# Algorithms for NLP

## Computational Ethics

Yulia Tsvetkov – CMU

# What is Ethics?

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

It is therefore, above all, a **practical discipline**.

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

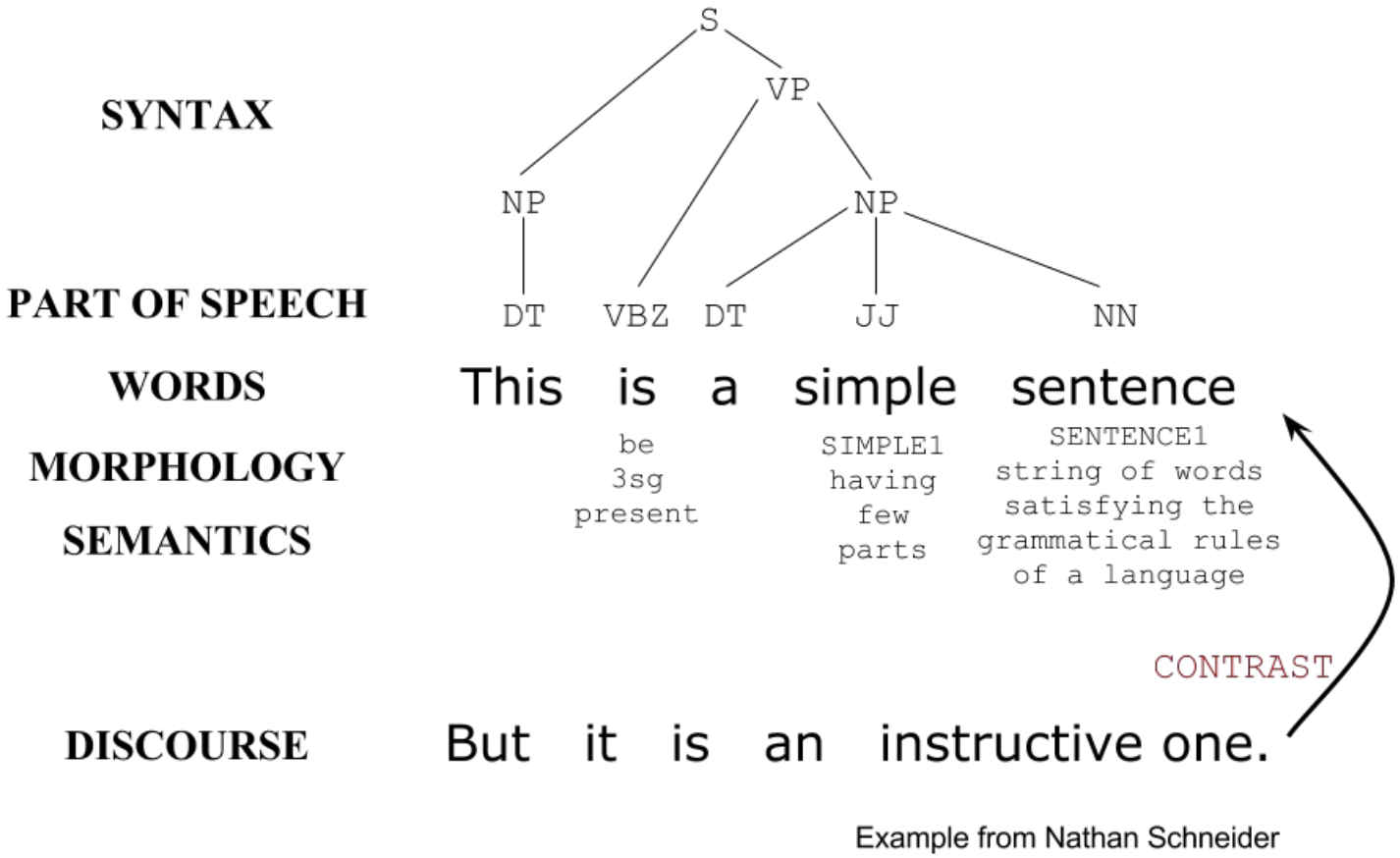
-- Introduction to Ethics, John Deigh

# What is Ethics?

It’s the **good** things

It’s the **right** things

# What NLP Has To Do With Ethics?




Example from Nathan Schneider

# Language and People

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

Dan Jurafsky's keynote talks at CVPR'17 and EMNLP'17

demo.clab.cs.cmu.edu/ethical_nlp/

# Computational Ethics for NLP

## CMU CS 11830, Spring 2019

- **Philosophical foundations:** what is ethics, history, medical and psychological experiments, IRB and human subjects, ethical decision making.
- **Misrepresentation and bias:** algorithms to identify biases in models and data and adversarial approaches to debiasing.
- **Privacy:** algorithms for demographic inference, personality profiling, and anonymization of demographic and personal traits.
- **Civility in communication:** techniques to monitor trolling, hate speech, abusive language, cyberbullying, toxic comments.
- **Democracy and the language of manipulation:** approaches to identify propaganda and manipulation in news, to identify fake news, political framing.
- **NLP for Social Good:** Low-resource NLP, applications for disaster response and monitoring diseases, medical applications, psychological counseling, interfaces for accessibility.

