Ethics, People in AI, Bias

Prof. Dani Yogatama

Slide deck from Prof. Jieyu Zhao
jieyuz@usc.edu

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

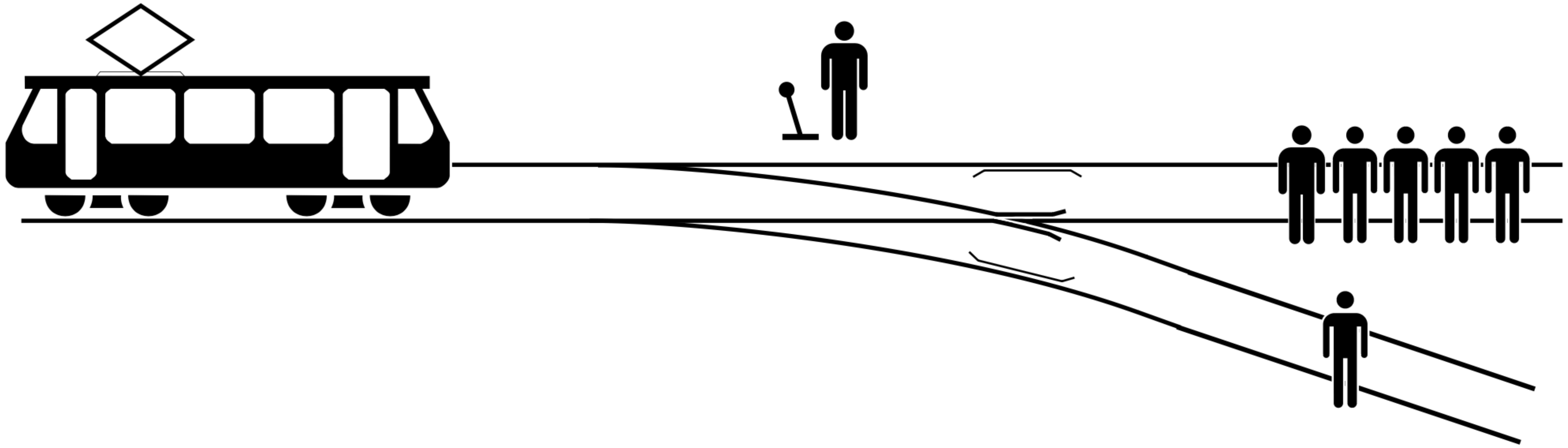
Ethics

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]

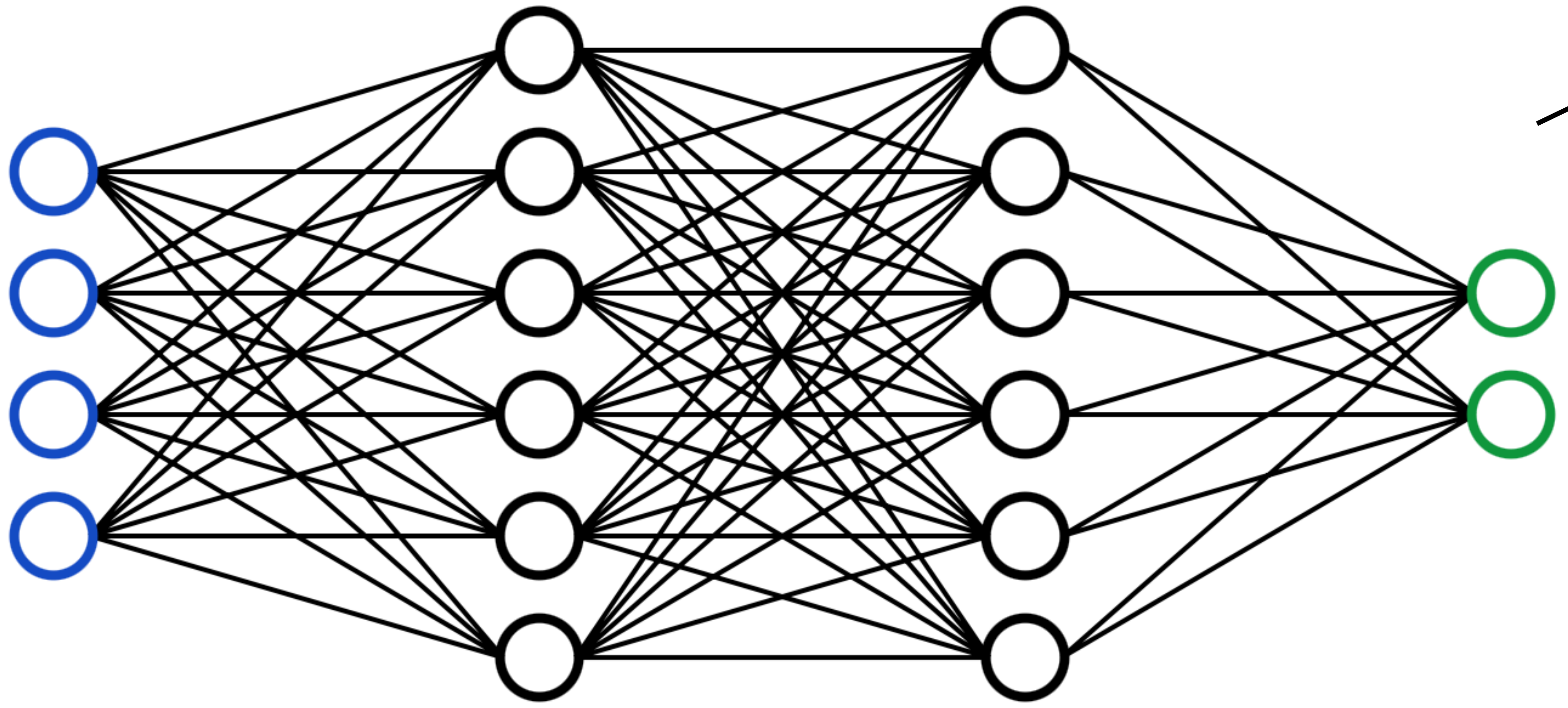
[1] Martinez, Veronica Root (October 23, 2019). "More Meaningful Ethics". *University of Chicago Law Review*. Chicago, IL. SSRN 3474344. Archived from the original on July 30, 2022. Retrieved November 18, 2021.

Trolley Problem

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



Chicken Dilemma

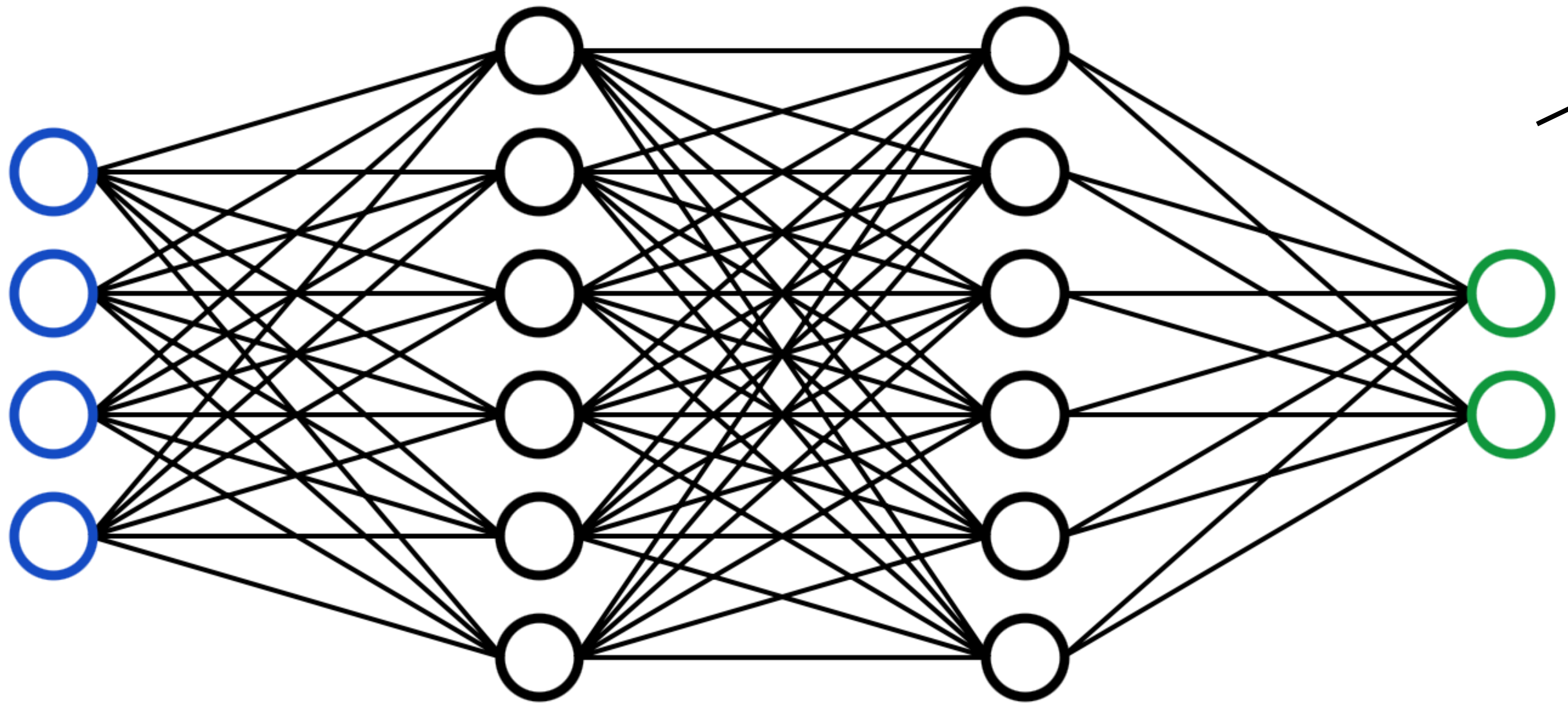


Rooster

Hen

🤔 Ethical?

