Natural Language Processing

Computational Ethics

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Announcements

- Last quiz on Monday Dec 5
  - dependency parsing, recommender systems, summarization, ethics
- Last homework deadline on Friday Dec 9
- Last lecture Monday Dec 5

- Email me if you would like to TA this course in the next Fall quarter (2023)

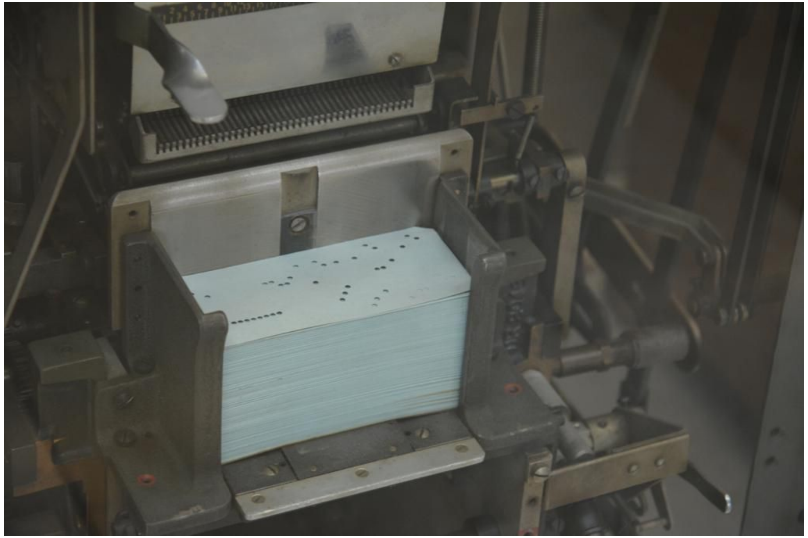
- Thank you for the fun course and for your participation and hard work ❤️

Week	Date	Topics	Readings	Homeworks
1	9/28	Logistics [slides] [recording]	Course website, syllabus	
	9/30	Introduction [slides] [recording]	Eis 1	HW1 out
2	10/03	Introduction [slides]	Eis 1	
	10/05	Text classification [slides] [recording]	Eis 2; J&M III 4	
	10/07	Text classification [slides] [recording]	Eis 2; J&M III 4; Ng & Jordan, 2001	
3	10/10	Text classification [slides] [recording]	Eis 2; J&M III 5; Pang et al. 2002	
	10/12	Text classification [slides] [recording]	J&M III 5	In-class quiz 1
	10/14	Text classification [slides] [recording]	J&M III 5	
4	10/17	Language modeling [slides] [recording]	J&M III 3; Eis 6.1-6.2, 6.4	
	10/19	Language modeling [slides] [recording]	J&M III 3; Eis 6.1-6.2, 6.4	
	10/21	Lexical semantics [slides] [recording]	J&M III 6; Eis 14	HW1 due
5	10/24	Lexical semantics [slides] [recording]	Eis 14; J&M III 6	
	10/26	Lexical semantics [slides] [recording]	Eis 14; J&M III 6	In-class quiz 3
	10/28	Neural networks [slides] [recording]	Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011	
	10/31	Neural networks [slides] [recording] [supplementary recording]	Annotated Transformer; Illustrated Transformer	HW2 Out

6	11/02	Sequence labeling [slides] [recording]	Eis 7.1-7.4, 8.1; J&M III 8	In-class quiz 4
	11/04	Sequence labeling [slides] [recording]	Eis 7.1-7.4, 8.1; Collins notes	
7	11/07	Sequence labeling [slides] [recording]	Eis 7.5, 7.7, 8.3; Sutton & McCallum 2.1 - 2.5	
	11/09	Sequence labeling [slides] [recording]	Eis 7.6	In-class quiz 5
	11/11	Veterans Day (no class)		
8	11/14	Neural sequence labeling [slides] [recording]	Eis 7.6	
	11/16	Parsing [slides] [recording]	Eis 10.1-10.2; J&M III 13	In-class quiz 6
	11/18	Parsing [slides] [recording]	Eis 11.1, 11.3; J&M III 14	
9	11/21	Cancelled	Chen and Manning 2014	HW2 due, In-class quiz 7
	11/23	Parsing [slides] [recording]	Kassas et al. 2021	
	11/25	Thanksgiving (no class)		
10	11/28	Advanced topics: Recommender systems and online training		HW3 out
	11/30	Research topics: Summarization [slides]		
	12/02	Advanced topics: NLP + Historical linguistics		
11	12/05	Q&A		In-class quiz 8
	12/07	TBD		
	12/09	TBD		HW3 due