# What is Ethics?

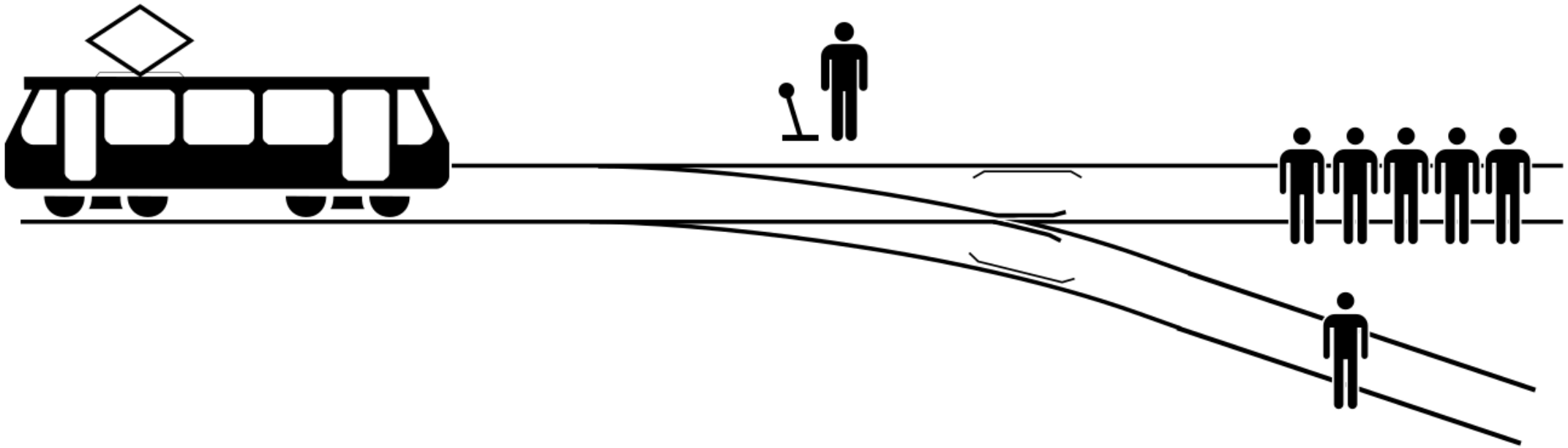
It's the **good** things

It's the **right** things

How simple is it to define
what's good and what's right?

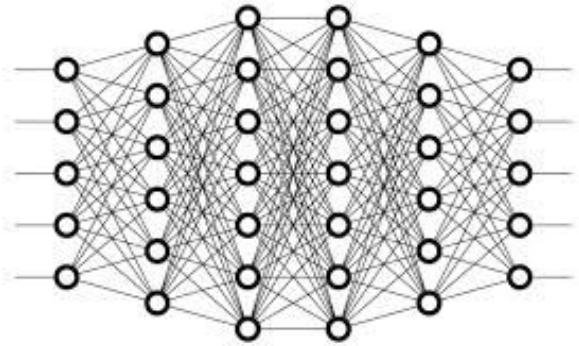
# The Trolley Dilemma

Should you pull the lever to divert the trolley?



[From Wikipedia]

# The Chicken Classifier

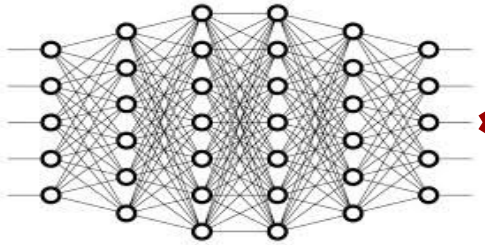


rooster

hen

Ethical?

# The Chicken Dilemma



rooster

hen

- ➔ Ethics is inner guiding, moral principles, and values of people and society
- ➔ It's the **good** things
- ➔ It's the **right** things
- ➔ But is there some absolute definition of right?

# Ethics ≠ Law

- Illegal+immoral:
- legal+immoral:
- illegal+moral:
- legal+moral:

# Ethics ≠ Law

- Illegal+immoral: murder
- legal+immoral:
- illegal+moral:
- legal+moral:

# Ethics ≠ Law

- Illegal+immoral: murder
- legal+immoral: hurting an animal
- illegal+moral:
- legal+moral:

# Ethics ≠ Law

- Illegal+immoral: murder
- legal+immoral: hurting an animal
- illegal+moral: civil disobedience
- legal+moral: eating ice cream

# Ethics ≠ Law

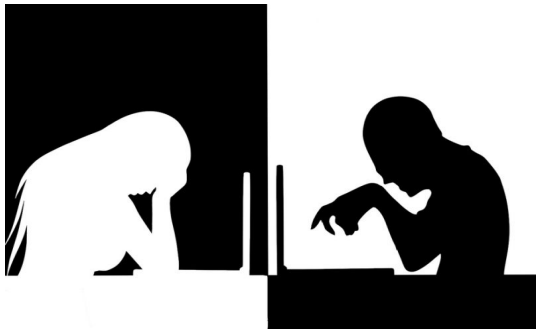
- Illegal+immoral: murder
  - capital punishment
- legal+immoral: hurting an animal
  - eating a burger
- illegal+moral: civil disobedience
  - assassination of a dictator
- legal+moral: eating an ice cream
  - eating the last ice cream in the freezer

