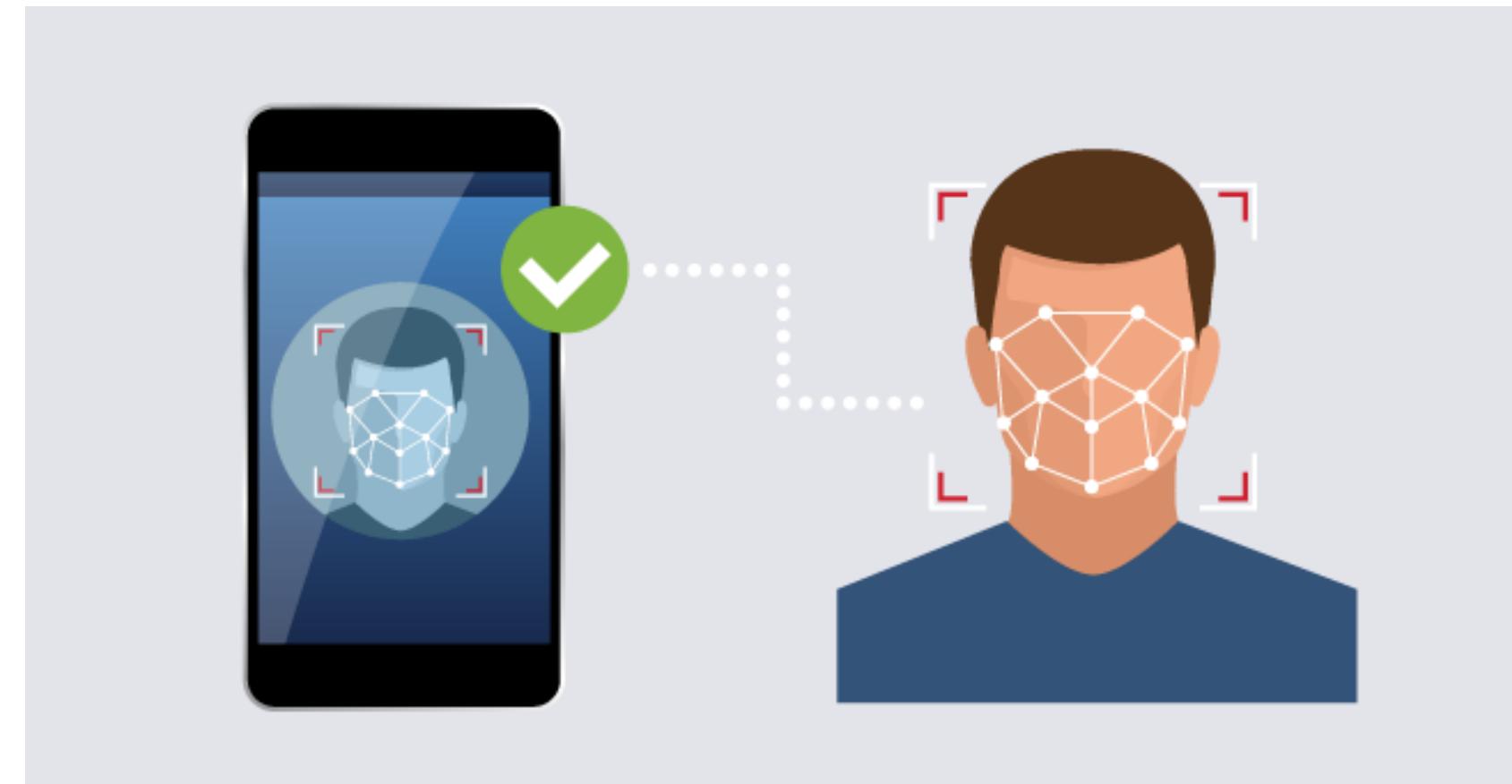
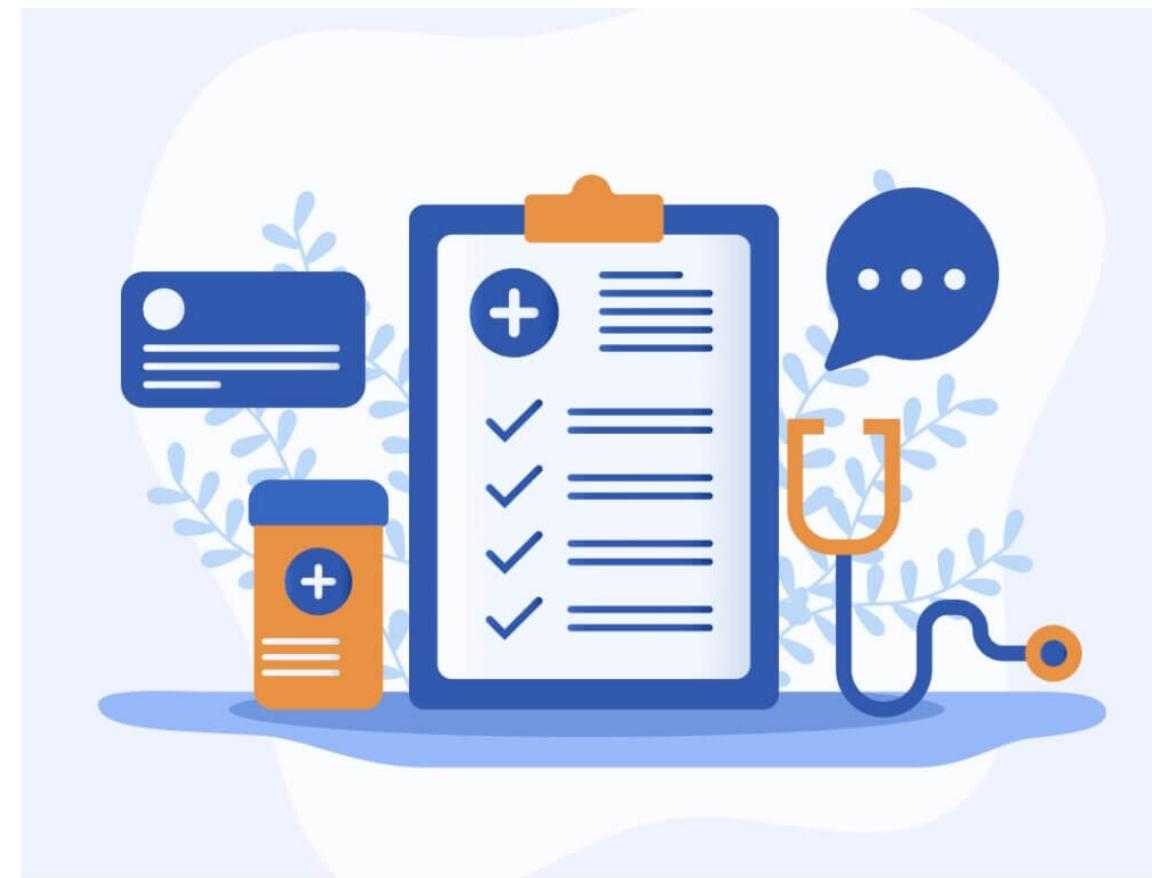
IQ Classifier – Risks