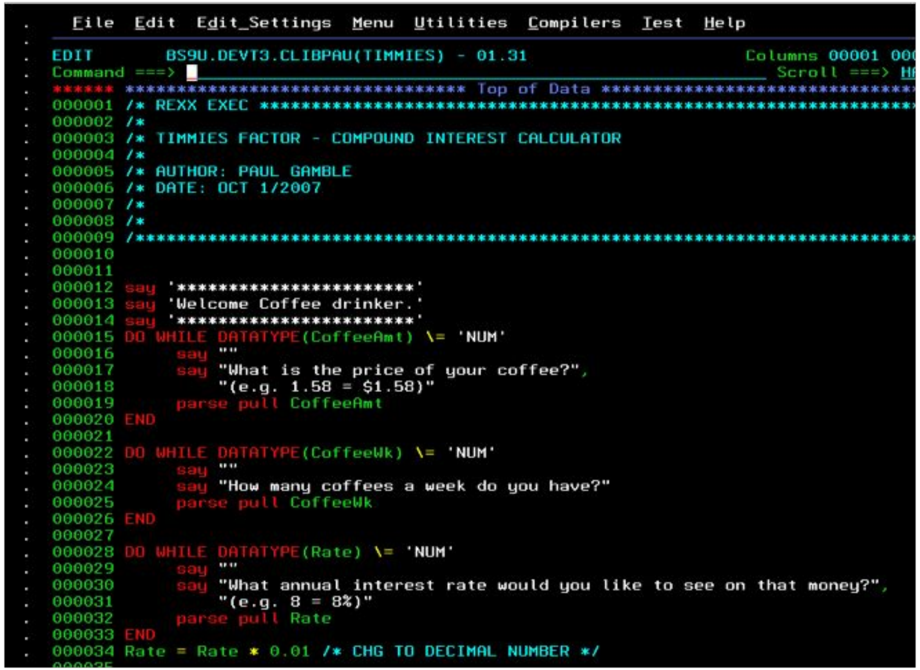
Communication with machines

- 50s-70s



Communication with machines

- 80s



NLP: Communication with machines

- Today

Language use is fundamentally a social activity

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