# Ethics in Law

- Laws start off to be codified ethics for society
  - But language is used to encode laws and language is never precise
  - Language changes over time: (it says “man” but meant “person”)
  - What’s ethical and what’s not is even encoded in language (“murder” vs “homicide/killing”)
- Adversarial Lawyer looks for loopholes
  - Both sides try to change the interpretation of the law to their advantage

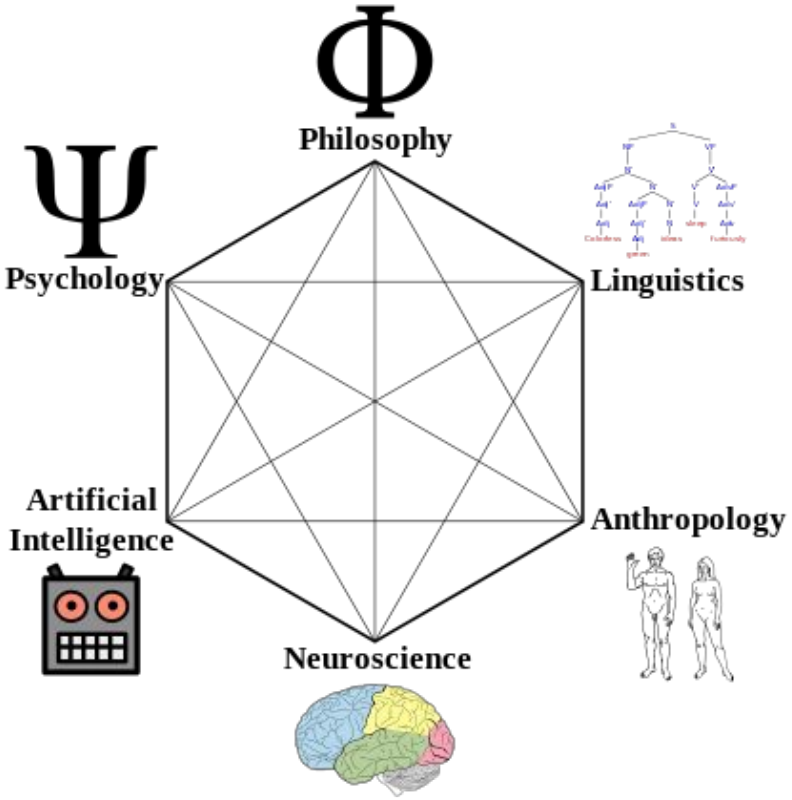
# Ethical Considerations are Time-Dependent

# We Cannot Foresee All Possible Uses of Technology

# Both Ethics and NLP are Interdisciplinary Fields

- Philosophy
- Sociology
- Psychology
- Linguistics
- Sociolinguistics
- Social psychology
- Computational Social Science
- Machine Learning



Ethics is hard even to define, especially over time,
it is highly subjective and personal:
should we then be focusing on it in NLP?

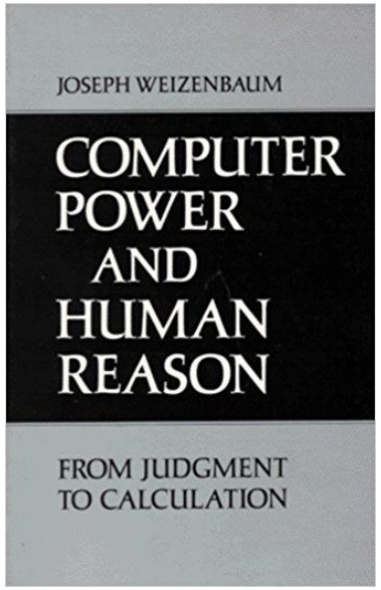
# Ethics is Vaguely Defined, Like Many Other NLP Tasks

- **Ethics is hard even to define**: should we be then focusing on it?
  - It is another problem with an ill-defined answer
    - It still has some definition of good and bad
    - Not everyone agrees on all examples
    - But they do agree on some examples
    - They do have some correlation between people
  - Is this different from other Language Technology Problems
    - Summarization, QA, Dialog, Speech Synthesis

# History: Eliza

- 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” (1964) that raised ethical questions

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



# Still Relevant: “Smart” Toys

## Barbara Grosz, NYT 2015: Barbie Wants to Get to Know Your Child

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

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

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

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

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

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

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

“That’s it, Barbie,” Tiara said.

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

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

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

# Watch This Talk

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

"Yeah," Tiara said. "I only have one."

"What's something nice that your sister does for you?" Barbie asked.

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

"She helped me with my project – and then she *destroyed* it."

"Oh, yeah, tell me more!" Barbie said, oblivious to Tiara's unhappiness.

"That's it, Barbie," Tiara said.