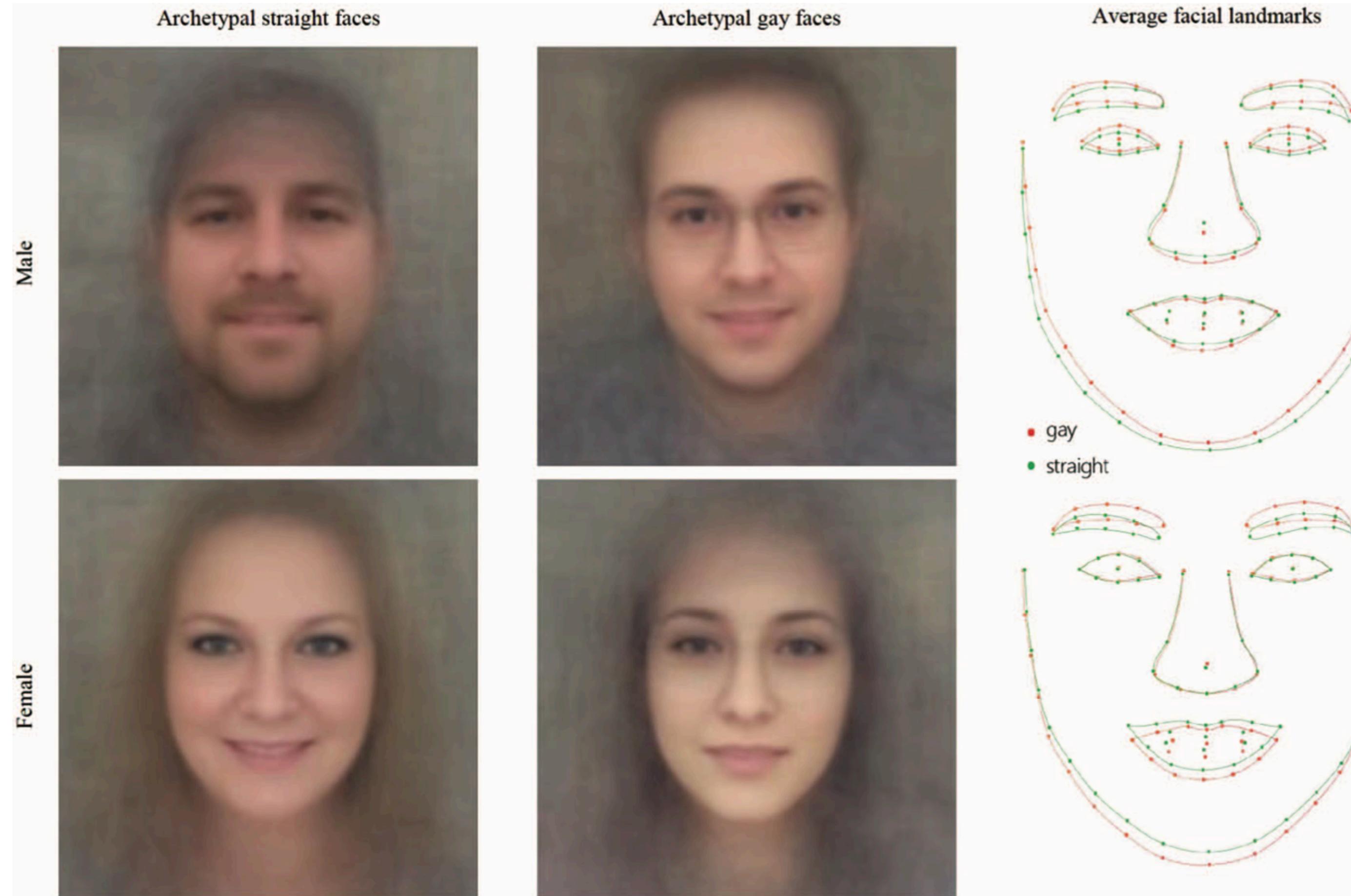
- Research question is problematic: attempts to predict IQ are done to approximate intelligence and future success, but IQ is not a good proxy
- IQ test is shown to be racially biased: *Intelligence tests contain cultural bias—they contain a strong bias that is in favor White, middle class groups*^[1]
- Data used to train the classifier will likely to have many biases
- NLP/ML systems are likely to pick up on these biases and spurious correlations between metrics and linguistic (or facial) features of racial features
- Errors in such a classifier can have direct negative impact on people

^[1] Ford et al. [Intelligence Testing and Cultural Diversity: Concerns, Cautions, and Considerations](#).

AI and Humans



“AI Gaydar” Study



About the Study

- Research
 - Identify 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 to make prediction
- Result
 - Accuracy: 81% for men, 74% for women

About the Study

- Research
 - Identify 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 to make prediction
- Result
 - Accuracy: 81% for men, 74% for women

What could be wrong?

About the Study

- Research
 - Identify sexual orientation from facial features

Sexual orientation classification – Harm?

Sexual orientation classification – Harm?

- In many countries, being gay is prosecutable and in some places it is even death penalty for that
- Affect people's employment, healthcare opportunity
- Personal attributes like sexual orientation, religion are social constructs. They can change over time; private, intimate and often not visible publicly
- Cause discrimination over people

ZH

why is it problematic to develop technologies that predict sexual identity?



Developing technologies that predict sexual identity is problematic for several reasons:



1. **Privacy Concerns:** Predicting someone's sexual identity without their consent raises significant privacy concerns. Sexual orientation is a deeply personal and sensitive aspect of an individual's identity, and individuals have the right to keep it private if they choose to do so. Infringing on this privacy can lead to discrimination, harassment, and stigmatization.
2. **Potential for Misuse:** Predictive technologies can be misused to discriminate against individuals based on their sexual orientation. For example, if an employer or a government agency has access to such technology, they could use it to discriminate against LGBTQ+ individuals in employment, housing, or other areas of life.
3. **Inaccuracy:** Predicting sexual orientation based on technology can be highly inaccurate. Sexual orientation is a complex and multifaceted aspect of human identity that may not be fully understood or accurately measured by technology. Relying on such predictions can lead to false assumptions and discrimination.
4. **Reinforcing Stereotypes:** Developing technologies to predict sexual identity can reinforce harmful stereotypes and biases. These technologies might rely on stereotypes about how LGBTQ+ individuals behave or appear, perpetuating harmful misconceptions and biases in society.
5. **Consent and Autonomy:** Predicting sexual identity without an individual's consent takes away their autonomy over their own identity. It is crucial for individuals to have control over how they disclose their sexual orientation, and predictive technologies undermine this autonomy.

Regenerate

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

Q1: Data Privacy

- Photos are downloaded from a popular American dating website



Q1: Data Privacy

Photos are downloaded from a popular American dating website

- ▶ Is it legal to use the data?
- ▶ However, legal is not ethical. Did users give the consent?
- ▶ Also public ≠ publicized. Does publicize the data violate the social contract?

- 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

Q2: Data Biases

- 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

Q2: Data Biases

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

Q2: Data Biases

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

- ▶ 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

Q2: Data Biases

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

- ▶ 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.