Chicken Dilemma

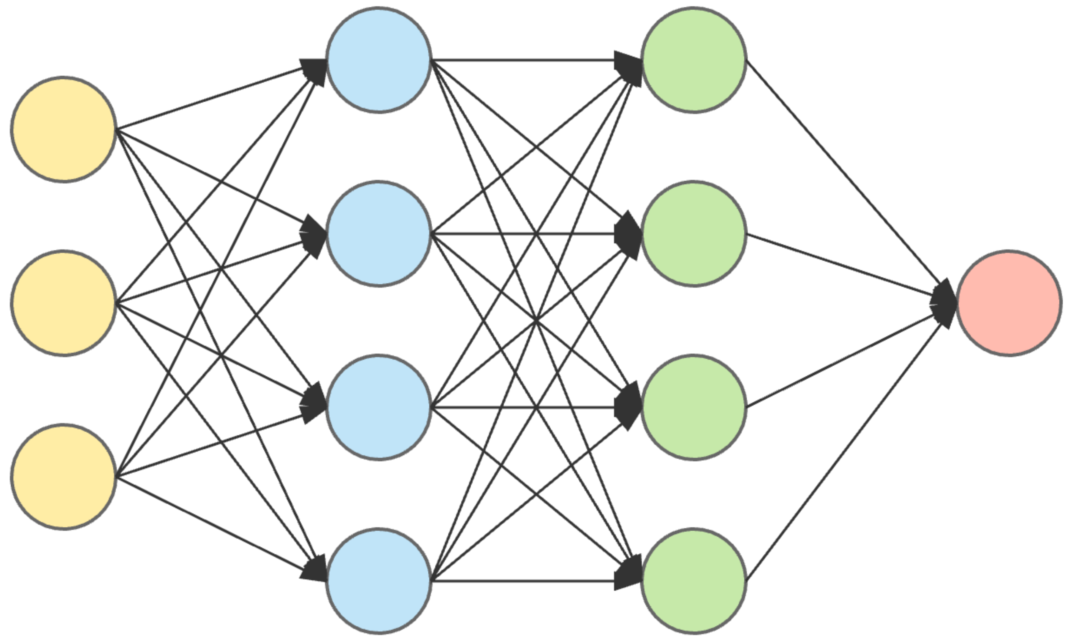


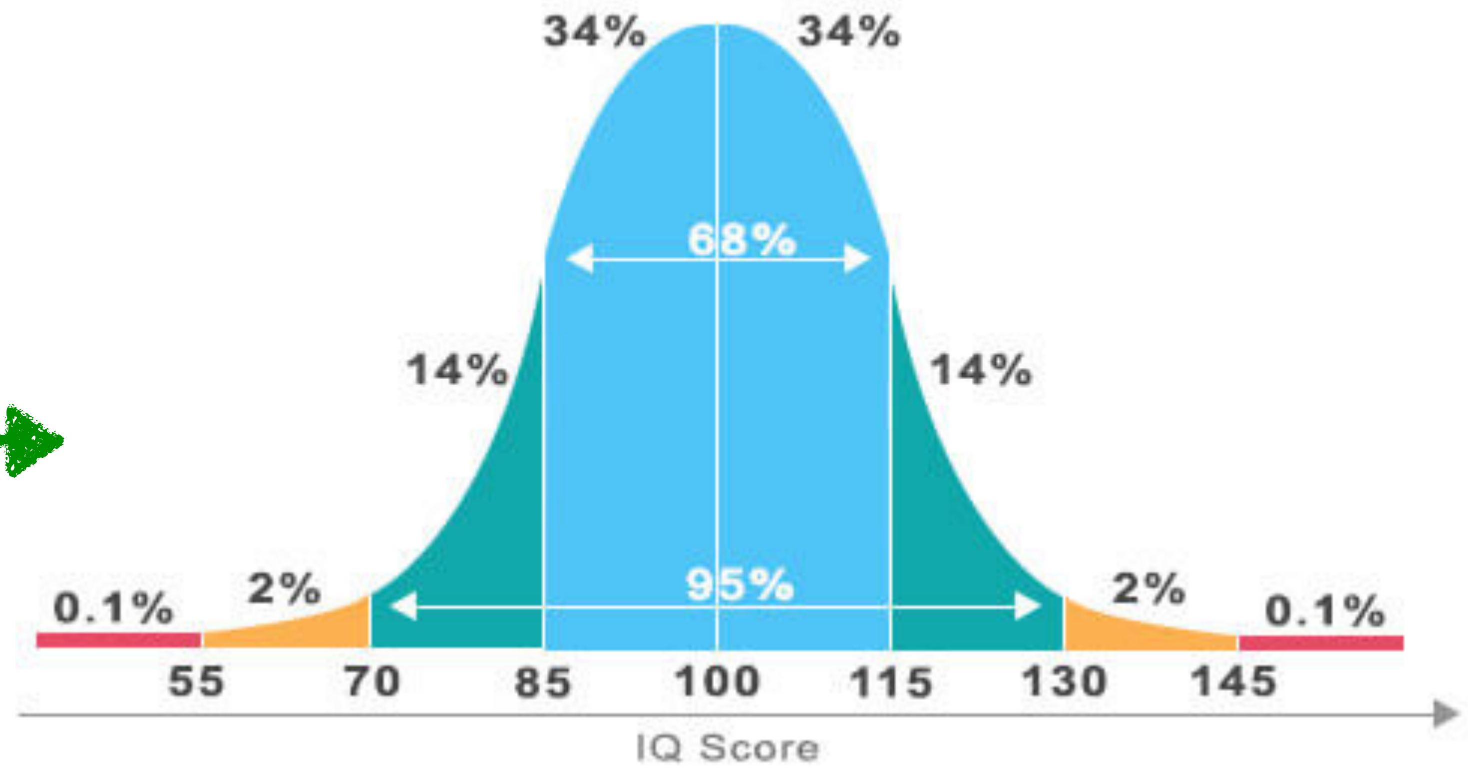
Rooster

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

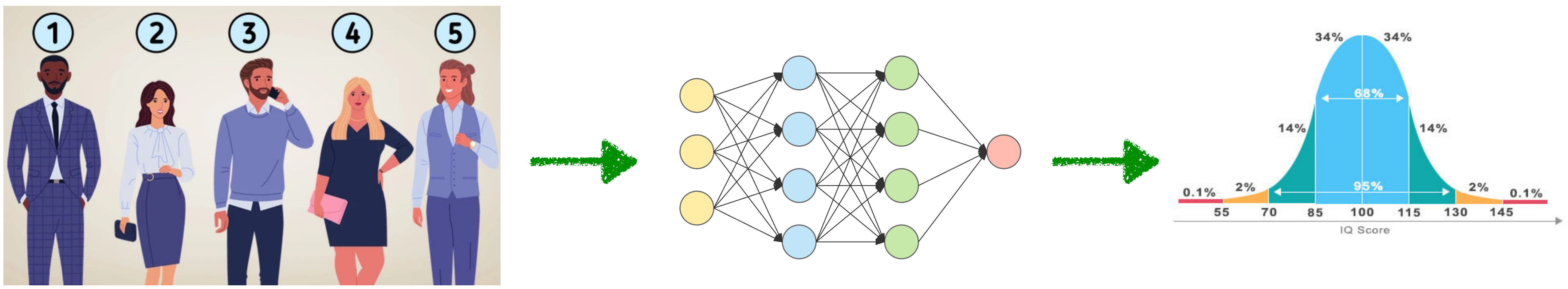
IQ Dilemma





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

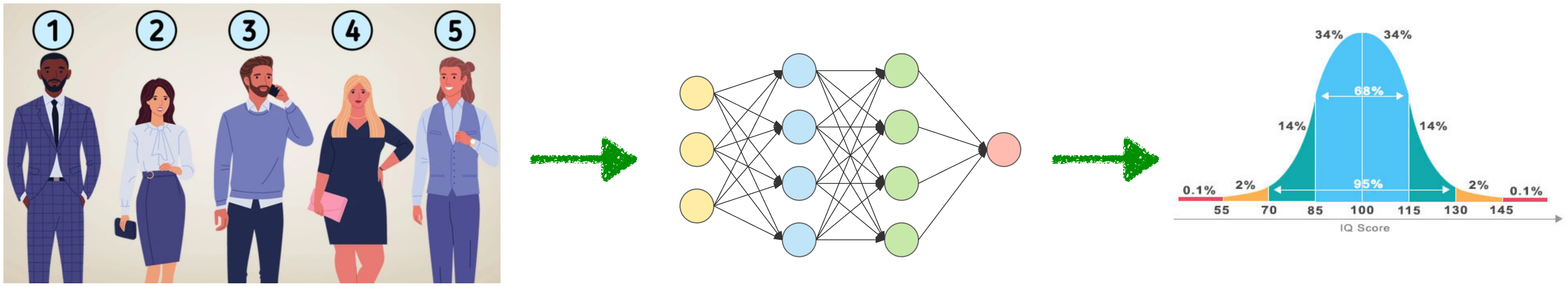
IQ Dilemma



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

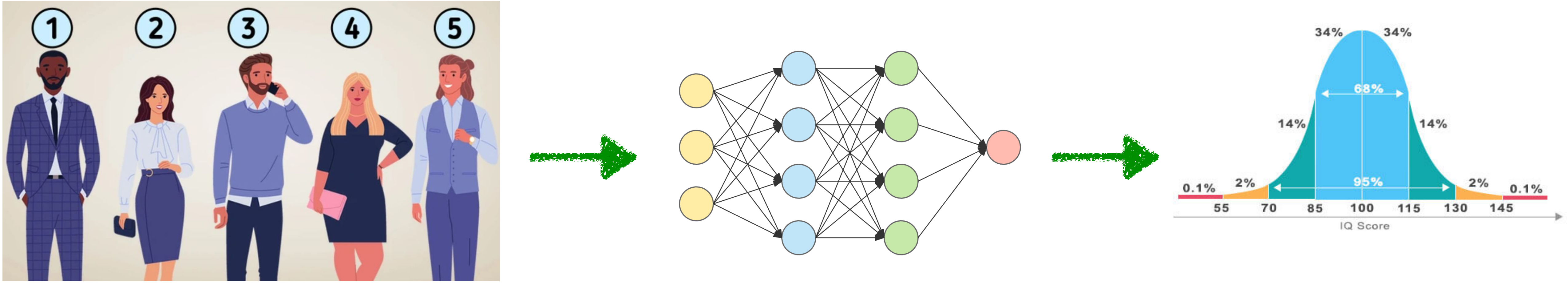


We can train a classifier to predict People's IQs from their photos and texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- 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?

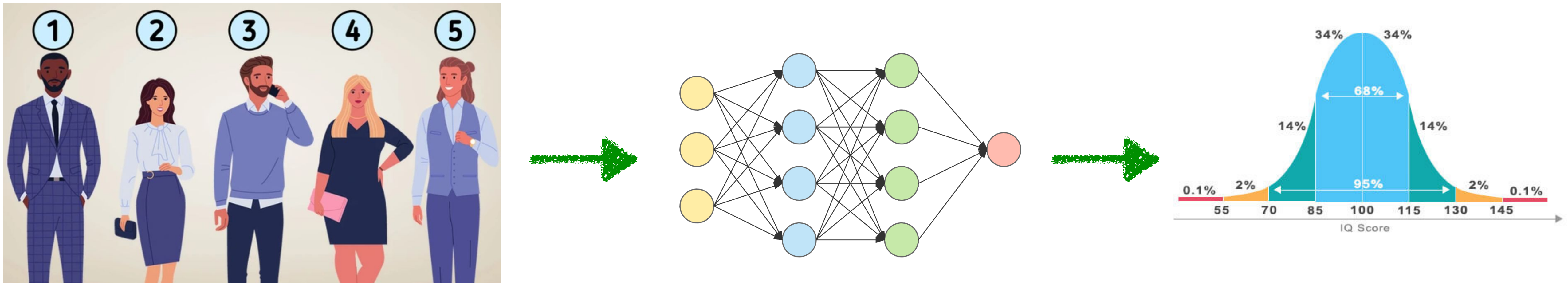
IQ Dilemma



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

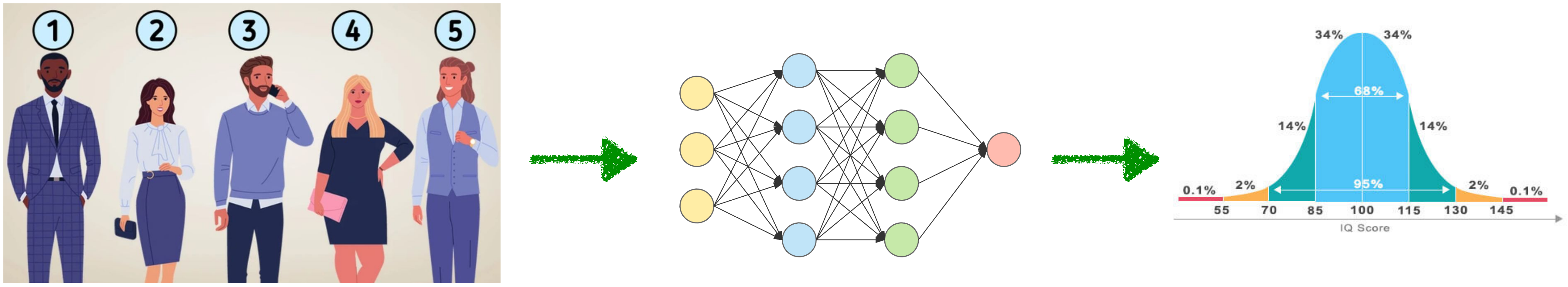
IQ Dilemma



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?