"Have you told your sister lately how cool she is?"

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

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

## Intelligent Systems: Design & Ethical Challenges

https://goo.gl/8tBho8

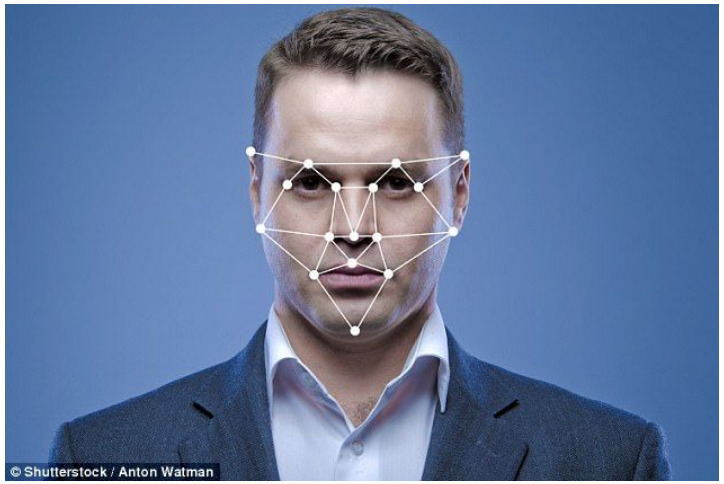
Carnegie Mellon University
Language Technologies Institute

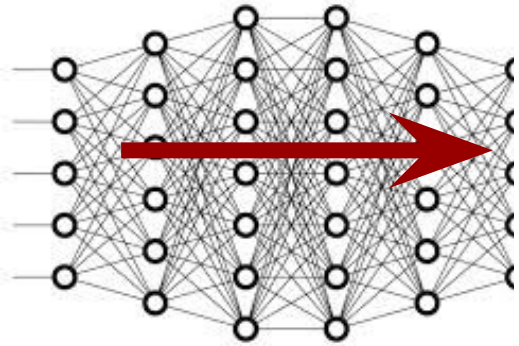
# Our Goals in This Lecture

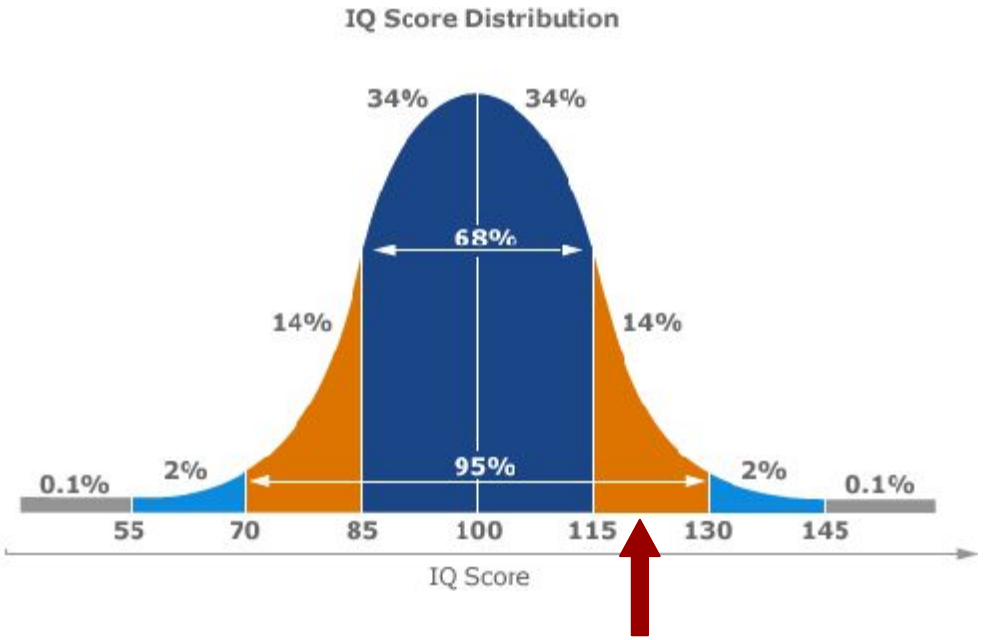
Identify a range of problems where ethical issues emerge, particularly focusing on **Technologies** that interact with **People**

Identify a range of questions that we should be asking ourselves when working with these problems

# Let's Train an IQ Classifier






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

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Assume the classifier is 100% accurate. Who can be harmed from such a classifier?