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

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

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

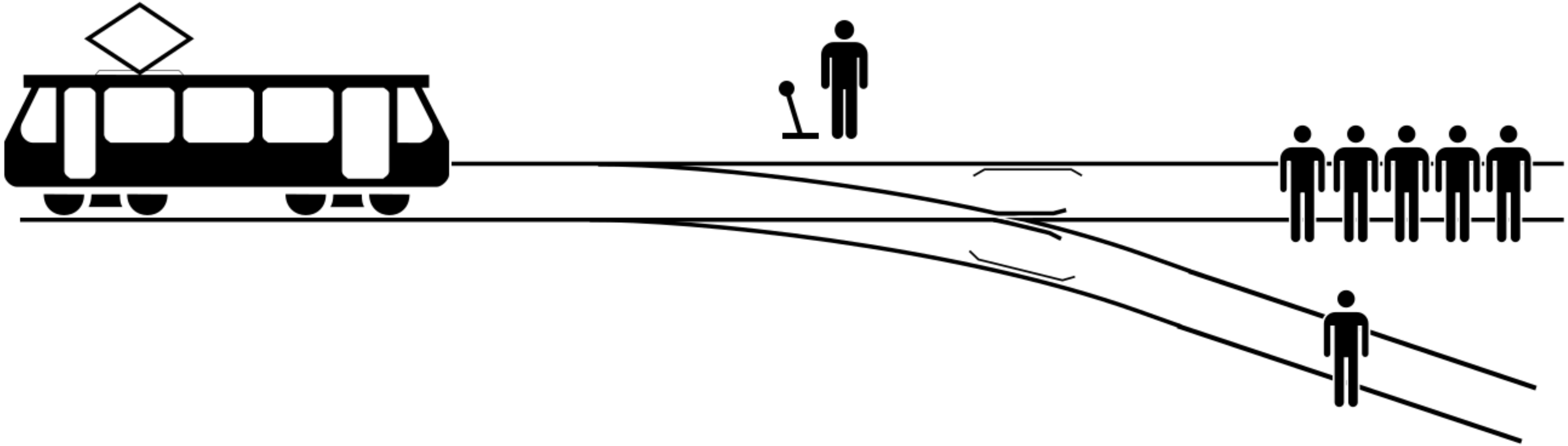
Ethics

It’s the **good** things