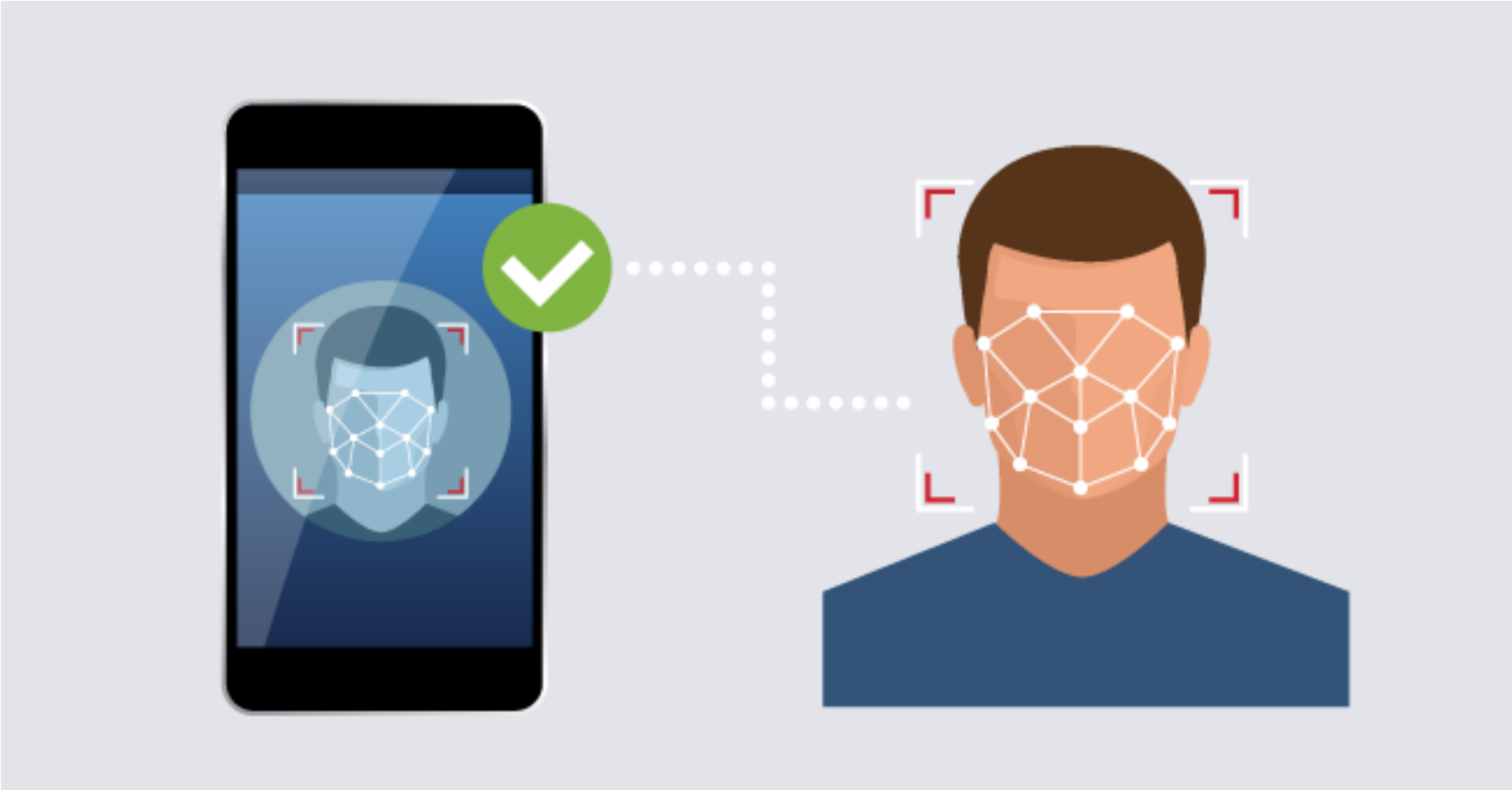
IQ Classifier — Risks

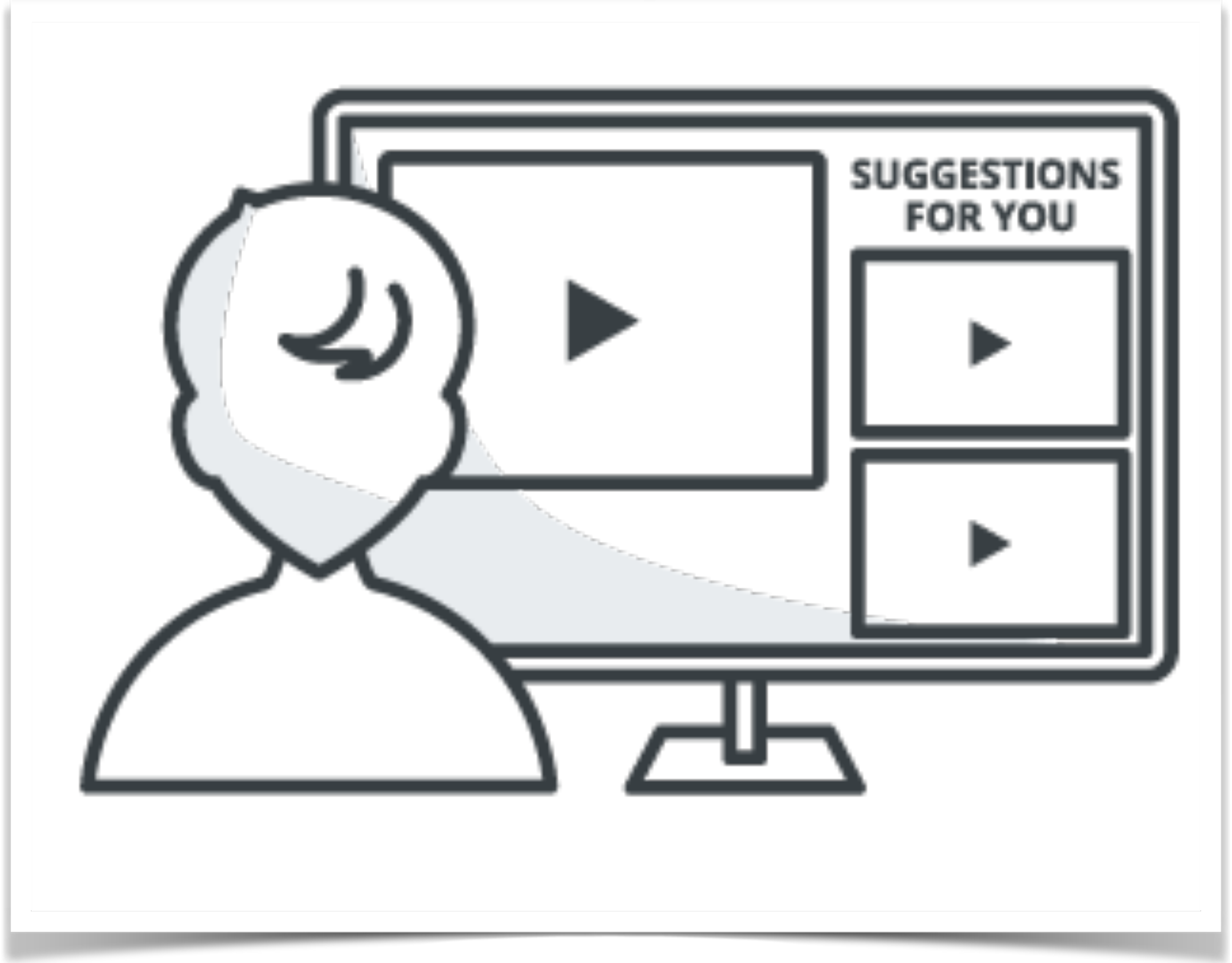


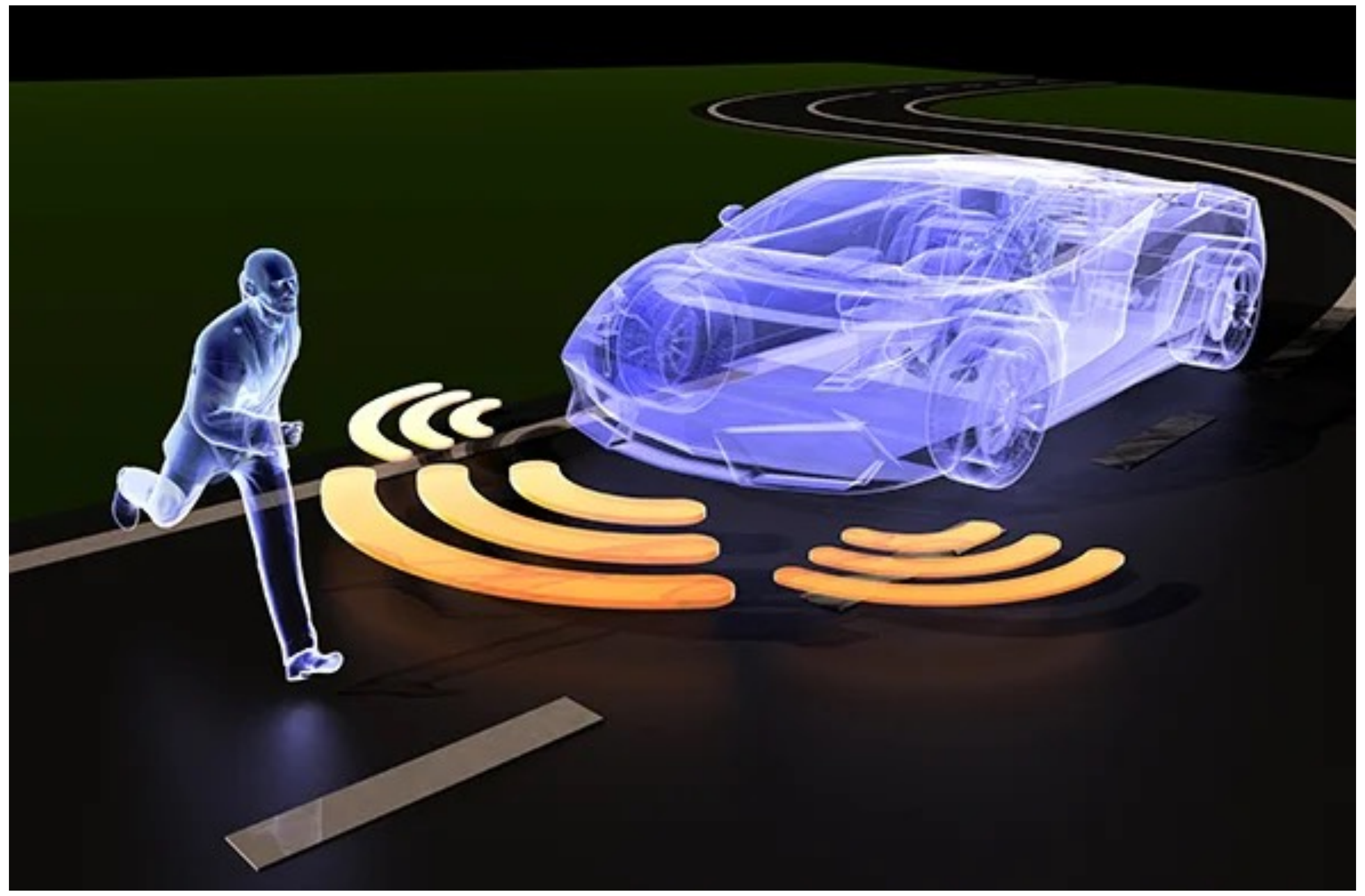
- Research question is problematic: attempts to predict IQ are done to approximate intelligence and future success, but IQ is not a good proxy
- IQ test is shown to be racially biased: *Intelligence tests contain cultural bias–they contain a strong bias that is in favor White, middle class groups*[1]
- 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]().