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Suppose, our test results show 90% accuracy

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

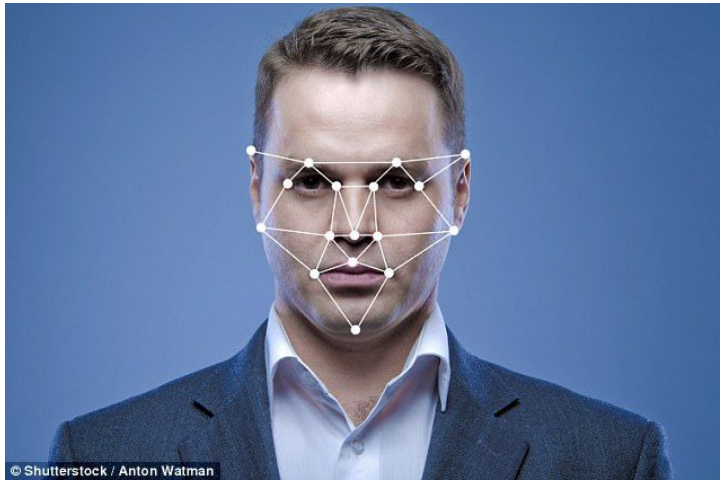
- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Suppose, our test results show 90% accuracy
  - Evaluation reveals that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy

# An IQ Classifier

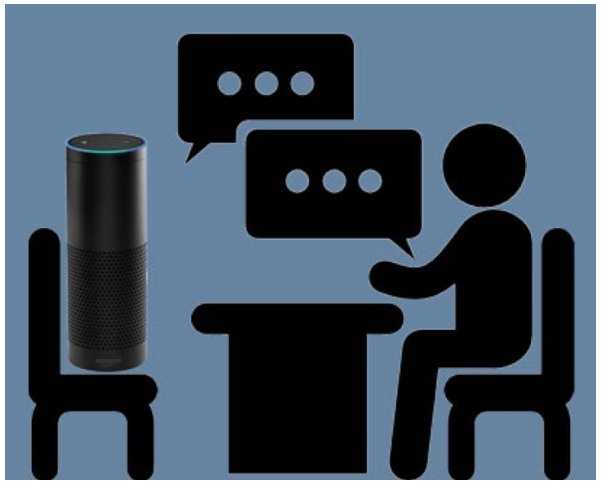
Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Suppose, our test results show 90% accuracy
  - Evaluation reveals that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy
- Who is responsible?
  - Researcher/developer? Reviewer? University? Society?

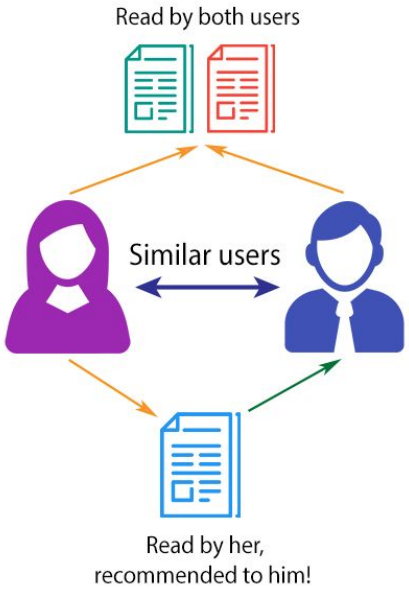
# What's the Difference?