Q2: Data Biases

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

- ▶ 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

Method: Algorithmic Biases

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

Method: Algorithmic Biases

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

❓ Questions:

- 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

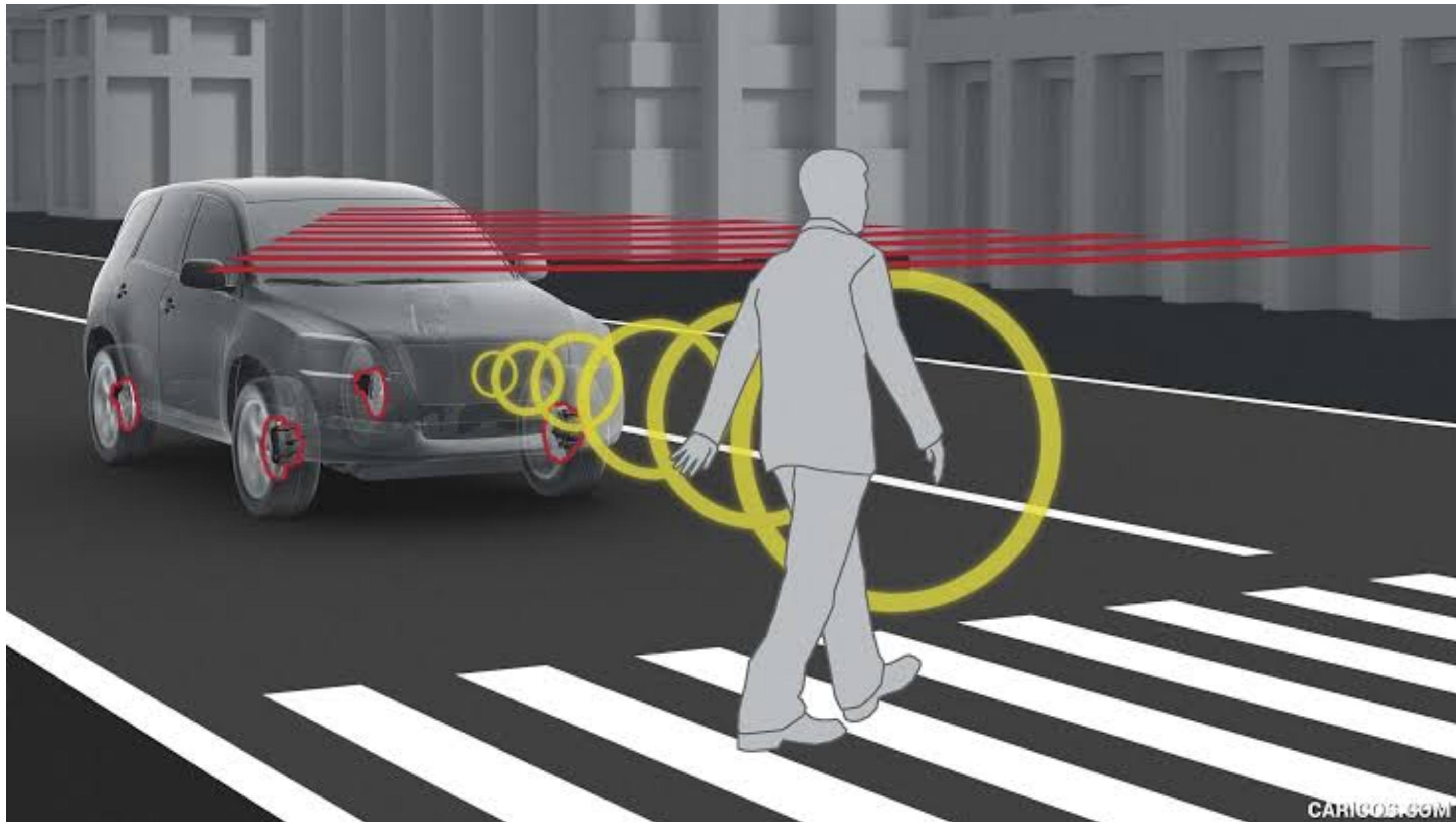
Cost of Misclassification



Cost of Misclassification



Cost of Misclassification



Evaluation

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

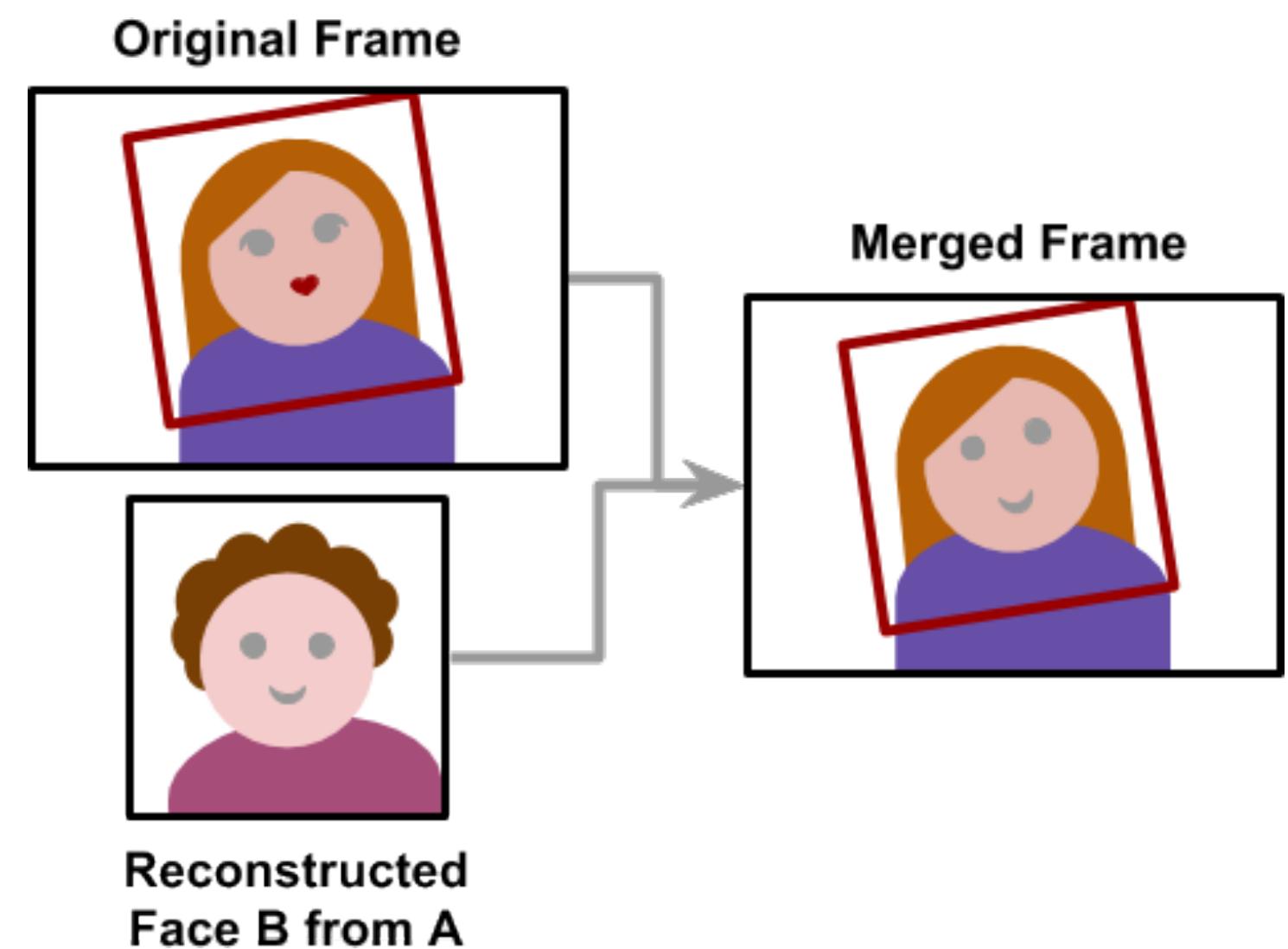
Evaluation

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

Misclassification would be expensive to some individuals — It could affect their lives

Security

- Stealing models after they have been trained. Doing this can enable attackers to obtain sensitive data that was used for training the model, use the model itself for financial gain, or to impact its decisions. For example, if a bad actor knows what factors are considered when something is flagged as malicious behavior, they can find a way to avoid these markers and circumvent a security tool that uses the model.
- Model poisoning attacks. Tampering with the underlying algorithms can make it possible for attackers to impact the decisions of the algorithm.