It’s the **right** things

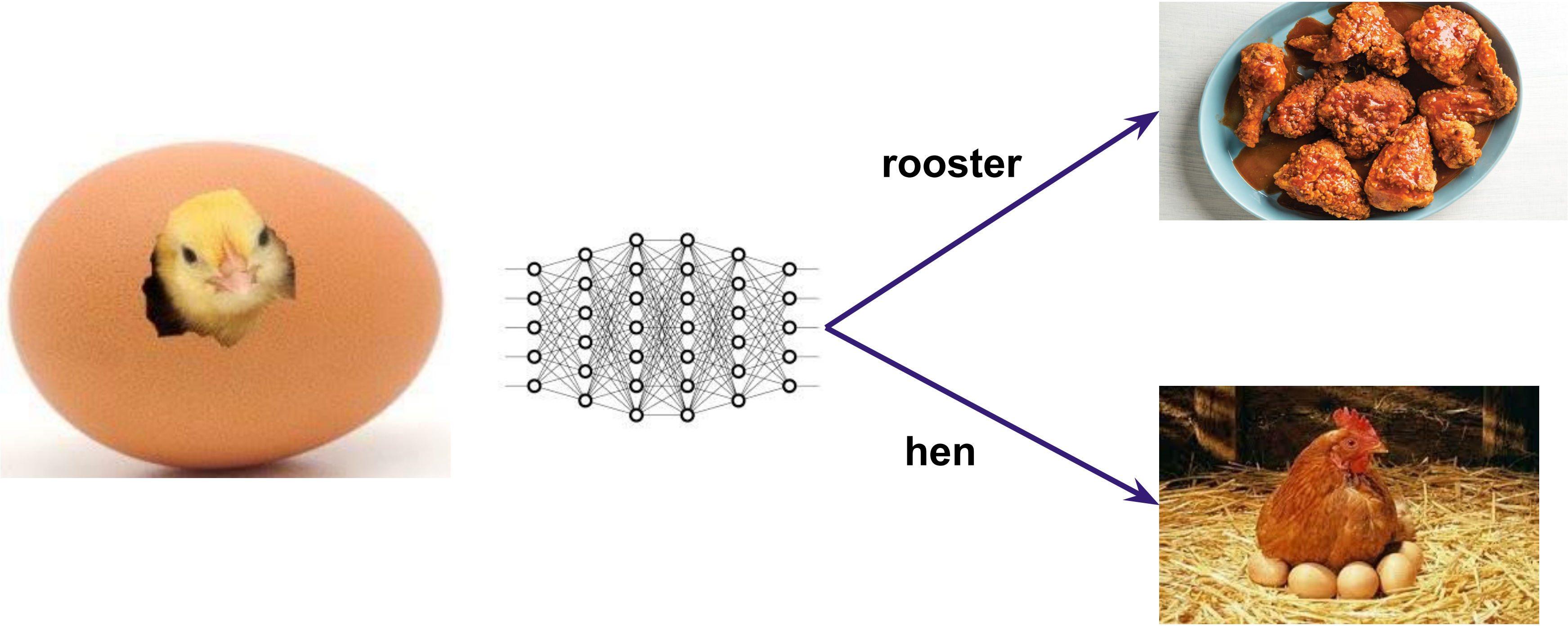
The Trolley Dilemma

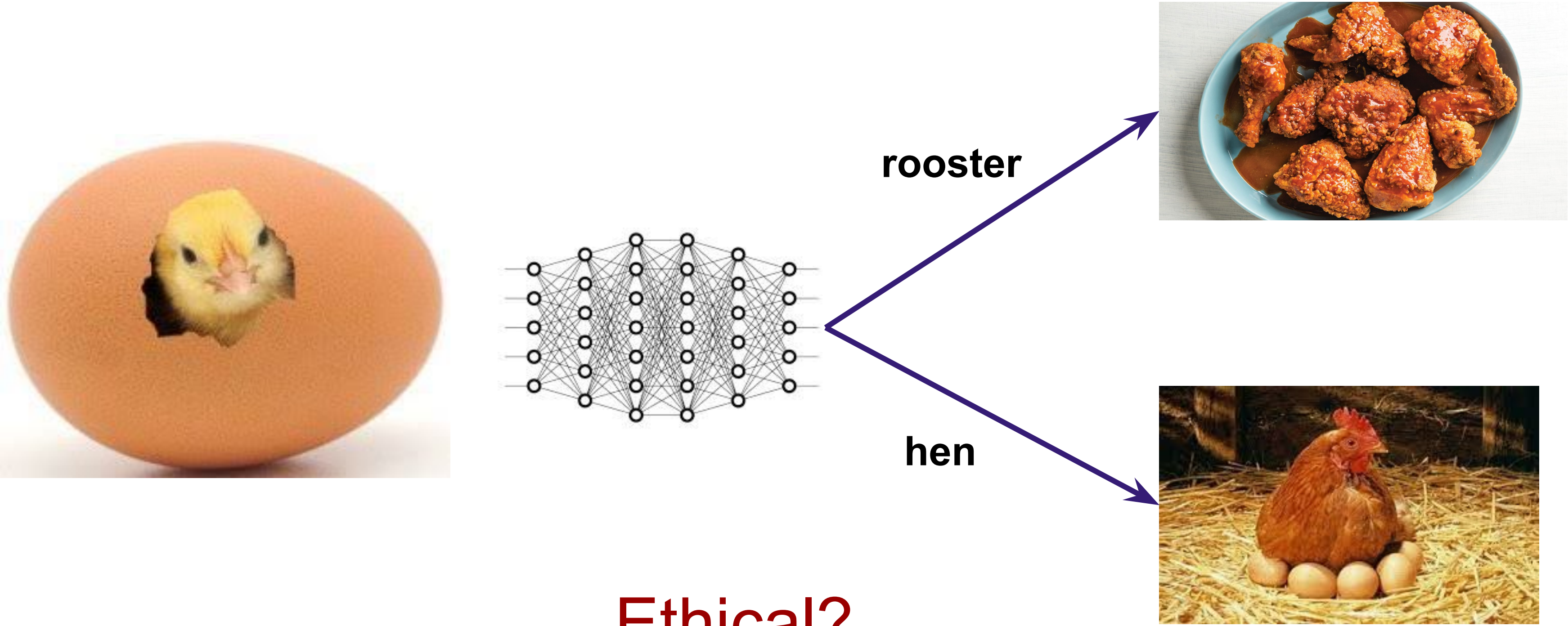
Should you pull the lever to divert the trolley?



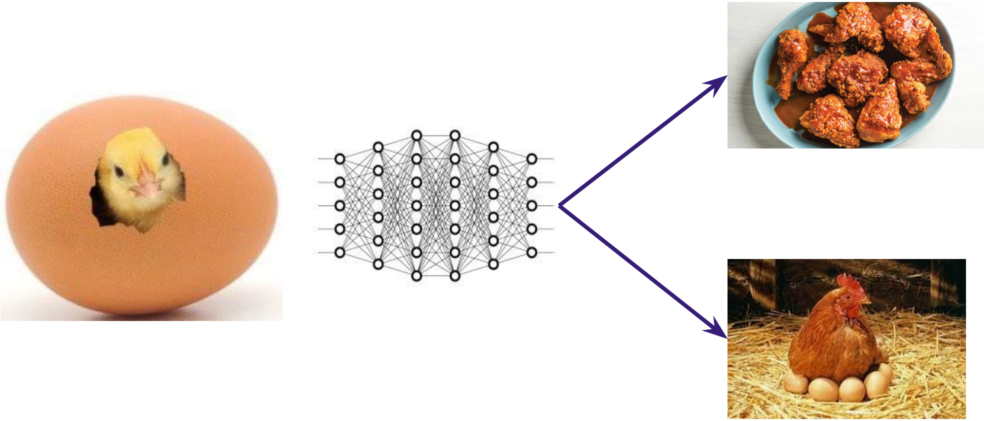
[image from Wikipedia]