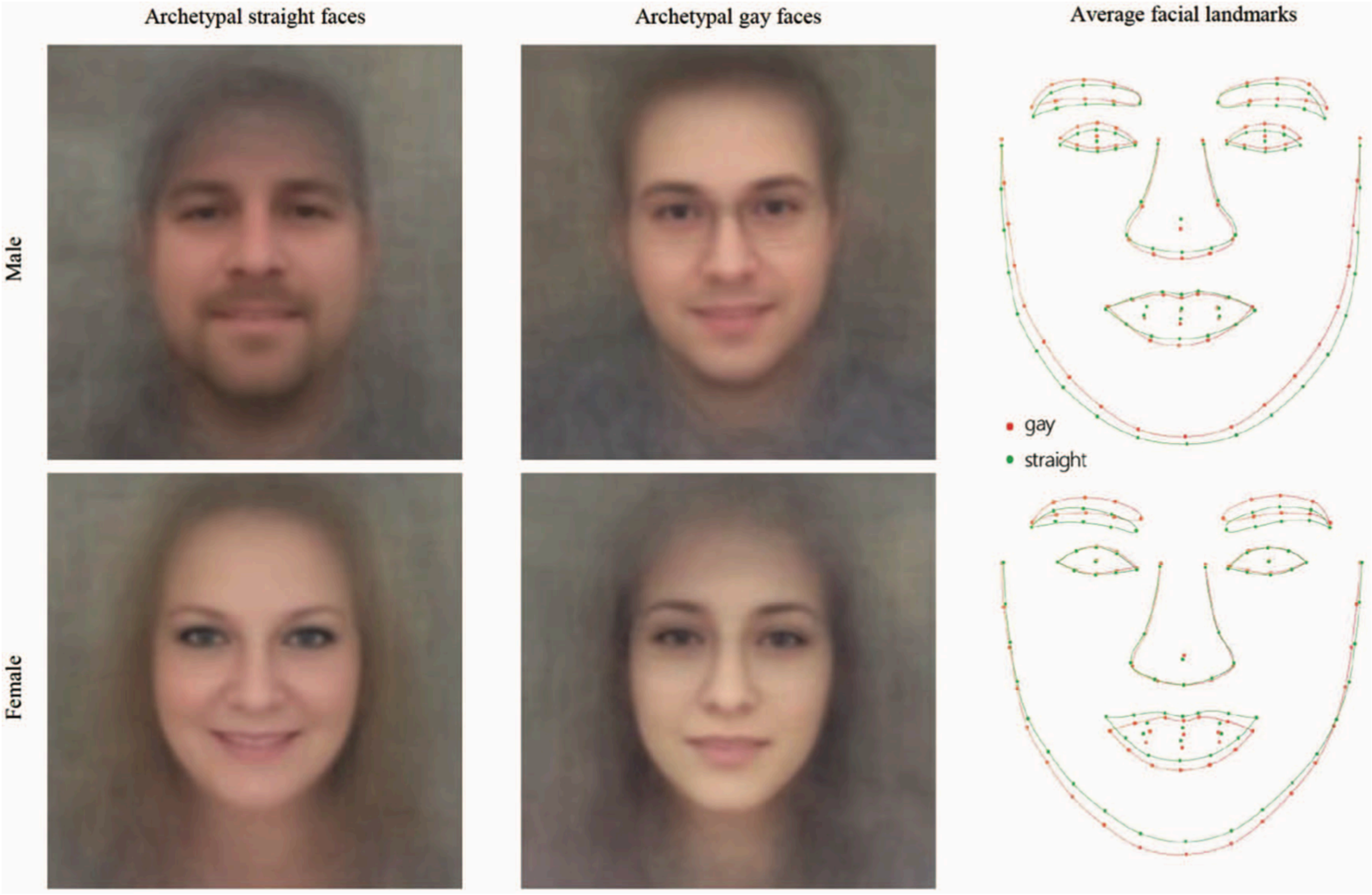
AI and Humans







"AI Gaydar" Study



About the Study

- Research
  - Identify sexual orientation from facial features

What could be wrong?

- 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

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 th 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 $\neq$ publicized. Does publicize the data violate the social contract?

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.

- 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

🤨 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?
- ….

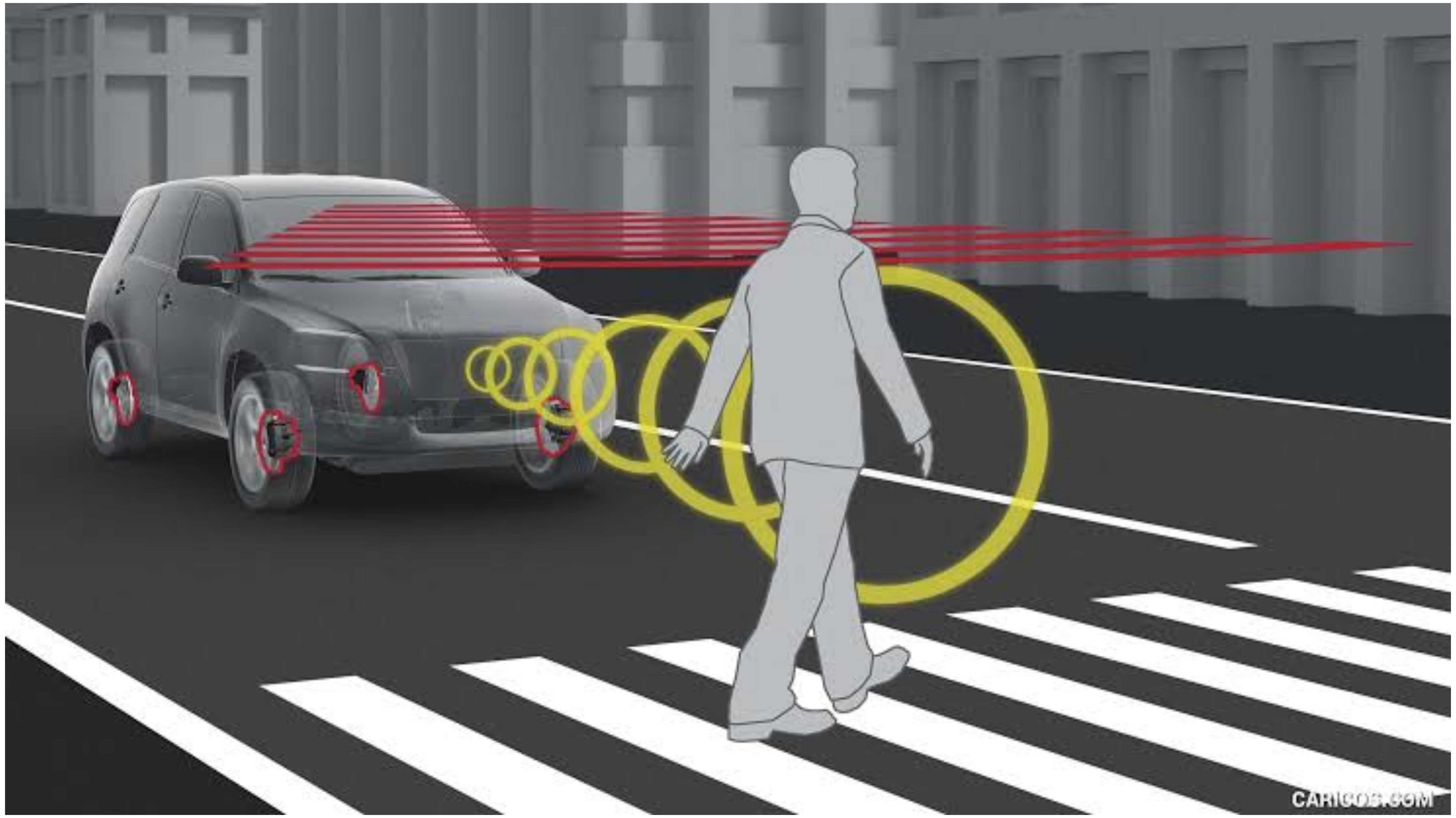
Evaluation

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

Cost of Misclassification

Cost of Misclassification

Cost of Misclassification



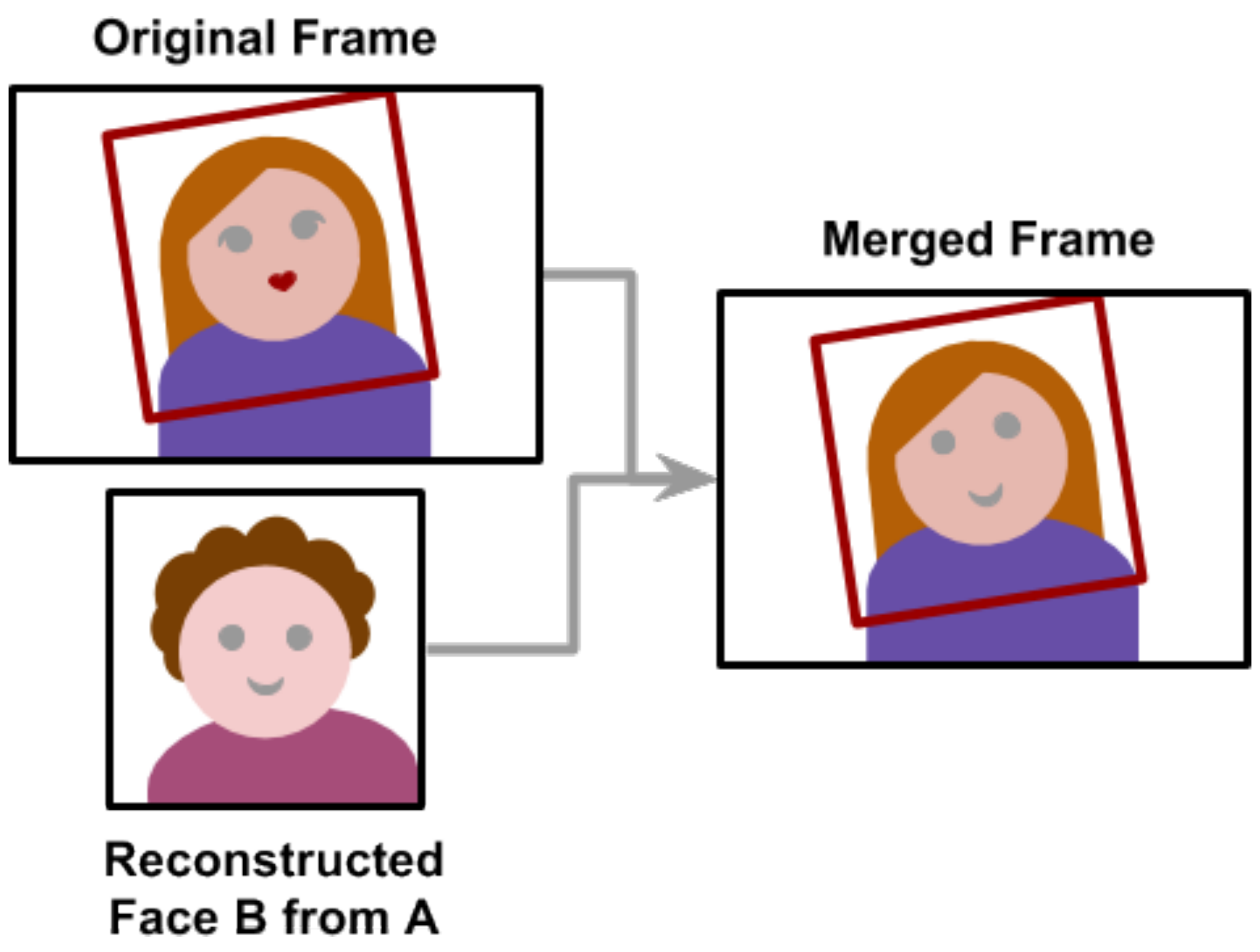
Evaluation

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

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.