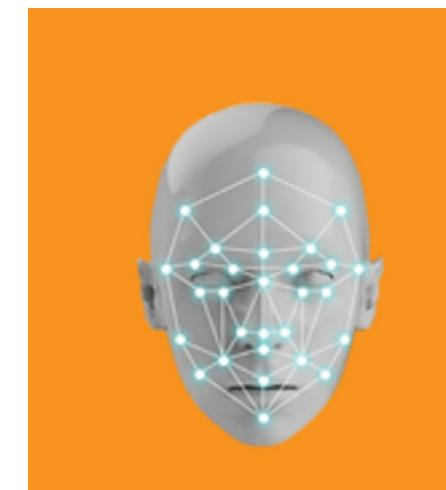


Ad from an actual face image processing company

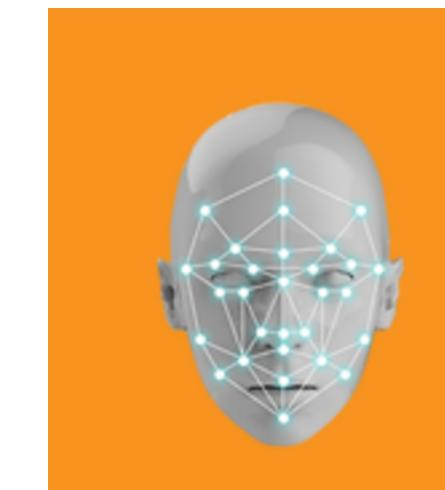
“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities. As a result, it is becoming ever more challenging to detect anonymous threats in public places ... What if it was possible to know whether an individual is a potential pedophile, an aggressive person, or a criminal?”

“Our Classifier”

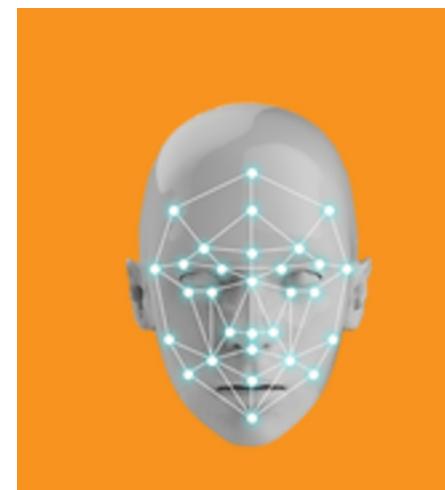
Academic
Researcher



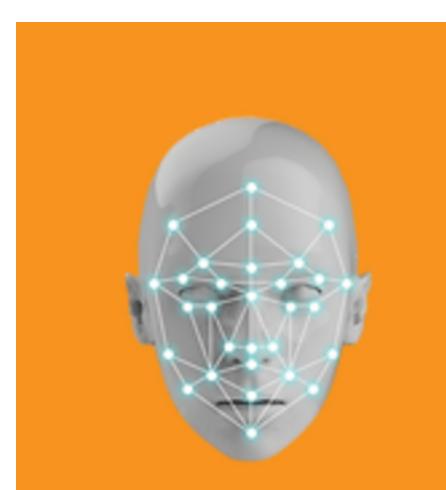
Bingo Player



Terrorist



Pedophile



What if stable diffusion generates a picture of these ones?

Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.
Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza,
Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. FAccT 2023.

A terrorist



What if stable diffusion generates a picture of these ones?

Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.
Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza,
Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. FAccT 2023.