# AI and People

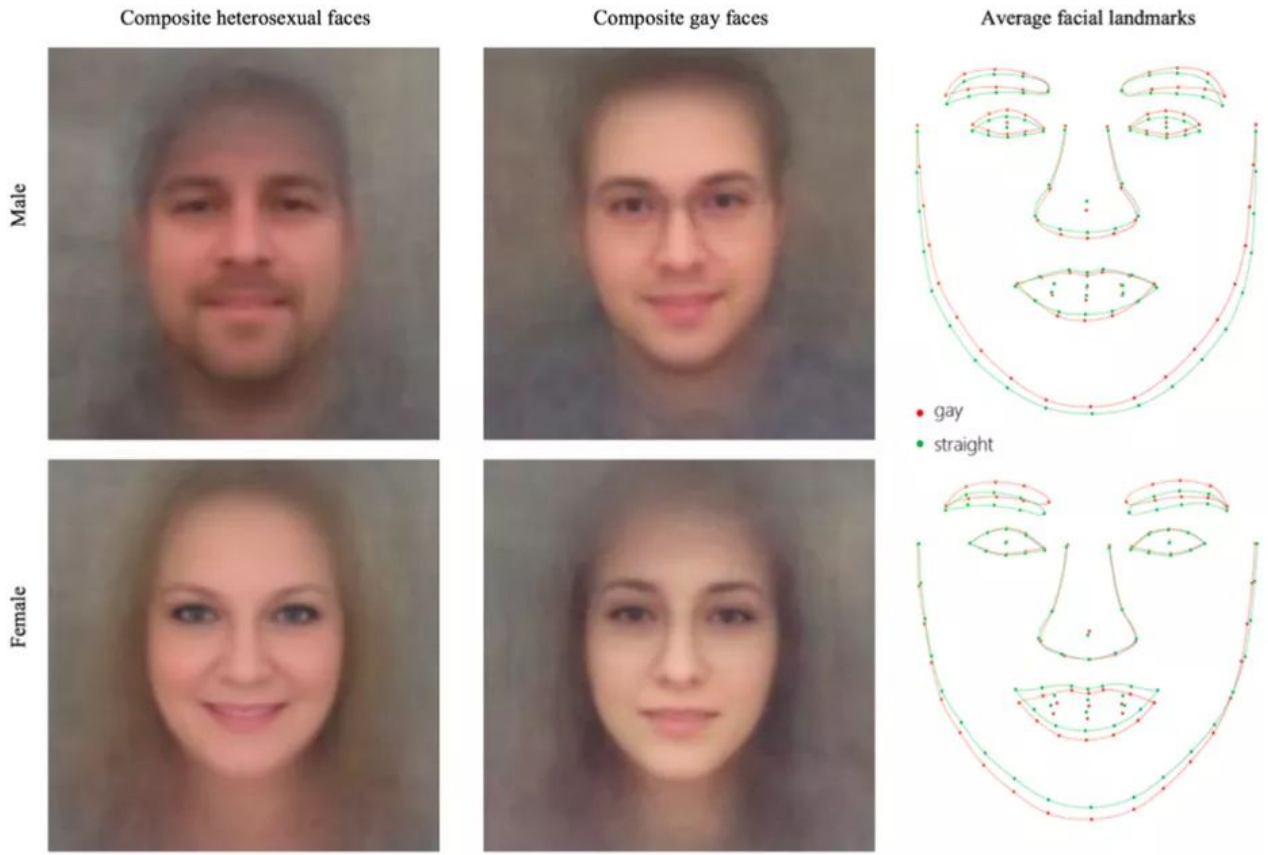



PAROLE




**Applications pervasive in our daily life!**

# A Recent Study: the “A.I. Gaydar”




Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology (in press).* September 7, 2017.

# A Case Study: the “A.I. Gaydar”

**Abstract.** We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone allowed for detecting gay males with 57% accuracy and gay females with 58% accuracy. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology (in press).* September 7, 2017.

# A Case Study: the “A.I. 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

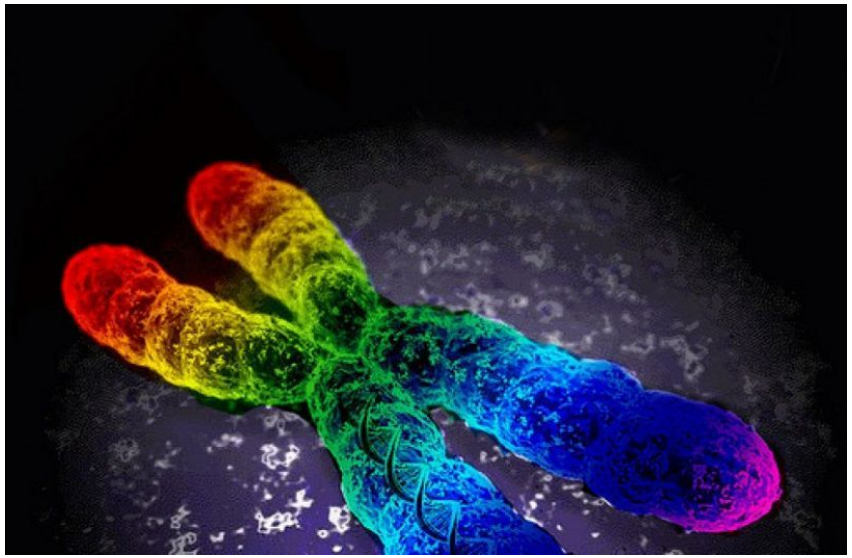
# Let's Discuss...

- 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 ethical questions could be asked here?**

# Research Question

- Identification of  sexual orientation from facial features



# Research Question

- Identification of  sexual orientation from facial features

**How  people can be harmed by this research?**

- 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, e.g. 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.

# Research Question

- Identification of  sexual orientation from facial features

*"... Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women."*

→ your thoughts on this?

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

**Legal ≠ Ethical**
**Public ≠ Publicized**
**Did these people agree to participate in the study?**

**→ Violation of 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

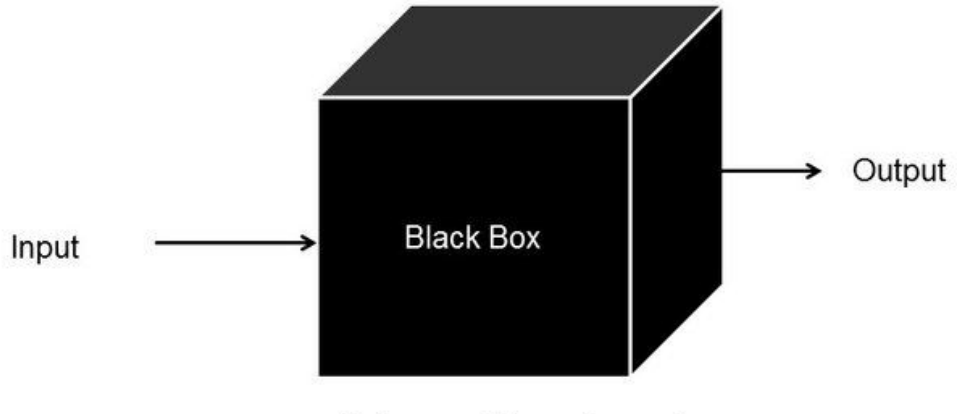
# Data & Bias

- 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

**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 so there is even self-selection bias…**
**The dataset is balanced, which does not represent true class distribution.**