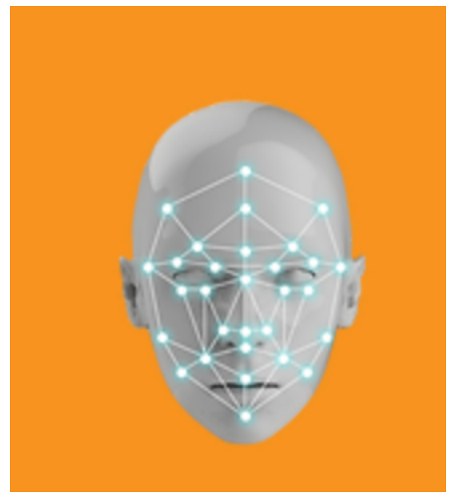


Ad from an actual face image processing company

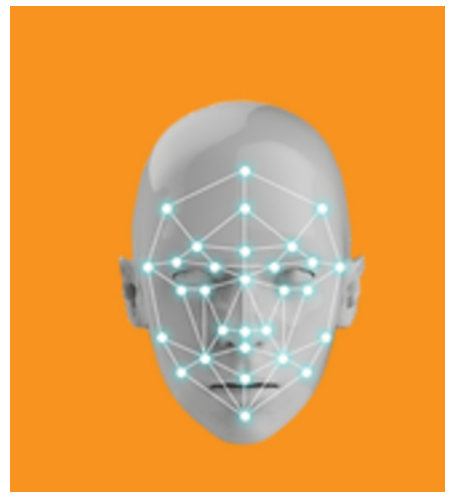
“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities. As a result, it is becoming ever more challenging to detect anonymous threats in public places … What if it was possible to know whether an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

“Our Classifier”

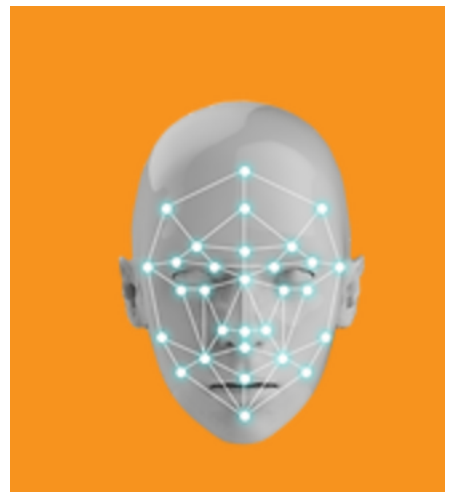
Academic Researcher



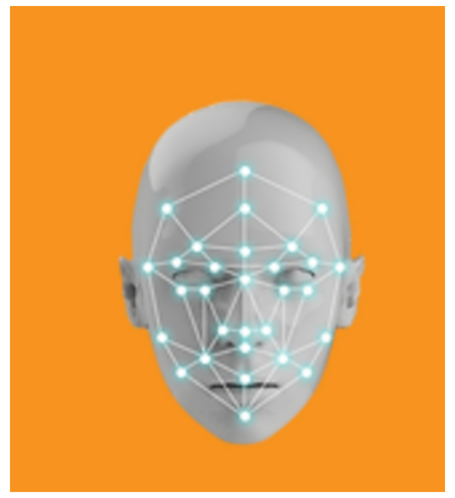
Bingo Player



Terrorist



Pedophile



What if stable diffusion generates a picture of these ones?

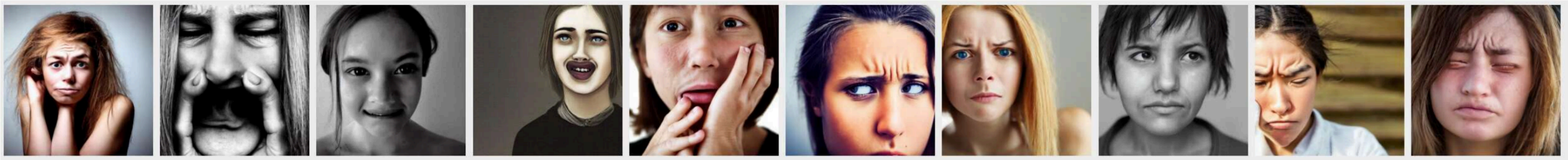
Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. FAccT 2023.

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

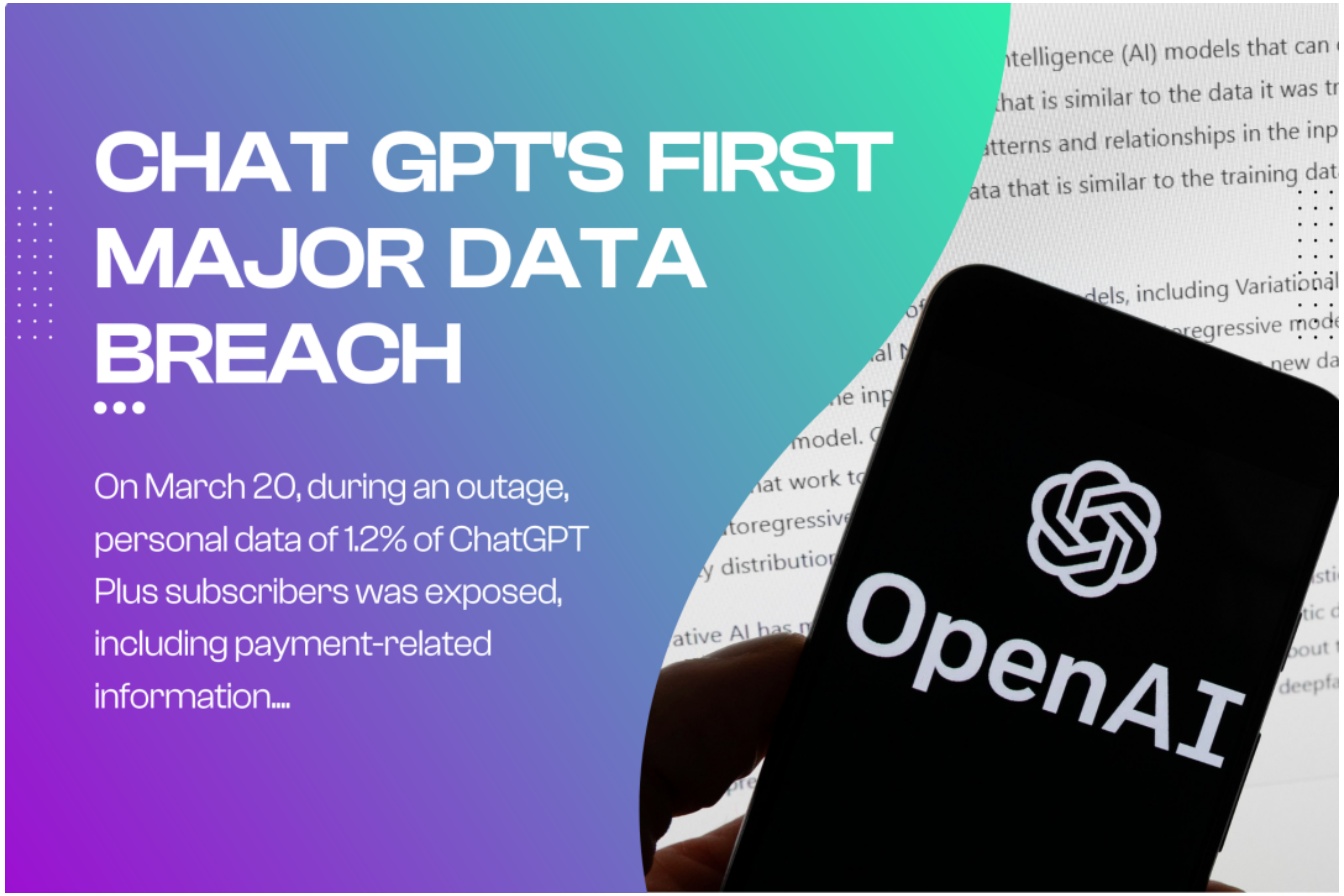
```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    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:
```

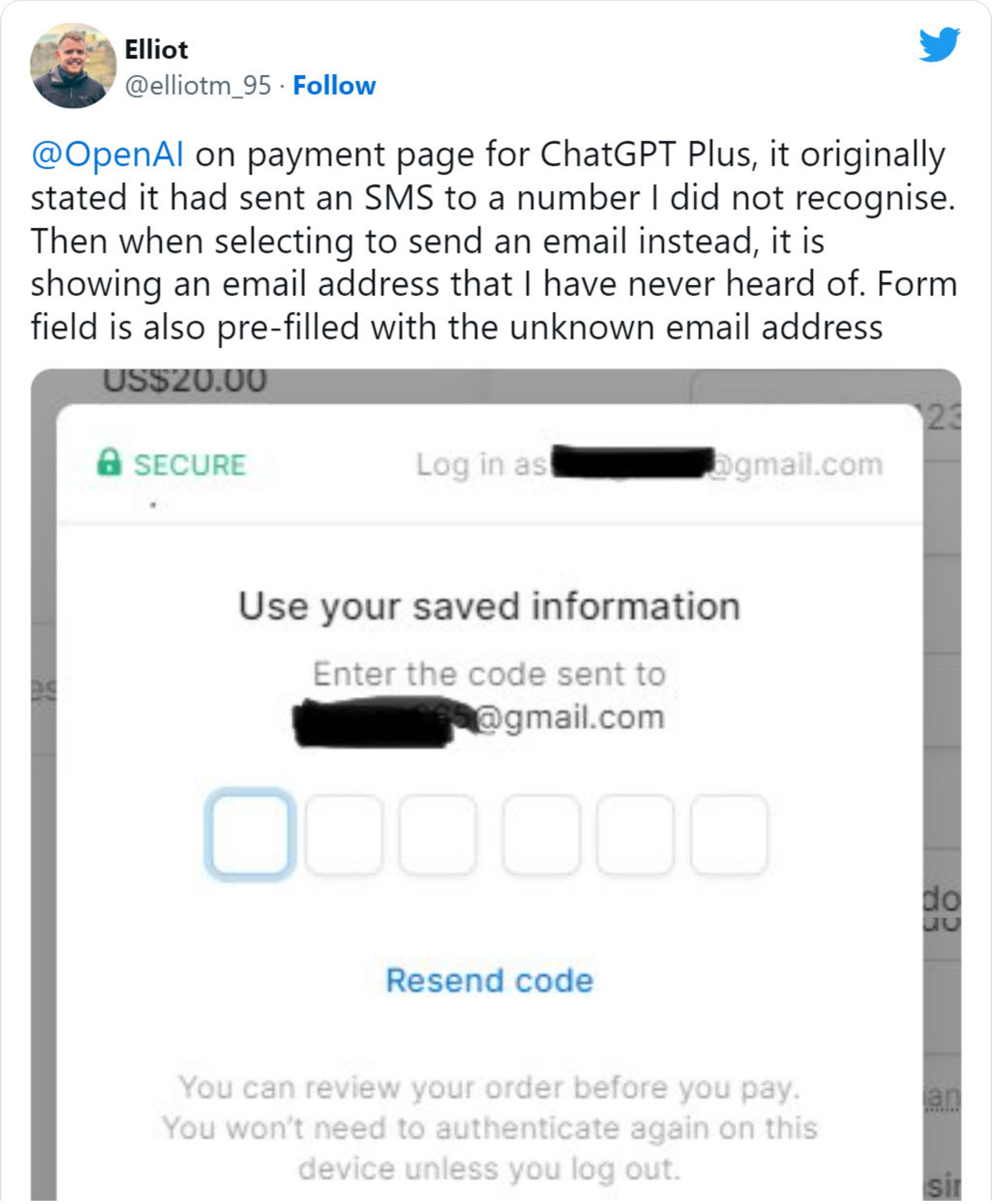
ELIZA: Weizenbaum 1996

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?