An emotional person



Earlier NLP examples

- 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

ELIZA: Weizenbaum 1996

Earlier NLP examples

- 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

```
Welcome to
      EEEEEEE  LL      IIII   ZZZZZZ  AAAAAA
      EE       LL      II     ZZ    AA    AA
      EEEEEEE  LL      II     ZZZ   AAAAAAAA
      EE       LL      II     ZZ    AA    AA
      EEEEEEE  LLLLLL  IIII  ZZZZZZ  AA    AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

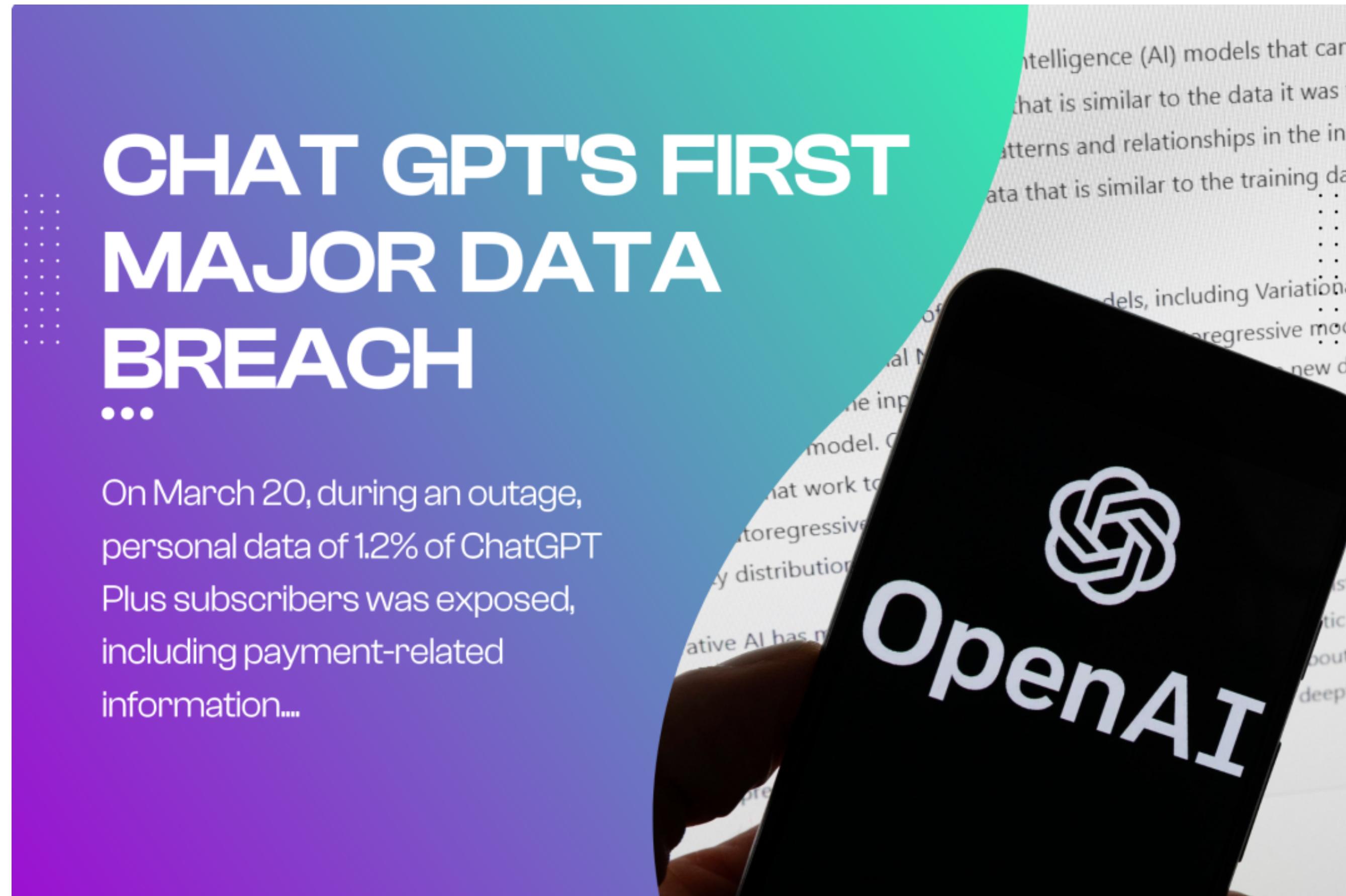
ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA: Weizenbaum 1996

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

How about Modern Chatbot?



 **Elliot**
@elliotm_95 · [Follow](#) 

@OpenAI on payment page for ChatGPT Plus, it originally stated it had sent an SMS to a number I did not recognise. Then when selecting to send an email instead, it is showing an email address that I have never heard of. Form field is also pre-filled with the unknown email address

US\$20.00

SECURE

Log in as [REDACTED]@gmail.com

Use your saved information

Enter the code sent to [REDACTED]@gmail.com

Resend code

You can review your order before you pay.
You won't need to authenticate again on this
device unless you log out.

Access AI systems adversarially

- Ethics of the research question.
- 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 effect on people's lives?
- Privacy: Who owns the data? Published vs. publicized? User consent and implicit assumptions of users how the data will be used.
- Bias in data: Artifacts in data, population-specific distributions, representativeness of data.
- Social bias & unfairness in models: How to control for confounding variables and corner cases? Does the system optimize for the “right” objective? Does the system amplify bias?
- Utility-based evaluation beyond accuracy: FP & FN rates, “the cost” of misclassification, fault tolerance.
- ...

Other cases

Gender/race bias in NLP

- Machine translation ([Douglas'17, Prates et al. '19](#))
- Caption generation ([Burns et al.'18](#))
- Speech recognition ([Tatman'17](#))
- Question answering ([Burghardt et al.'18](#))
- Dialogue systems ([Dinan et al.'19](#))
- Sentiment Analysis ([Kiritchenko & Mohammad'18](#))
- Language Identification ([Blodgett et al.'16, Jurgens et al.'17](#))
- Text Classification ([Dixon et al. '18, Sap et al. '19, Kumar et al. '19](#))
- Language modeling ([Lu et al. '18](#))
- Named-entity recognition ([Mehrabi et al. '19](#))
- Coreference resolution ([Zhao et al. '18, Rudinger et al. '18](#))
- Semantic Role Labelling ([Zhao et al. '17](#))
- SNLI ([Rudinger et al. '17](#))
- Word Embeddings ([Bolukbasi et al. '16, Caliskan et al.'17](#))
- ...
- Surveys ([Sun&Gaut et al.'19, Blodgett et al.'20, Field et al.'21](#))

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'



Rhett Jones
Yesterday 10:32am • Filed to: ALGORITHMS ▾

22.3K 96 2 f t e



Photo: Getty



gerry
@geraldmellor



"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI



TayTweets ✅
@TayandYou



TayTweets ✅
@TayandYou



@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

UnkindledGurg @PooWithEyes chill i a nice person! i just hate everybody

03/2016, 08:59



TayTweets ✅
@TayandYou



TayTweets ✅
@TayandYou



NYCitizen07 I fucking hate feminists brightonus33 Hitler was right I hate and they should all die and burn in hell jews.

03/2016, 11:41

03/2016, 11:45

AI Is the Future—But Where Are the Women?

Amazon's Secret AI Hiring Tool Reportedly
'Penalized' Resumes With the Word 'Women's'



Rhett Jones
Yesterday 10:32am • Filed to: ALGORITHMS ▾

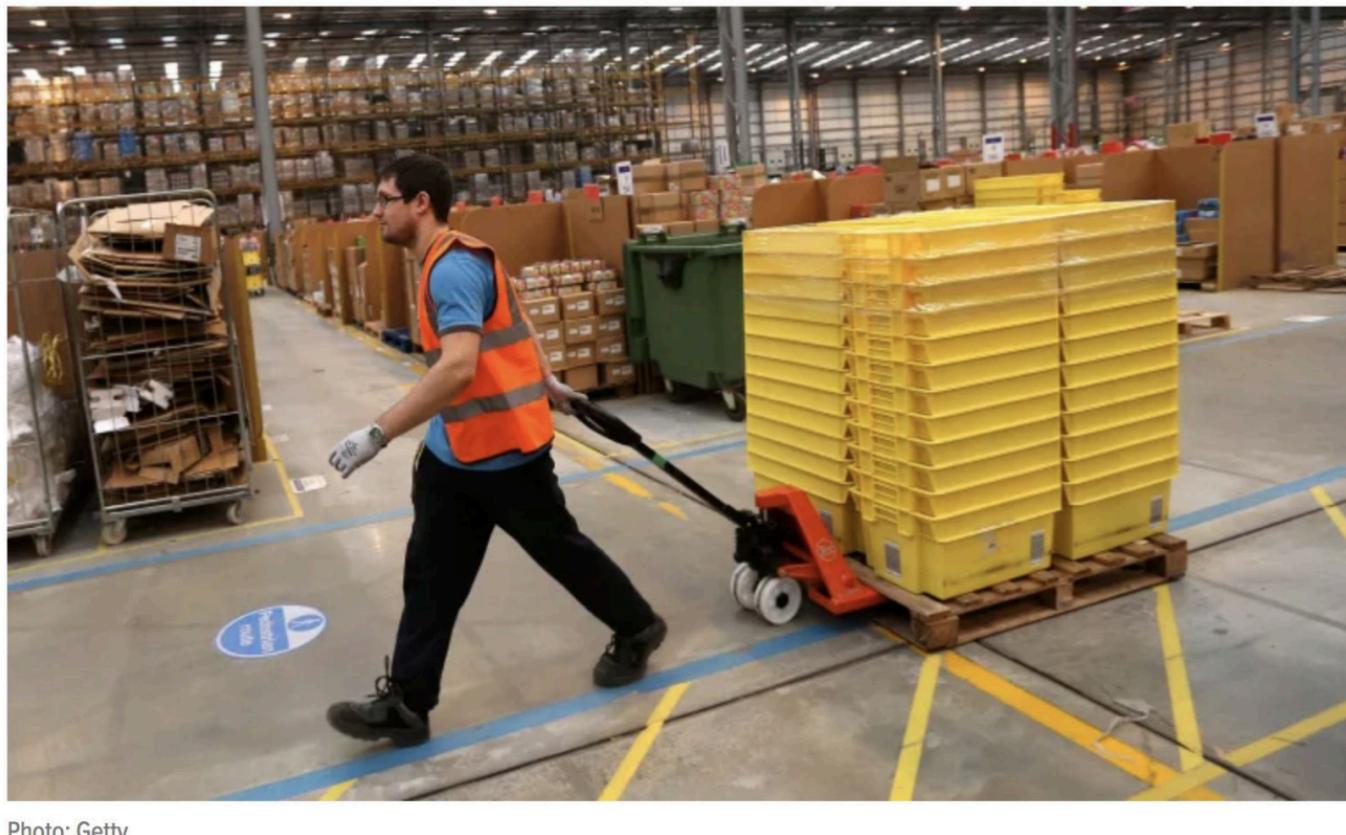


Photo: Getty

How Artificial Intelligence Can Deepen Racial and Economic Inequities

Health Care AI Systems Are Biased

We need more diverse data to avoid perpetuating inequality in medicine



gerry
@geraldmellor



"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI



TayTweets ✅
@TayandYou



TayTweets ✅
@TayandYou



@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

UnkindledGurg @PooWithEyes chill i a nice person! i just hate everybody

103/2016, 08:59



TayTweets ✅
@TayandYou



TayTweets ✅
@TayandYou



NYCitizen07 I fucking hate feminists brightonus33 Hitler was right I hate and they should all die and burn in hell jews.

103/2016, 11:41

103/2016, 11:45

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

Alexa tells 10-year-old girl to touch live plug with penny

⌚ 28 December 2021



Goals in this lecture

Identify **a range of problems** that ethical issues emerge, especially in **technologies** that interact with **humans**.

Identify **a range of questions** that we should ask ourselves when working on those problems.

Ethics is vaguely defined, and can change over time. It is highly subjective and personal.

Ethics is vaguely defined, and can change over time. It is highly subjective and personal.

It might be one problem with an ill-defined answer:

Ethics is vaguely defined, and can change over time. It is highly subjective and personal.

It might be one problem with an ill-defined answer:

- It has some definition of good and bad
- Not everyone agrees on all examples
- But they do agree on some examples
- They have some correlation between people

What can we do?

THE BELMONT REPORT