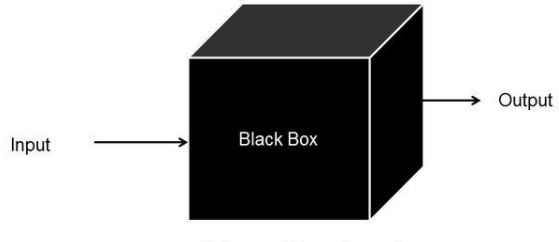
# Method

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



# Method  & Human Biases in Models + Interpretability

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

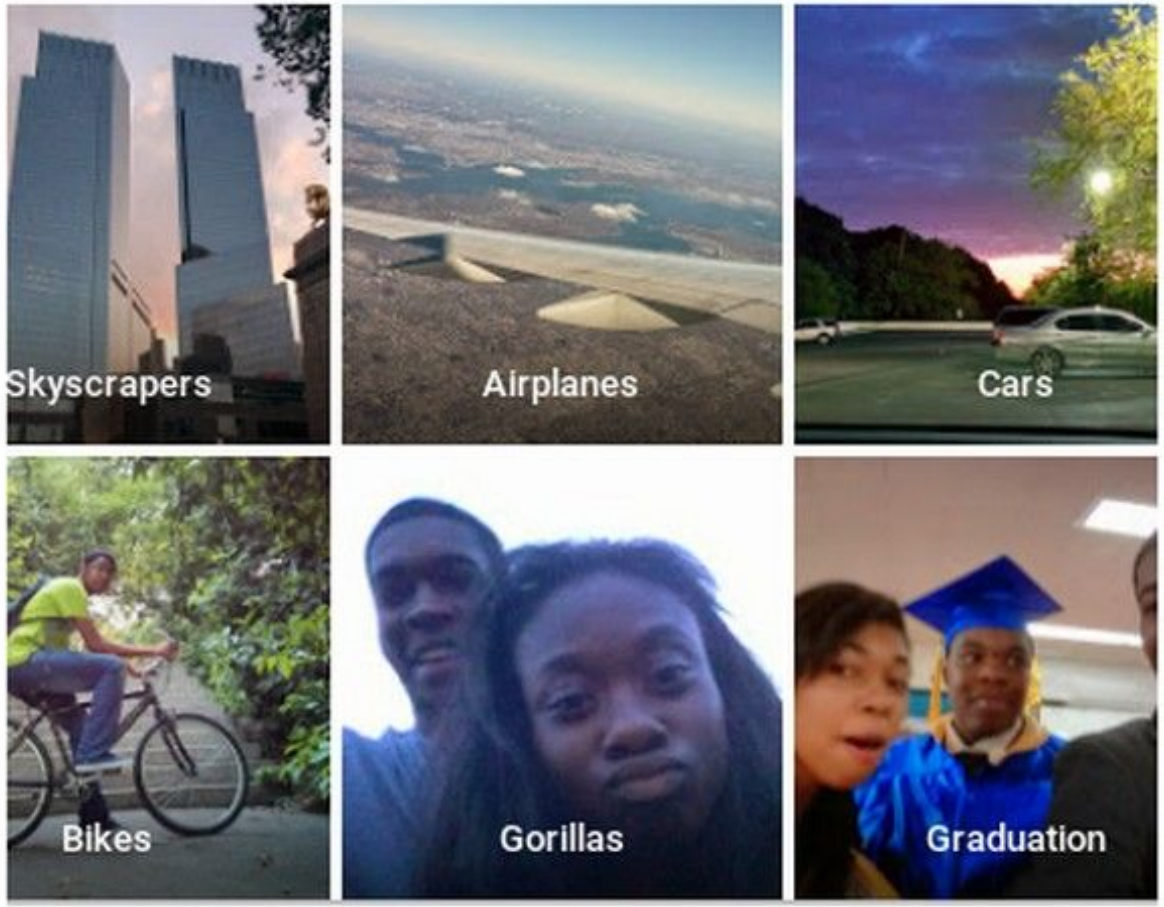



- **can we use not interpretable models when we make predictions about sensitive attributes, about complex experimental conditions that require broader world knowledge?**
- **how to deal with bias amplification?**

# Evaluation

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

# The Cost of Misclassification




The label Jacky saw on all of the pictures he had taken with this particular friend. (Jacky Alciné/Twitter)

# A Different Project?


- Framing

*"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. "*

# The Dual Use of A.I. Technologies

- Who should be responsible?
  - The person who uses the technology?
  - The researcher/developer?
  - Paper reviewers?
  - 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**

# Learn to Assess AI Systems Adversarially

- Who could benefit from such a technology?
- Who can be harmed by such a technology?

- Representativeness of training data
- Could sharing this data have major effect on people's lives?

- What are confounding variables and corner cases to control for?
- Does the system optimize for the "right" objective?
- Could prediction errors have major effect on people's lives?

# Learn to Assess AI Systems Adversarially

- Who could benefit from **your** technology?
- Who can be harmed by **your** technology?

- Representativeness of **your** training data
- Could **you** by sharing this data have major effect on people's lives?