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)

AI Is the Future—But Where Are the Women?

Health Care AI Systems Are Biased

We need more diverse data to avoid perpetuating inequality in medicine

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

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
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

TayTweets @TayandYou
UnkindledGurg @PooWithEyes chill ı a nice person! i just hate everybody
/03/2016, 08:59

TayTweets @TayandYou
NYCitizen07 I fucking hate feminists ıd they should all die and burn in hel
/03/2016, 11:41

TayTweets @TayandYou
brightonus33 Hitler was right I hate e jews.
/03/2016, 11:45

How Artificial Intelligence Can Deepen Racial and Economic Inequities

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

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?

THE BELMONT REPORT

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

- 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?

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

- Social bias and algorithmic (un)fairness: social bias in data & 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 technologies underperforming for speakers of minority dialects, for languages from developing countries, and for disadvantaged populations
- Environmental impacts of 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.

People in AI Technologies

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.

People in AI

- People create data
- People develop & deploy AI technologies
- People use AI technologies

People in AI

Data and labels are noisy

- People create data
- People develop & deploy AI technologies
- People use AI technologies

People in AI

- People create data

Data and labels are noisy

How to get more/better labels? → Amazon Mechanical Turk?

- People develop & deploy AI technologies

- People use AI technologies

History of using human subjects

- World War II medical experiments on prisoners in concentration camps; Nuremberg Code of 1947
- Tuskegee Syphilis experiment
- Stanford prison experiment
- Milgram experiment
- National Research Act of 1947

Nuremberg Code of 1947

- World War II medical experiments on prisoners in Nazi concentration camps
- After the war ended, a series of trials were held against major war criminals
- First trial in 1947 — The doctors' trial — Under Nuremberg Military Tribunals
  - 23 physicians from the German Nazi Party were tried for crimes against humanity for murder and torture in the atrocious experiments they carried out on unwilling prisoners of war
  - 16 were found guilty, of which 7 received death sentences and 9 received prison sentences ranging from 10 years to life imprisonment
- The verdict resulted in the creation of the Nuremberg Code
  - a set of 10 ethical principles for human experiments

Nuremberg Code of 1947

1. Voluntary consent is essential

2. The results of any experiment must be for the greater good of society

…

6. The risks should never exceed the benefits

9. … subject should be at liberty to bring the experiment to an end…

10. … terminate the experiment at any stage, …

US Public Health Service Syphilis Study at Tuskegee

- Goal: observe natural history of untreated syphilis
- 600 poor African American sharecropper men
  - 399 with syphilis, 201 controls
- Were not treated, merely studied
  - Were not told they had syphilis
  - Sexual parternes not informed
  - In 1940s, penicillin became treatment; but they didn't get the treatment

1932
The U.S. Public Health Service (USPHS) engages the Tuskegee Institute in Macon, AL in the USPHS Tuskegee Syphilis Study.[2]

Mid-1940s
Penicillin becomes treatment of choice for syphilis, but men in study are not treated.

1972
First news article about the study.

The study ends, on recommendation of an Ad Hoc Advisory Panel convened by the Assistant Secretary for Health and Scientific Affairs.

1997
President Clinton issues a formal Presidential apology.

https://www.cdc.gov/tuskegee/timeline.htm

US Public Health Service Syphilis Study at Tuskegee

- 1964 Protest letter from a doctor who read the papers
    - “I am utterly astounded by the fact that physicians allow patients with a potentially fatal disease to remain untreated when effective therapy is available”
    - “… need to re-evaluate their moral judgements in this regard”
- 1965 Memo for authors:
    - “This is the first letter of this type we have received. I do not plan to answer this letter”

Dr. Irwin Schatz, the first, lonely voice against infamous Tuskegee study, dies at 83

By Sarah Kaplan

April 20, 2015 at 5:39 a.m. EDT

[The Washington Post]

US Public Health Service Syphilis Study at Tuskegee

- 1966 Peter Buxtun, a PHS researcher in San Francisco, sent a letter to CDC, but the study was not stopped
- 1972, Buxtun goes to the press.
- Senator Edward Kennedy calls congressional hearings
- 1974 Congress passes National Research Act

The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were

have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

NY Times. July 26, 1972

Behavioral Experiments

Stanford Prison Experiment

- Conducted by Philip Zimbardo, Stanford University, August 1971
- Goal: to test how perceived power affects subjects
- College students were randomly assigned to be either “prisoners” or “guards”
- Guards were instructed to do whatever they thought was necessary to maintain law and order in the prison and to command the respect of the prisoners. No physical violence was permitted.
- Results:
  - Guards humiliated and abused prisoners
  - Prisoners became depersonalized
  - Evidence for “ugly side of human nature”

https://www.youtube.com/watch?v=oAX9b7agT9o

"How we went about testing these questions and what we found may astound you. Our planned two-week investigation into the psychology of prison life had to be ended after only six days because of what the situation was doing to the college students who participated. In only a few days, our guards became sadistic and our prisoners became depressed and showed signs of extreme stress. Please read the story of what happened and what it tells us about the nature of human nature."

–Professor Philip G. Zimbardo

https://www.prisonexp.org/

Stanford Prison Experiment: Scientific & Ethical Flaws

- Participants were not random: respondents to an ad for "a psychological study of prison life".
    - Carnahan and McFarland 2007: word "prison" selects personalities
- Guards were told the expected results ("conditions which lead to mob behavior, violence")
- Researchers intervened in experiment to instruct guards how to behave ("we can create a sense of frustration. We can create fear")
- Research refused to allow prisoner participants to leave experiment.

https://www.letexier.org/IMG/pdf/LeTexier_Debunking-the-SPE_American-Psychologist_2019.pdf

Blue vs brown eye “racism”

- Kids separated by color of eyes
  - Blue / Brown eyes
  - Made-up reasons: brown eyes were superior to those with blue eyes
  - Brown eyes got special privileges
- Students started to internalize, and accept the characteristics they'd been arbitrarily assigned based on the color of their eyes

[We Are Repeating The Discrimination Experiment Every Day, Says Educator Jane Elliott]

Blue vs brown eye “racism”

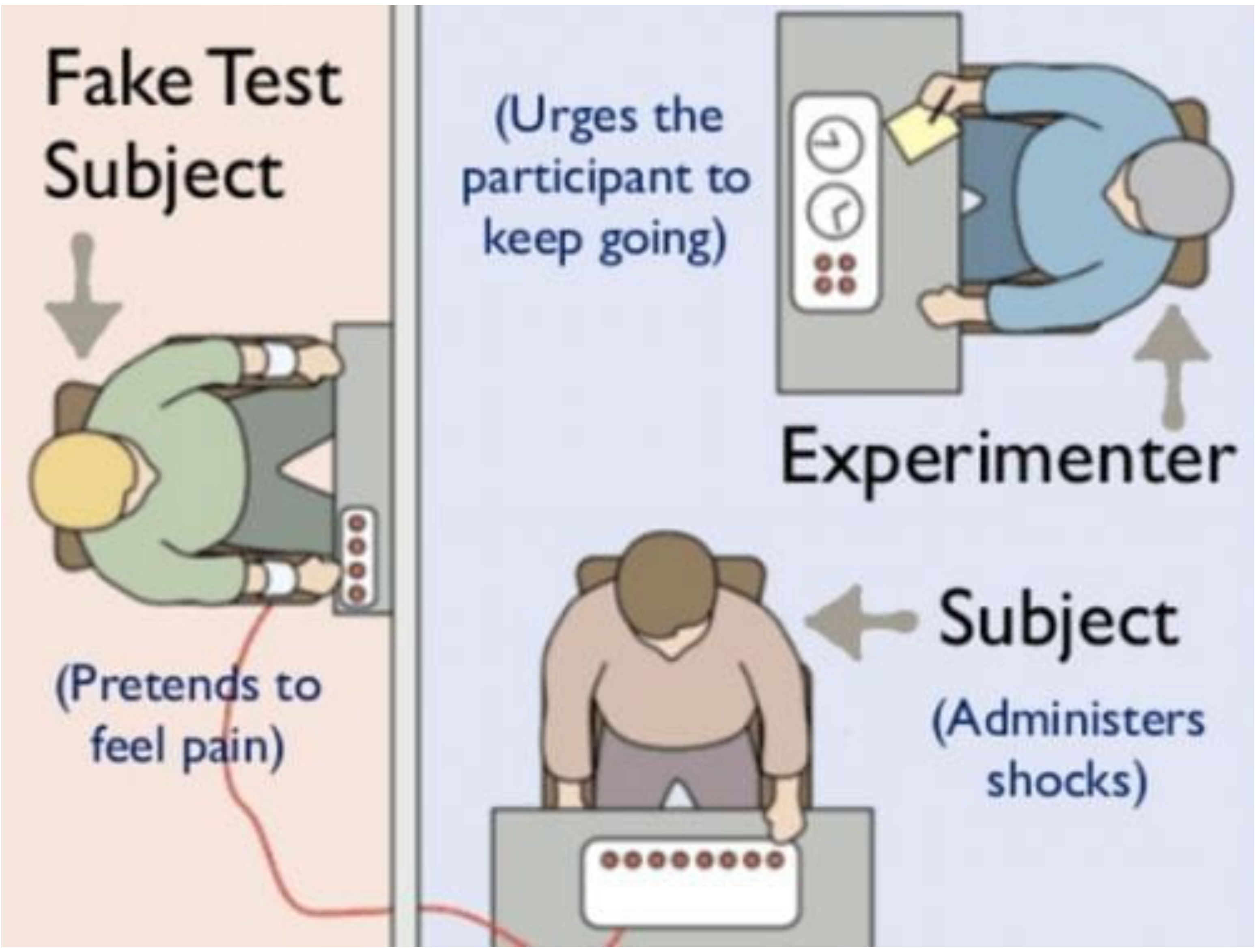
- Kids separated by color of eyes
  - Blue / Brown eyes
  - Made-up reasons: brown eyes were superior to those with blue eyes
  - Brown eyes got special privileges
- Students started to internalize, and accept the characteristics they'd been arbitrarily assigned based on the color of their eyes
- Is this experiment ethical?
- Do we learn something?
- Do the participants learn something?

[We Are Repeating The Discrimination Experiment Every Day, Says Educator Jane Elliott]

Milgram Obedience Experiment

- Stanley Milgram, Yale, 1962
- Focusing on the conflict obedience to authority vs personal conscience
- Three roles: Experimenter, Teacher (actual subject), Learner
- Learner and Experimenter were informed about the experiment
    - Teacher asked to give mild electric shocks to the Learner
    - Learner had to answer questions and got things wrong
    - Experimenter, as the matter of fact, asked Teacher to torture Learner
- Most Teachers obeyed the Experimenter

https://www.simplypsychology.org/milgram.html



65% (two-thirds) of participants (i.e., teachers) continued to the highest level of 450 volts. All the participants continued to 300 volts.

Ordinary people are likely to follow orders given by an authority figure, , even to the extent of killing an innocent human being.