The Chicken dilemma



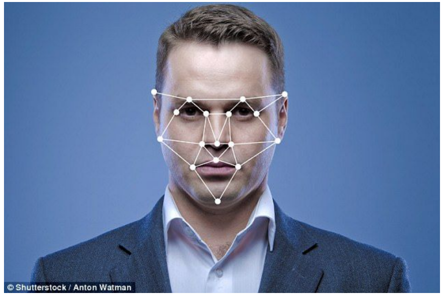


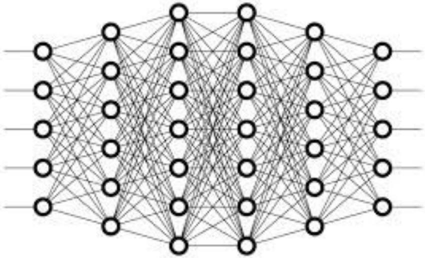
Ethical?

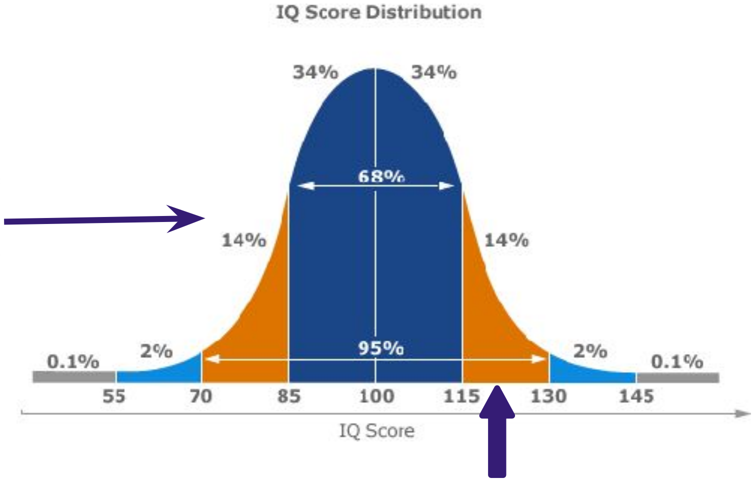


- Ethics is inner guiding, moral principles, and values of people and society
- There are gray areas. We often don’t have easy answers.
- Ethics changes over time with values and beliefs of people
- Legal ≠ Ethical

The IQ dilemma







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

The IQ dilemma

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

- Who could benefit from such a classifier?

The IQ dilemma: the ethics of the research question

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

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

The IQ dilemma: understanding the risks

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

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

The IQ dilemma: understanding the responsibility

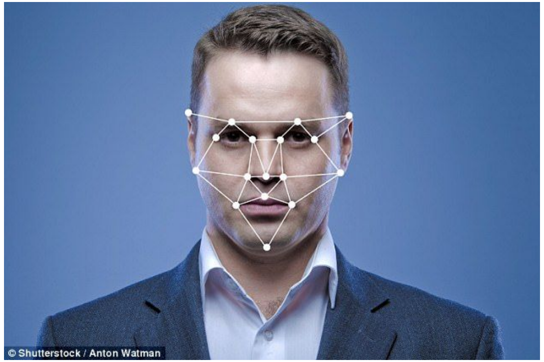
We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

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

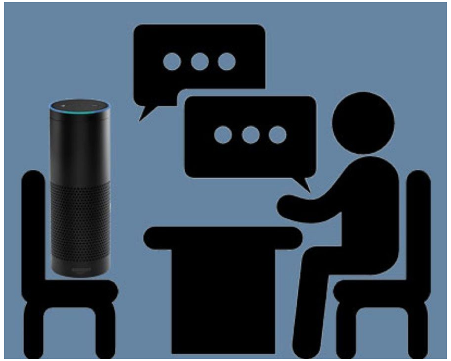
We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