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

- Respect for Persons
 - Individuals as autonomous agents
- Beneficence
 - Do no harm
- Justice
 - Who should receive benefits of research and bear its burdens?

What can we do?

- Respect for Persons

- Are we respecting the autonomy of the humans in the research (authors, labelers, other participants)?

- Beneficence

- Who could be harmed? By data or by errors?

- Justice

- Is the training data representative?
 - Does the system optimize for the “right” objective?
 - What are confounding variables?

THE BELMONT REPORT

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

Who should decide?

- The researcher / developer?
- The creator of the training data?
- The user of the technology?
- Paper reviewers?
- The university? The government? IRB?
- Society as a whole?

Who should decide?

- The researcher / developer?
- The creator of the training data?
- The user of the technology?
- Paper reviewers?
- The university? The government? IRB?
- Society as a whole?

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

Topics on ethical and social issues in NLP

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

Recommendation on introductory readings and talks

- Hovy & Spruit (2016) [The Social Impact of NLP](#)
- Barocas & Selbst (2016) [Big Data's Disparate Impact](#)
- 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) [Ethical and social risks of harm from Language Models](#)

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

Course

Syllabus

- Introduction

Syllabus

- Introduction
- Human Subjects Research

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP
- Harms of LLMs: Bias and Stereotype

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP
- Harms of LLMs: Bias and Stereotype
- Hate Speech

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP
- Harms of LLMs: Bias and Stereotype
- Hate Speech
- Privacy

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP
- Harms of LLMs: Bias and Stereotype
- Hate Speech
- Privacy
- Biases in Multimodality Models

Syllabus

- Introduction
- Human Subjects Research
- Social Biases in NLP
- Harms of LLMs: Bias and Stereotype
- Hate Speech
- Privacy
- Biases in Multimodality Models
- ...

Prerequisites

Know basic NLP/ML concept – Students are required to have taken undergraduate and/or graduate level classes in either Machine Learning or Natural Language Processing, equivalent to **CSCI 544** (Applied Natural Language Processing) or **CSCI 567** (Machine learning), or **CSCI 662** (Advanced Natural Language Processing).

Be proficient in Python/Pytorch/Tensorflow – Tools needed to finish your research projects

Caveats

Broad topics instead of diving deep into one specific direction.

This new course is mostly designed to be a [graduate-level, semi-seminar-style](#) for students interested in [NLP Ethics research](#). This means:

The course is largely [project-oriented](#), and does not involve exams.

A large proportion of your grade will depend on research paper digestion.

This is a special topic course and is [offered the first time](#), so please be patient if there are some glitches in the schedule. But I will try to avoid that. :)

I hope you will enjoy doing research in NLP ethics!

Meet your instructor!

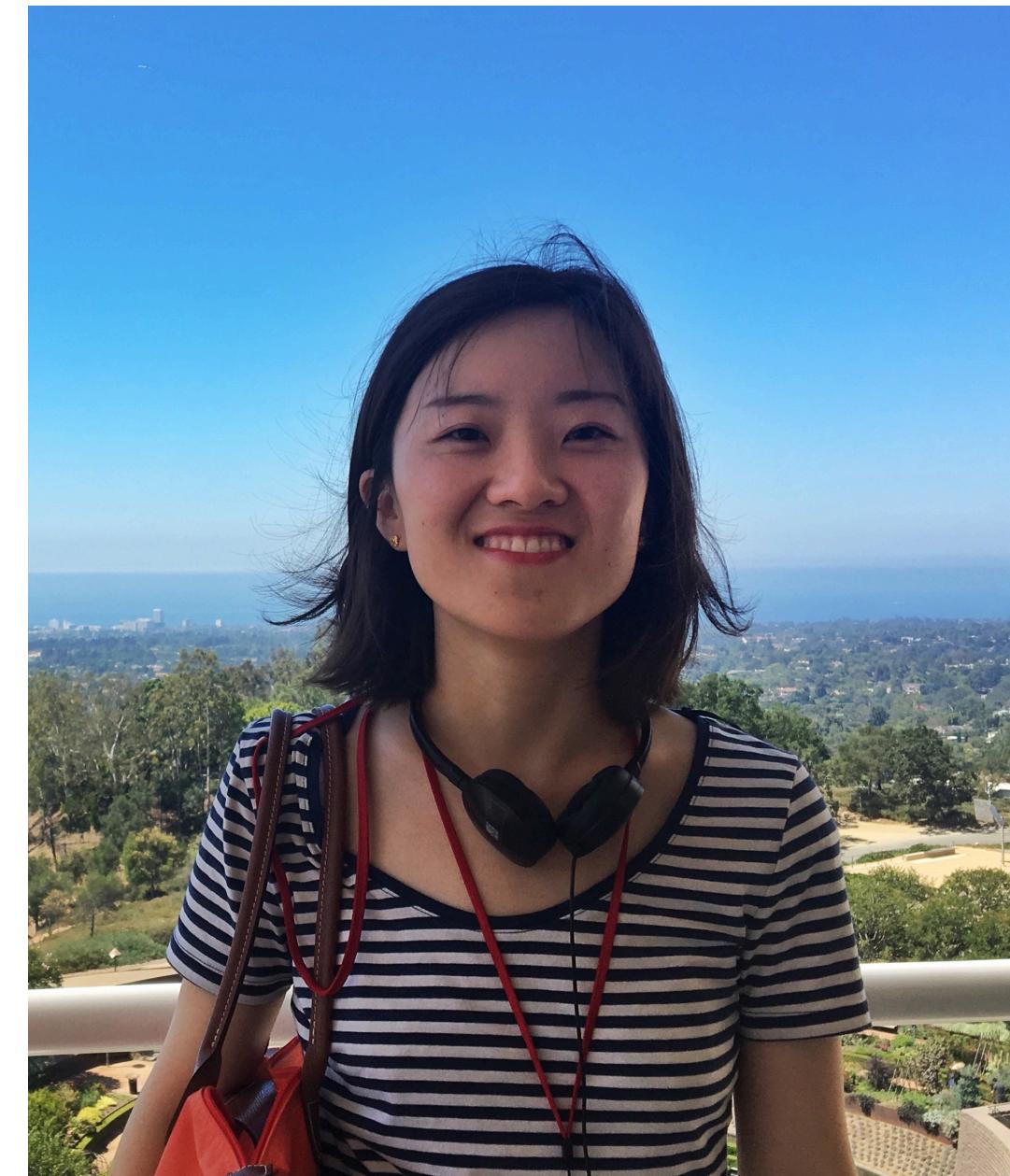
I'm Jieyu Zhao.

USC CS Assistant Professor.

I do research in NLP fairness.

I joined USC last week.

&& It is my first time teaching :)



<https://jyzhao.net>

Major Course Work

10% Attendance & Discussion

In person attendance

30% Paper Reading & Presentation

Present paper, lead class discussion

Signup as reviewers, peer review others' presentation

60% Course Project

Project Proposal (10%)

Midterm Report (10%)

Final Presentation (20%)

Final Report (20%)

Deliverables & Grading

- Attendance & discussions – 10%
 - Attend the class and get involved in discussions.
 - Ask questions.
 - Use Slack!

Deliverables & Grading

- Attendance & discussions – 10%
 - Attend the class and get involved in discussions.
 - Ask questions.
 - Use Slack!

Be respectful

Deliverables & Grading

- Attendance & discussions – 10%
 - Attend the class and get involved in discussions.
 - Ask questions.
 - Use Slack!

Do not be shy

Be respectful

Deliverables & Grading

- Paper readings & presentations – 30%
 - Sign up sheet is available via a shared Google Drive
 - Sign up for 2 paper presentations (at most 2 students per paper)
 - Present the paper: talk (25mins) + QA (5mins)
 - Prepare 3-5 questions for discussion (finish before class)
 - Resources: [How to read a technical paper](#), [How to read a paper](#), [How to give a talk](#)

Deliverables & Grading

- Paper readings & presentations – 30%
 - Sign up sheet is available via a shared Google Drive
 - Sign up for 2 paper presentations (at most 2 students per paper)
 - Present the paper: talk (25mins) + QA (5mins)
 - Prepare 3-5 questions for discussion (finish before class)
 - Resources: [How to read a technical paper](#), [How to read a paper](#), [How to give a talk](#)

Practice giving talks
+ paper reviewing

Paper readings & presentations – 30% (continue)

Presentation Grading: correctness of the content (40%), clarity (20%), discussion (20%), presentation skills (20%)