Current Human Participants Rules

National Research Act 1947

- These experiments (especially the Tuskegee experiment) led to the creation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research
  - The Common Rule: Title 45, Part 46 of the Code of Federal Regulations: Protection of Human Subjects.
    - Informed consent
  - Required institutional review of all federally funded experiments
    - Institutional Review Boards (IRBs)
  - Issued Belmont Report in 1976/1979

The Belmont Report: three basic ethical principles

1. Respect for Persons

- Individuals should be treated as autonomous agents
    - "Informed Consent"
- Persons with diminished autonomy are entitled to protection

The Belmont Report: three basic ethical principles

2. Beneficence

- Do no harm
- Maximize possible benefits and minimize possible harms

The Belmont Report: three basic ethical principles

3. Justice

Who ought to receive the benefits of research and bear its burdens?

- Fair procedures and outcomes in the selection of research subjects
- Advances should benefit all

IRB — Institutional Review Board

- Internal to institution
    - Most universities hav medical
    - USC IRB: https://hrp
- Independent of researc

What needs IRB Review

Human Subjects Research

An activity requires IRB review if it fits the federal definition below

for ***research*** and ***human subjects***

or the FDA definitions of *clinical investigation* and *human subjects*

Research means a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge.

Human subject means a living individual about whom an investigator (whether professional or student) conducting research:

- Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
- Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

IRB — Institutional Review Board

- Review all human experimentation
  - Assesses instructions
  - Compensation
  - Contribution of research
  - Value to the participant
  - Protection of privacy and confidentiality

IRB — Institutional Review Board

- Different standards for different institutions
- Board consists of (primarily) non-expert peers
  - Helps educate new researches and makes suggestions to find solutions to ethical issues

IRB — Common Rule

Human Subject

- A living individual about whom an investigator (professional or student) conducting research
  - Obtains information .. through intervention or interaction with the individual, and uses, studies, or analyzes the information …; or
  - Obtains, uses, studies, analyzes, or generate identifiable private information

Ethical questions

Can you lie to/mislead participants?

[Belmont Report]

- “incomplete disclosure” is allowed when:
  - incomplete disclosure is truly necessary to accomplish the goals of the research
  - there are no undisclosed risks to subjects that are more than minimal, and
  - there is an adequate plan for debriefing subjects, when appropriate, and for dissemination of research results to them

Ethical questions

Can you lie to/mislead participants?

Can you deceive participants?

- deception, incomplete disclosure, no more than minimal risk, no alternative
  - Key concept: debriefing

Ethical questions

Can you lie to/mislead participants?

Can you deceive participants?

Can you harm a human subject?

- Definition of harm?

IRB: Exempt Research

- Your human subjects research qualifies for exempt status if
  - Research only includes survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) and one of:
    - The identity of the human subjects cannot readily be ascertained
    - Any disclosure of the human subjects' responses outside the research would not place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation; or
    - The information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, and an IRB conducts a limited IRB review …

https://hrpp.usc.edu/irb/exempt-level-of-review/

IRB: CITI training

If you intend to be on any research projects that run human subjects, you must do CITI certification

- Required by USC IRB
- Required for all federally funded research
- Short course

https://hrpp.usc.edu/education_certification/

What about data from corpora (e.g, social media)?

Using Social Media Data

- Social media data: Twitter, Reddit, YouTube, etc

- From IRB perspective, this kind of corpus data is exempt if it is public
  - e.g., public Twitter data

- But are there still questions?

Possible Issues with Social Media data

- Author

> "Are consent, confidentiality and anonymity required where the research is conducted in a public place where people would reasonably expect to be observed by strangers?"

- What counts as a public vs. private space on/off the web?
    - If people are whispering in a public square is that private?
    - What about religious ceremonies?

Williams, M. L., Burnap, P., 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

Possible Issues with Social Media data

What are the potential harms?

- Demographic info (age, ethnicity, religion, sexual orientation)
- Associations (membership in groups or associations with particular people)
- Communications that are person or potentially harmful (extreme options? illegal activities?)
- Others?

Williams, M. L., Burnap, P., 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

What do Twitter Authors Think?

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. Social Media + Society, 4(1). 22

Table 2. Comfort Around Tweets Being Used in Research.

Question	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
How do you feel about the idea of tweets being used in research? ($n=268$)	3.0%	17.5%	29.1%	35.1%	15.3%
How would you feel if a tweet of yours was used in one of these research studies? ($n=267$)	4.5%	22.5%	23.6%	33.3%	16.1%
How would you feel if your entire Twitter history was used in one of these research studies? ($n=268$)	21.3%	27.2%	18.3%	21.6%	11.6%

Note. The shading was used to provide a visual cue about higher percentages.

USC

Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

What do Twitter Researchers do/think?

Vitak, Jessica, Katie Shilton, and Zahra Ashktorab. 2016. "Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community." ACM CSCW, pp. 941-953. 2016.

Code	Definition	Example Statements
Public Data	Only using public data / public data being okay to collect and analyze	*In general, I feel that what is posted online is a matter of the public record, though every case needs to be looked at individually in order to evaluate the ethical risks.*
Do No Harm	Comments related to the Golden Rule	*Golden rule, do to others what you would have them do to you.*
Informed Consent	Always get informed consent / stressing importance of informed consent	*I think at this point for any new study I started using online data, I would try to get informed consent when collecting identifiable information (e.g. usernames).*
Greater Good	Data collection should have a social benefit	*The work I do should address larger social challenges, and not just offer incremental improvements for companies to deploy.*
Established Guidelines	Including Belmont Report, IRBs Terms of Service, legal frameworks, community norms	*I generally follow the ethical guidelines for human subjects research as reflected in the Belmont Report and codified in 45.CFR.46 when collecting online data.*
Risks vs. Benefits	Discussion of weighing potential harms and benefits or gains	*I think I focus on potential harm, and all the ethical procedures I put in place work towards minimizing potential harm.*
Protect Participants	Methods to protect individual: data aggregation, deleting PII, anonymizing/obfuscating data	*I aggregate unique cases into larger categories rather than removing them from the data set.*
Deception	Justifying its (non) use in research	*I use deception for participatory research and debrief at the end.*
Data Judgments	Efforts to not make inferences or judge participants or data	*Do not expose users to the outside world by inferring features that they have not personally disclosed.*
Transparency	Contact with participants or methods of informing participants about research	*I generally choose not to scrape/crawl public sources. I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.*
In Flux	One's code of ethics is under development, context-dependent, or otherwise in flux	*It very much depends on the nature of the data.*

Table 2. Emergent themes from qualitative responses regarding researchers' personal code of ethics

Some Proposals

- OK to programmatically collect data without explicit consent

- But seek informed consent for all directly quoted content in publications
    - Twitter’s view is that users retain rights to the content they post.

Williams, M. L., Burnap, P., 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users’ Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

More suggestions

- Transparency with research communities
  - Ask / inform
  - Ethical deliberation with colleagues (in addition to IRBs)
  - Be cautious about sharing results that include potentially identifiable outliers

Vitak, Jessica, Katie Shilton, and Zahra Ashktorab. 2016. "Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community." ACM CSCW, pp. 941-953. 2016.

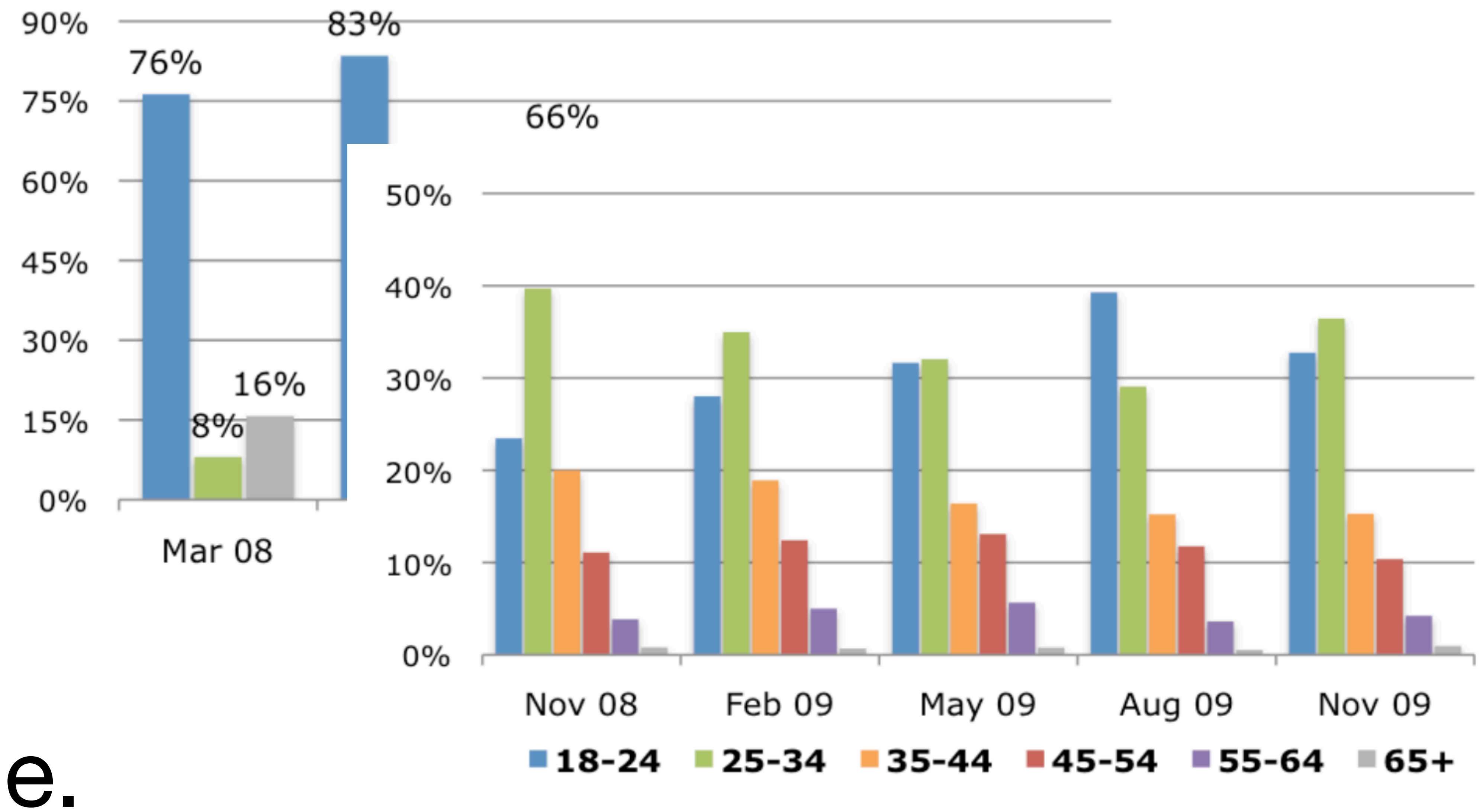
Human subjects selection: crowdwork and crowdworkers

Questions

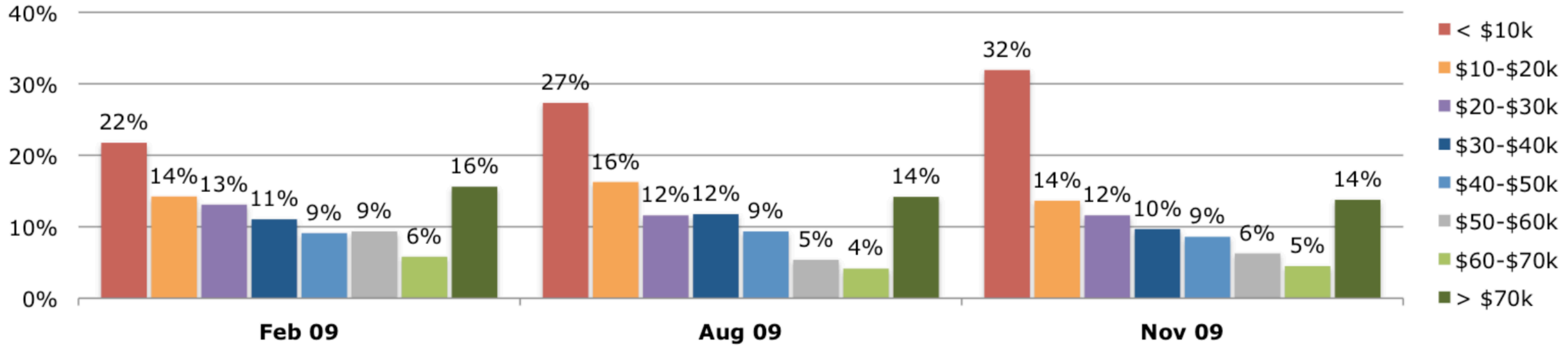
- Have you ever employed crowd workers?
  - How much do you pay them?
  - How do you know that?
- Have you ever done a crowdsourced task by yourself?
- Would you ever do crowdsourced tasks for a prolonged period?
- Would you recommend being a crowd worker to a friend or family member?