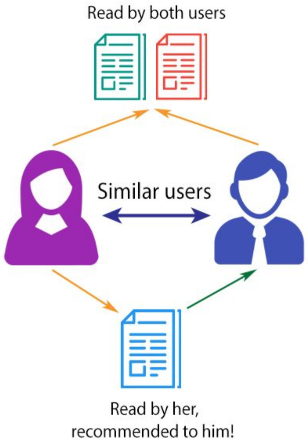
IQ classifier - risks

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

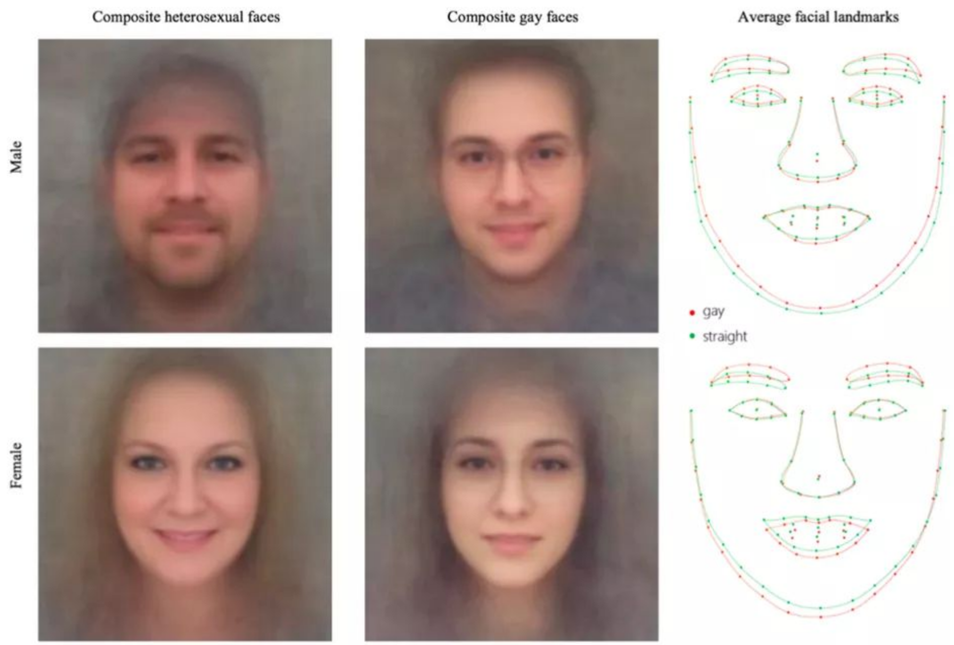

© Shutterstock / Anton Watman

AI and people





A recent study: the “AI Gaydar”, 2017



A recent study: the “AI Gaydar”

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

Let's discuss...

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

What went wrong?

Questioning the ethics of the research question

- Identification of sexual orientation from facial features

Sexual orientation classifier - who can be harmed?

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

Dual framing in predictive analytics

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

Data

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

Data privacy

- Photos downloaded from a popular American dating website

Data privacy

- Photos downloaded from a popular American dating website

Questions to ask:

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

Data

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

Data biases

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

Questions to ask:

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

⟶ this dataset contains many types of biases

Method

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

Algorithmic biases

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

Questions to ask:

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

Evaluation

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

The cost of misclassification

The cost of misclassification

Learn to assess 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.

Beyond decision-support tools and human-centered analytics

Gender/race bias in NLP

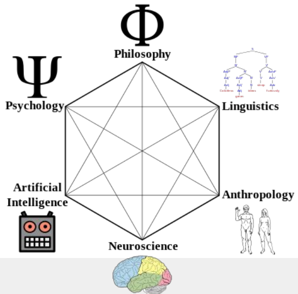
- Machine translation (Douglas'17, Prates et al. '19)
- Caption generation (Burns et al.'18)
- Speech recognition (Tatman'17)
- Question answering (Burghardt et al.'18)
- Dialogue systems (Dinan et al.'19)
- Sentiment Analysis (Kiritchenko & Mohammad'18)
- Language Identification (Blodgett et al.'16, 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)

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

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



Recommended 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

https://tinyurl.com/Readings-CompEthicsInNLP-2022

Spring 2023 – Ethics in AI – CSE 582