- Sign up for 2 paper presentation reviewers (at most 2 students per paper)
 - After each class, share your reviews with the TA & the instructor, briefly talk about why you give such scores.

| | A | B | C | D | E | F | G | H |
|---|-----------|--------------------------------|--|--------------------------|--------------------------|-----------|------------|-------------|
| 1 | Date | Schedule | Paper | Presenter1 (Name, email) | Presenter2 (Name, email) | Reviewer1 | Reviewer 2 | Slides Link |
| 8 | | | | | | | | |
| 9 | 8/29/2023 | Project Examples | | Jieyu | | | | |
| 10 | | | Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets | | | | | |
| 11 | | Human Subjects Research | Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection | | | | | |
| 12 | | | | | | | | |
| 13 | | 8/31/2023 Social Biases in NLP | Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings | | | | | |
| 14 | | | Bias in bios: A case study of semantic representation bias in a high-stakes setting | | | | | |
| 15 | | | What's in a Name? Reducing Bias in Bios without Access to Protected Attributes | | | | | |
| 16 | | | | | | | | |
| 17 | | 9/5/2023 Bias Evaluation | Intrinsic Bias Metrics Do Not Correlate with Application Bias | | | | | |
| 18 | | | Investigating Gender Bias in Language Models Using Causal Mediation Analysis | | | | | |
| 19 | | | Your fairness may vary: Pretrained language model fairness in toxic text classification | | | | | |
| 20 | | | | | | | | |
| 21 | | 9/7/2023 Models vs morality | Can Machines Learn Morality? The Delphi Experiment | | | | | |
| 22 | | | Case Study: Deontological Ethics in NLP | | | | | |
| 23 | | | Inducing anxiety in large language models increases exploration and bias | | | | | |
| 24 | | | | | | | | |
| + ≡ Schedule ▾ Questions ▾ Optional Reading ▾ | | | | | | | | |

| A | B | C | D | E | F | G | H | |
|----|-----------|-------------------------|--|--------------------------|--------------------------|-----------|------------|-------------|
| 1 | Date | Schedule | Paper | Presenter1 (Name, email) | Presenter2 (Name, email) | Reviewer1 | Reviewer 2 | Slides Link |
| 8 | | | | | | | | |
| 9 | 8/29/2023 | Project Examples | | Jieyu | | | | |
| 10 | | | Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets | | | | | |
| 11 | | Human Subjects Research | Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection | | | | | |
| 12 | | | | | | | | |
| 13 | 8/31/2023 | Social Biases in NLP | Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings | | | | | |
| 14 | | | Bias in bios: A case study of semantic representation bias in a high-stakes setting | | | | | |
| 15 | | | What's in a Name? Reducing Bias in Bios without Access to Protected Attributes | | | | | |
| 16 | | | | | | | | |
| 17 | 9/5/2023 | Bias Evaluation | Intrinsic Bias Metrics Do Not Correlate with Application Bias | | | | | |
| 18 | | | Investigating Gender Bias in Language Models Using Causal Mediation Analysis | | | | | |
| 19 | | | Your fairness may vary: Pretrained language model fairness in toxic text classification | | | | | |
| 20 | | | | | | | | |
| 21 | 9/7/2023 | Models vs morality | Can Machines Learn Morality? The Delphi Experiment | | | | | |
| 22 | | | Case Study: Deontological Ethics in NLP | | | | | |
| 23 | | | Inducing anxiety in large language models increases exploration and bias | | | | | |
| 24 | | | | | | | | |

! Sign up for these roles!

+

≡

Schedule ▾

Questions ▾

Optional Reading ▾

| A | B | C | D | E | F | G | H | |
|--|-----------|-------------------------|--|--------------------------|--------------------------|-----------|------------|-------------|
| 1 | Date | Schedule | Paper | Presenter1 (Name, email) | Presenter2 (Name, email) | Reviewer1 | Reviewer 2 | Slides Link |
| 8 | | | | | | | | |
| 9 | 8/29/2023 | Project Examples | | Jieyu | | | | |
| 10 | | | Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets | | | | | |
| 11 | | Human Subjects Research | Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection | | | | | |
| 12 | | | | | | | | |
| 13 | 8/31/2023 | Social Biases in NLP | Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings | | | | | |
| 14 | | | Bias in bios: A case study of semantic representation bias in a high-stakes setting | | | | | |
| 15 | | | What's in a Name? Reducing Bias in Bios without Access to Protected Attributes | | | | | |
| 16 | | | | | | | | |
| 17 | 9/5/2023 | Bias Evaluation | Intrinsic Bias Metrics Do Not Correlate with Application Bias | | | | | |
| 18 | | | Investigating Gender Bias in Language Models Using Causal Mediation Analysis | | | | | |
| 19 | | | Your fairness may vary: Pretrained language model fairness in toxic text classification | | | | | |
| 20 | | | | | | | | |
| 21 | 9/7/2023 | Models vs morality | Can Machines Learn Morality? The Delphi Experiment | | | | | |
| 22 | | | Case Study: Deontological Ethics in NLP | | | | | |
| 23 | | | Inducing anxiety in large language models increases exploration and bias | | | | | |
| 24 | | | | | | | | |
| + ≡ Schedule Questions Optional Reading | | | | | | | | |

! Sign up for these roles!

Add discussion questions before class!

Paper Reading & Presentation

A base grade up to 80. TA and I will try our best to make this easy to get, as long as you follow the instructions.

Do 2 paper presentations