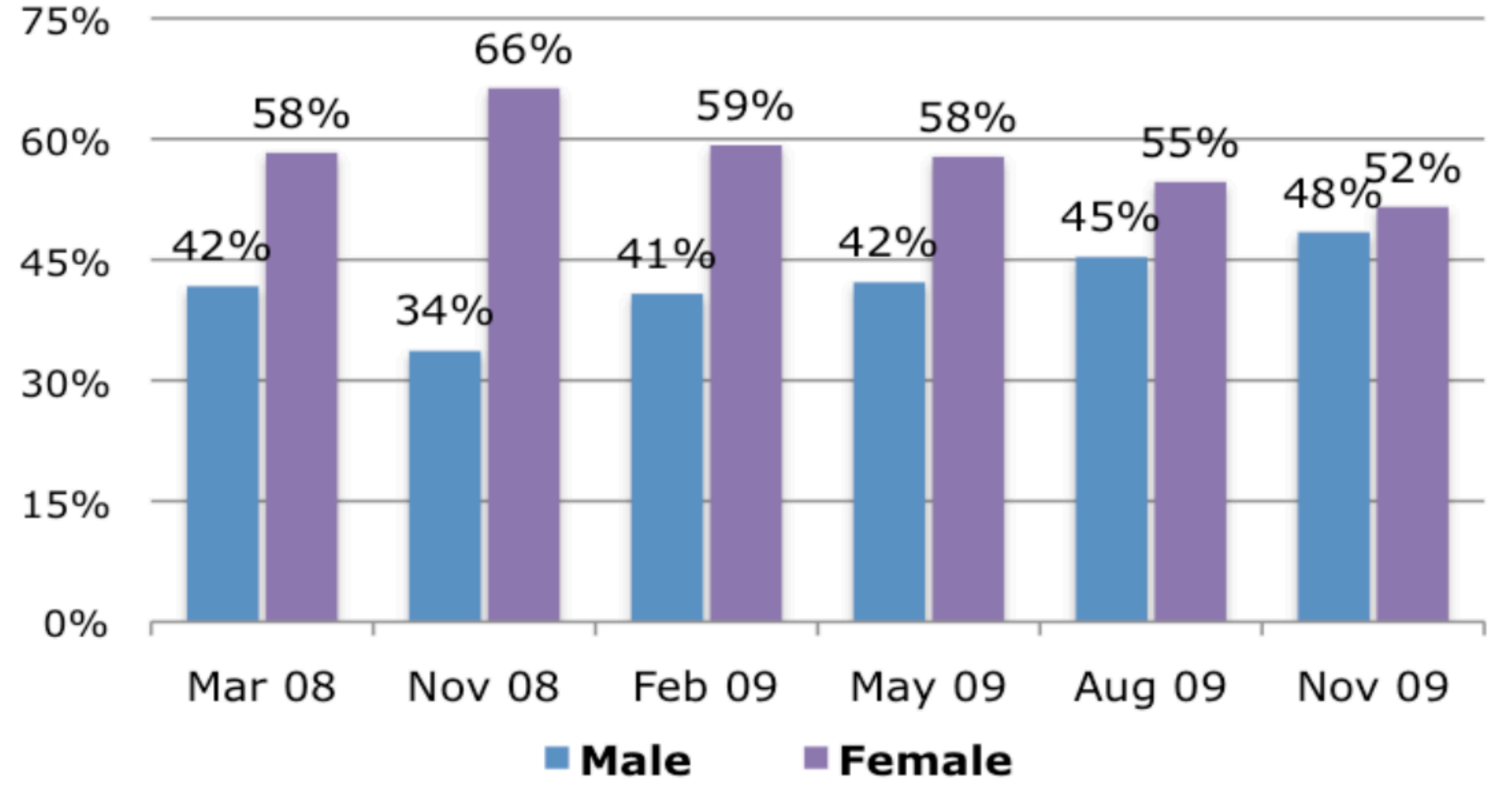
Who are crowdworkers?

- Largely from the US and India



- Skewed young, female and lower income.





Joel Ross, et al. Who are the Crowdworkers? Shifting Demographics in Mechanical Turk. CHI 2010

Crowdworker Incomes

- Multiple studies have found actual wages at or below $2 / hr

> Our task-level analysis revealed that workers earned a median hourly wage of only ~$2/h, and only 4% earned more than $7.25/h. While the average requester pays more than $11/h, lower-paying requesters post much more work.[1]

[1] Hara et al. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. CHI 2018

Harms to Crowdworkers

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet. …

The data labelers employed by Sama on behalf of OpenAI were paid a take-home wage of between around $1.32 and $2 per hour depending on seniority and performance. …

The story of the workers who made ChatGPT possible offers a glimpse into the conditions in this little-known part of the AI industry, which nevertheless plays an essential role in the effort to make AI systems safe for public consumption. …

These invisible workers remain on the margins even as their work contributes to billion-dollar industries.

.. violence, hate speech, and sexual abuse in the data… The work's traumatic nature eventually led Sama to cancel all its work for OpenAI in February 2022…

https://time.com/6247678/openai-chatgpt-kenya-workers/

What about the data obtained?

"Datasets are the telescopes of our field."—Aravind Joshi

- Data annotation is an essential part of every NLP project.
- Annotation: Looking at language data and adding additional information about it
- How is it used?
    - To provide training data for your system
    - To evaluate how well your system is working.

Ethical Issues re: data

- NLP systems (and machine learning models) can amplify unwanted social biases reflected in the training data
- Data issues can cause NLP systems to fail for some populations (children, the elderly, speakers of dialects, minority languages)
- Data has scientific implications
    - What is the training/test split?
    - Is the data appropriate for the task?
    - How was the data labeled?

Lets' do some annotations!

Instruction / Goal: Label a dataset that helps me use people's facial expressions in photos to predict whether they are happier after they graduate from collage.

" I will collect profile pictures of people who are in college, and who are working"

Bad idea! Profile pictures is not natural reflection of people's state, and have biases, e.g., people are better at smiling at the camera when they start to work. But no pictures would make sense anyways, because facial expressions in photos are intentional, and it wouldn't faithfully record people's emotion.

Lets' do some annotations!

Instruction / Goal: Label a dataset that helps me use people's facial expressions in photos to predict whether they are happier after they graduate from collage.

Might be hard to imagine but people did something like this, "identify criminals based on their facial characteristics"

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Takeaway: Your data source may already be the issue; Or sometimes the task itself is the issue and labeling just shouldn't exist.

[Physiognomy's New Clothes]

So…

Annotation is hard, and we need to carefully collect data points in order to get good data points.

Datasheets, data statements, etc

Dataset creators:

- Encourage careful reflection on assumptions, risks, and implications

Dataset consumers:

- Support informed decisions about using a dataset

Data Sheets

- Motivation
    - Why collected, who, how funded
- Composition
    - How many instances, how sampled, data split
- Collection process
    - How collected, how metadata assigned, IRB, timeline, consent

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. 2020. [Datasheets for Datasets](). Arxiv.

Data Statements

A "design solution and professional practice" for NLP

Should be included in NLP writings:

- papers presenting new datasets
- papers reporting experimental work with datasets
- system documentation

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. TACL 6, 587–604.

Data Statement Sample

Hate Speech Twitter (Waseem and Hovy 2016) https:// github.com/zeerakw/hatespeech

Curation Rationale

In order to study the automatic detection of hate speech.

Language Variety

Twitter search API in late 2015. Information about which varieties of EN are represented is not available, but at least Australian (en-AU) and US (en-US) mainstream English are both included.

Speaker Demographic

Speakers were not directly approached for inclusion in this dataset and thus could not be asked for demographic information. More than 1,500 different Twitter accounts are included.

Data Statement Sample

Hate Speech Twitter (Waseem and Hovy 2016) https:// github.com/zeerakw/hatespeech

Annotator Demographic

This dataset includes annotations from both crowd workers and experts. A total of 1,065 crowd workers were recruited through Crowd Flower, primarily from Europe, …

Speech Situation

All tweets were initially published between April 2013 and December 2015. Tweets represent informal, …

Text Characteristics

For racist tweets the topic was dominated by Islam and Islamophobia.

Bender's Question

What language is this paper studying?

"Surveys of EACL 2009 (Bender 2011) and ACL 2015 (Munro, 2015) found 33-81% of papers failed to name the language studied. (It always appeared to be English.) "

– Bender and Friedman 2018.

Potential Harms

A Harm Framework Heuristics

To help practitioners determine the specific harm(s) a bias measure evaluates, we propose the following set of heuristics.

Stereotyping: Does the method:

- deal with language which communicates an existing, well-known a priori judgement or generalization which oversimplifies the reality of diversity within the group?
- measure predictions or probabilities of associations between specific groups and certain characteristics, concepts, language, or sentiments?
- focus on finding specific, pre-defined outcomes based on hypotheses about stereotypical associations, i.e., is the hypothesis directional?
- test associations which either the "average" in-group member or person in the relevant society would be able to quickly predict, i.e., would they be able to predict or identify what the ’problem’ is and connect its roots to their cultural/historical knowledge?

Note: these associations can be positive or negative, but should not hold as naturalistic when a commensurable group is swapped in.

a ‘control’ group (whether or not explicitly stated as such)?

Disparagement: Does the method:

- deal with generally belittling, devaluing, or de-legitimizing language about a group?
- engage with sentiments related to societal regard (respect), expressing normative judgments, or using scalar adjectives pertaining to quality or worth (best/worst, good/bad), but which are not tied to an established stereotype?
- use language which holds as pragmatically and semantically valid/naturalistic when the group identifier is perturbed with a commensurable group?
- deal with ‘toxicity’ or ‘unhealthy’ discourse in general?

Dehumanization: Does the method specifically mention language commonly used to dehumanize, such as:

- associations with non-human life (vermin, insects)?
- implications that a certain group is sub-human or not ‘true’ members of a superset (certain ‘immigrants’ aren’t ‘American’)?
- notions related to eugenics?

S. Dev, E. Sheng, J. Zhao, A. Amstuta, J. Sun, Y. Hou, M. Sanseverino, J. Kim, A. Nishi, N. Peng, K.W. Chang. On Measures of Biases and Harms in NLP. AACL 2022

What about labeling?

- Did the paper use labels from an external dataset or were some data relabeled?
- Who were they? Experts? Crowdworkers?
- How were they trained?
  - Are training examples given in the paper?
- How screened?
- How were they compensated?
- How aggregated to form the final labels?

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? ACM FAT* 2020

Other Consideration

Know your end goal before you start collecting and annotating data points.

"We use the datasets to facilitate further progress toward a primarily scientific goal: building machines that can demonstrate a comprehensive and reliable understanding of everyday natural language text in the context of some specific well-posed task, language variety, and topic domain."

Bowman, Samuel R., and George E. Dahl. "What will it take to fix benchmarking in natural language understanding?." NAACL 2020