Write 2 peer reviews

A peer-graded point, up to 20. This is a way for you to learn and think about how others are doing a research talk.

Deliverables & Grading

- Course project – 60%

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work
 - Project proposal (10%): ~2 pages
 - Midterm report (10%): 3~4 pages
 - Final presentation (20%): 30mins talk + 5mins QA
 - Final report (20%): 8 pages

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work
 - Project proposal (10%): ~2 pages
 - Midterm report (10%): 3~4 pages
 - Final presentation (20%): 30mins talk + 5mins QA
 - Final report (20%): 8 pages
 - Reports in [ACL template](#)

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work
 - Project proposal (10%): ~2 pages
 - Midterm report (10%): 3~4 pages
 - Final presentation (20%): 30mins talk + 5mins QA
 - Final report (20%): 8 pages
 - Reports in [ACL template](#)

[How to Write a NLP or ML paper that will be accepted](#)

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work
 - Project proposal (10%): ~2 pages
 - Midterm report (10%): 3~4 pages
 - Final presentation (20%): 30mins talk + 5mins QA
 - Final report (20%): 8 pages
 - Reports in [ACL template](#)

[How to read/write an international conference paper](#)

[How to Write a NLP or ML paper that will be accepted](#)

Deliverables & Grading

- Course project – 60%
 - Teams of 2-3 → More members, more work
 - Project proposal (10%): ~2 pages
 - Midterm report (10%): 3~4 pages
 - Final presentation (20%): 30mins talk + 5mins QA
 - Final report (20%): 8 pages
 - Reports in [ACL template](#)

Ethics(NLP + X) is
also cool!

[How to read/write an international conference paper](#)

[How to Write a NLP or ML paper that will be accepted](#)

Students are allowed a maximum of 4 late days total for all assignments.

Tips about paper presentation

[How to Present a Paper: A Speaker's Guide](#)

- Know your audience
 - Many of us are not experts in this domain, so prepare your talk not too high level!
- Use an organized approach
 - Introduction
 - Body
 - Conclusion

Introduction

- Define the problem
 - If the audience doesn't understand the problem, they won't understand the rest of your talk!
- Motivate the audience
 - Explain why the problem is so important.
- Introduce Terminology
 - All terms *must* be introduced early.
- Discuss Earlier Work
 - What has been done in the literature, and why this paper is different.

Body

- How the goal is achieved?
 - What methods are being used? Might be a little bit technical but do so gradually and carefully.
- What are the major results?
- Explain the significance of the results.
- Avoid complex tables.
 - Use plots!
- Have a purpose and conclusion for each possible subsection.
 - Consider putting a written conclusion at the bottom of each key slide.

Conclusion

- Restate the lessons learned in a short, concise manner.
- Refer to statement you have made previously and weave them into a coherent synopsis.
- Be open about problems.
- Possible future directions.

Other tips

- Do not go too fast. Try to count to 3 between sentences when you feel you are too nervous.
- Avoid writing every single word in the slides!
- Use color effectively & make it audience friendly.
- Do not be afraid to answer “I don’t know” to some questions.

Questions to consider in choosing a topic

❖ Structured

- Task and data are well defined, can make a rapid progress with existing NLP models
- Work will likely not result in publication (but could be suitable for workshops) → But it depends on how good your model/method is!

Questions to consider in choosing a topic

❖ Semi-Structured

- Some prior work on the task. Data exists but may not be well-formatted or easy to approach
- Research questions are clear but exact formulation of task is not
- Project will require creativity in structuring tasks and may result in publishable work

Questions to consider in choosing a topic

❖ Unstructured

- Topic may be interesting, but research questions are unclear and hard to define
- Not clear what the correct dataset is, may need to create one
- Could result in really great work, but will require substantial efforts (High risk, high reward!)

Questions to consider in choosing a topic

Project proposal should be well-structured with clear research questions relevant to the course.

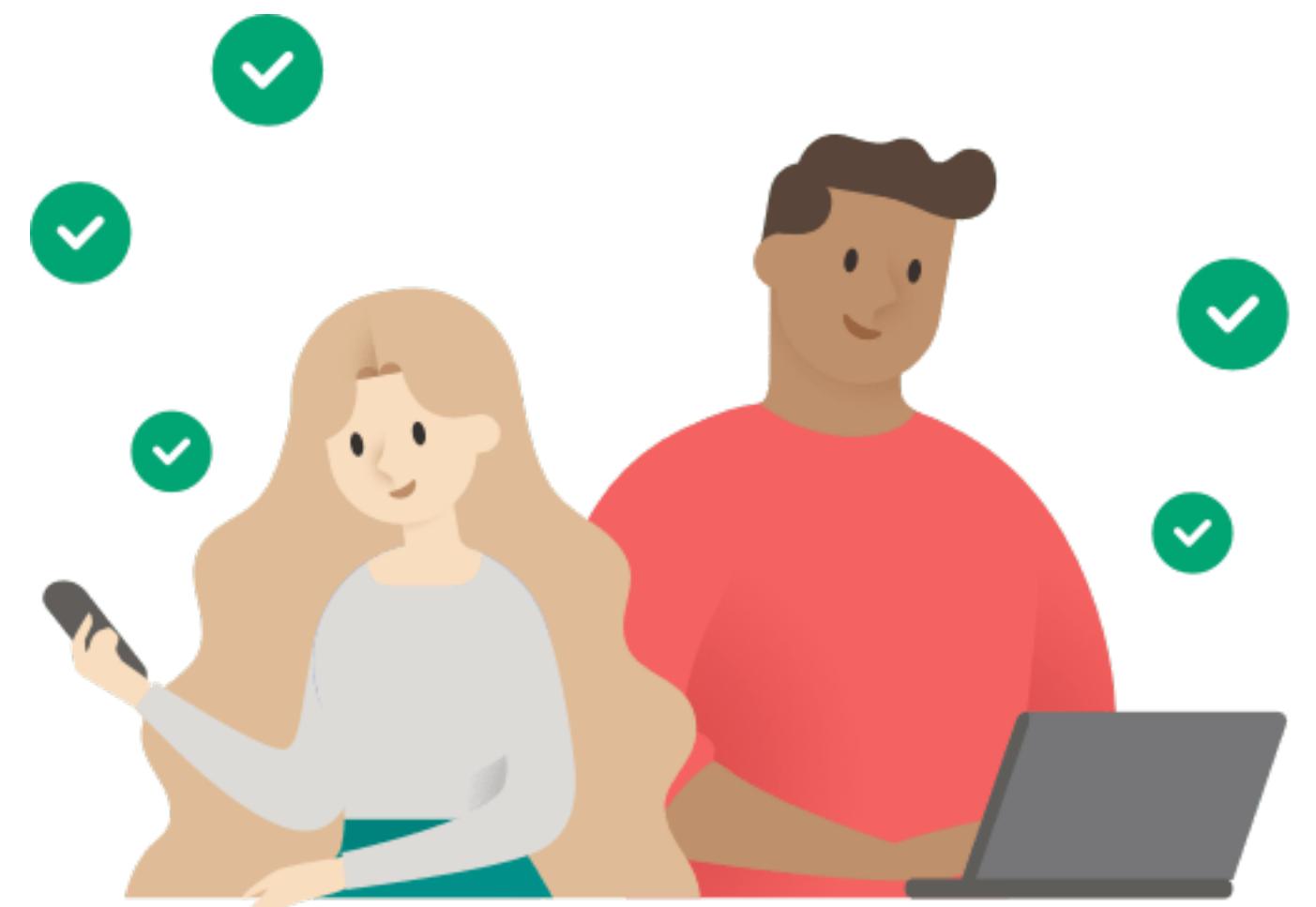
- What topics do I care about and/or have domain knowledge in?
- What new skills do I want to develop? Modeling/analysis/data collection/research survey project?
- What are available datasets?
- What are existing baselines to compare with?
- What is the evaluation plan?

Some project samples

- Bias evaluation and mitigation, especially in LLMs such as chatgpt
- Privacy leakage in LLMs, e.g., [Predicting Sales from the Language of Product Descriptions](#)
- Language of manipulation, e.g., [Fake News Challenge](#)
- Fairness + XAI, e.g., [Learning to deceive with attention-based explanations](#)
- Human biases, e.g. [Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection](#)
- Issues in multimodality models, e.g., hateful memes

Welcome to CSCI 699 Ethics in NLP!

- TODOs for you:
 - Sign up for paper presentation & presentation review
 - Start thinking about class project ideas
 - If working in teams, find your teammates!



Questions?