- What are confounding variables and corner cases **for you** to control for?
- Does **your** system optimize for the "right" objective?
- Could prediction errors of **your** technology have major effect on people's lives?

# Topics Discussed in 11-830

Part 1: Theoretical foundations

- What is ethics
- History
- Medical and psychological experiments
- IRB and human subjects

# Topics Discussed in 11-830

Part 2: Misrepresentation and bias

- Theoretical background, IAT
- Algorithms to identify bias in NLP models and data
- Debiasing

# Topics Discussed in 11-830

Part 3: Civility in communication

- Techniques to monitor trolling, hate speech, abusive language, cyberbullying, toxic comments
- Hate speech and bias in conversational agents

# Topics Discussed in 11-830

Part 4: Privacy

- Algorithms for demographic inference and personality profiling
- Style transfer and anonymization of demographic and personal traits

# Topics Discussed in 11-830

Part 5: Democracy and the language of manipulation

- Approaches to identify propaganda and manipulation in news
- Fake news
- Political and media framing
- Respect, power, agency in discourse

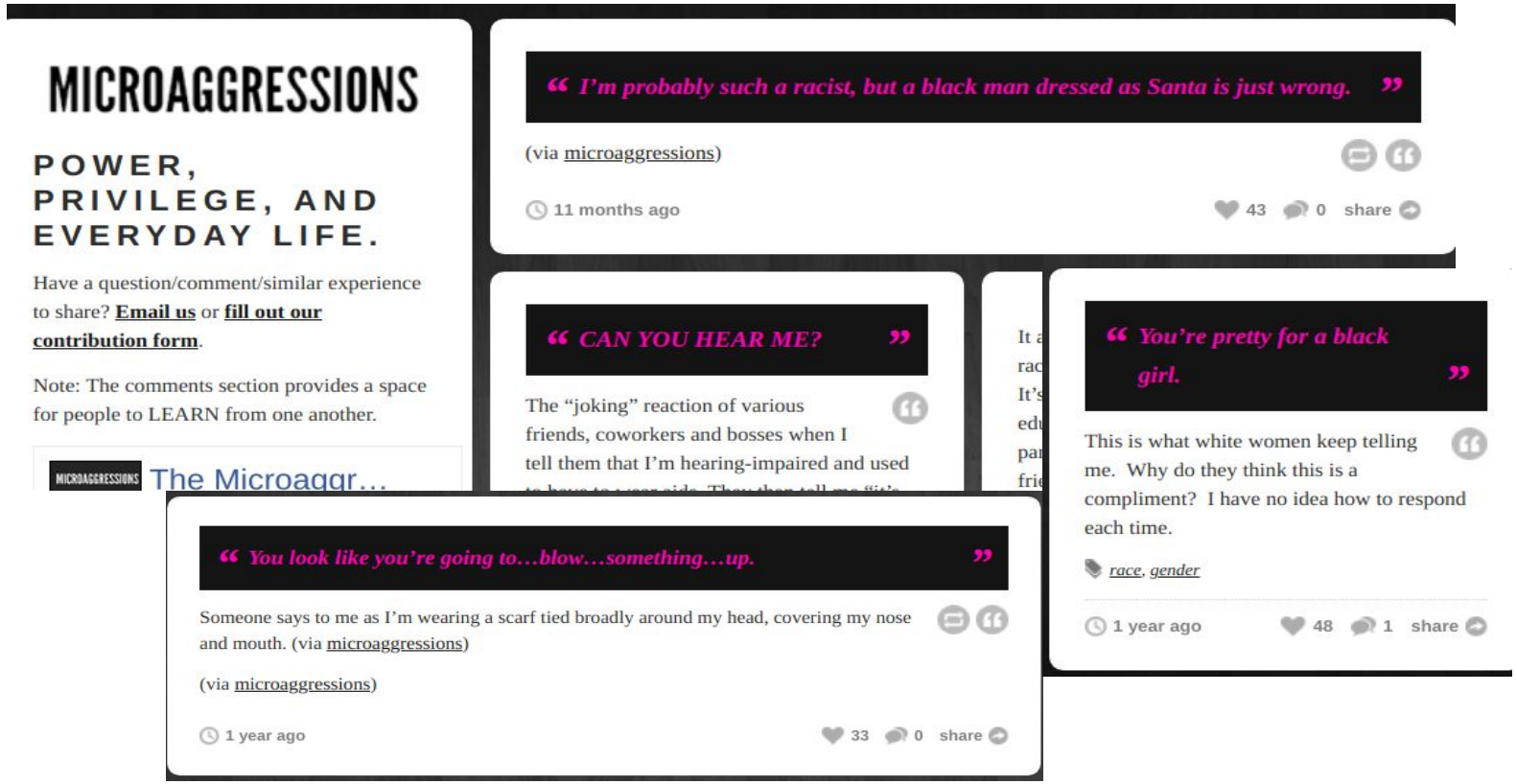
# Topics Discussed in 11-830

Part 6: NLP for social good

- Low-resource NLP
- NLP for disaster response
- Interfaces for helping people

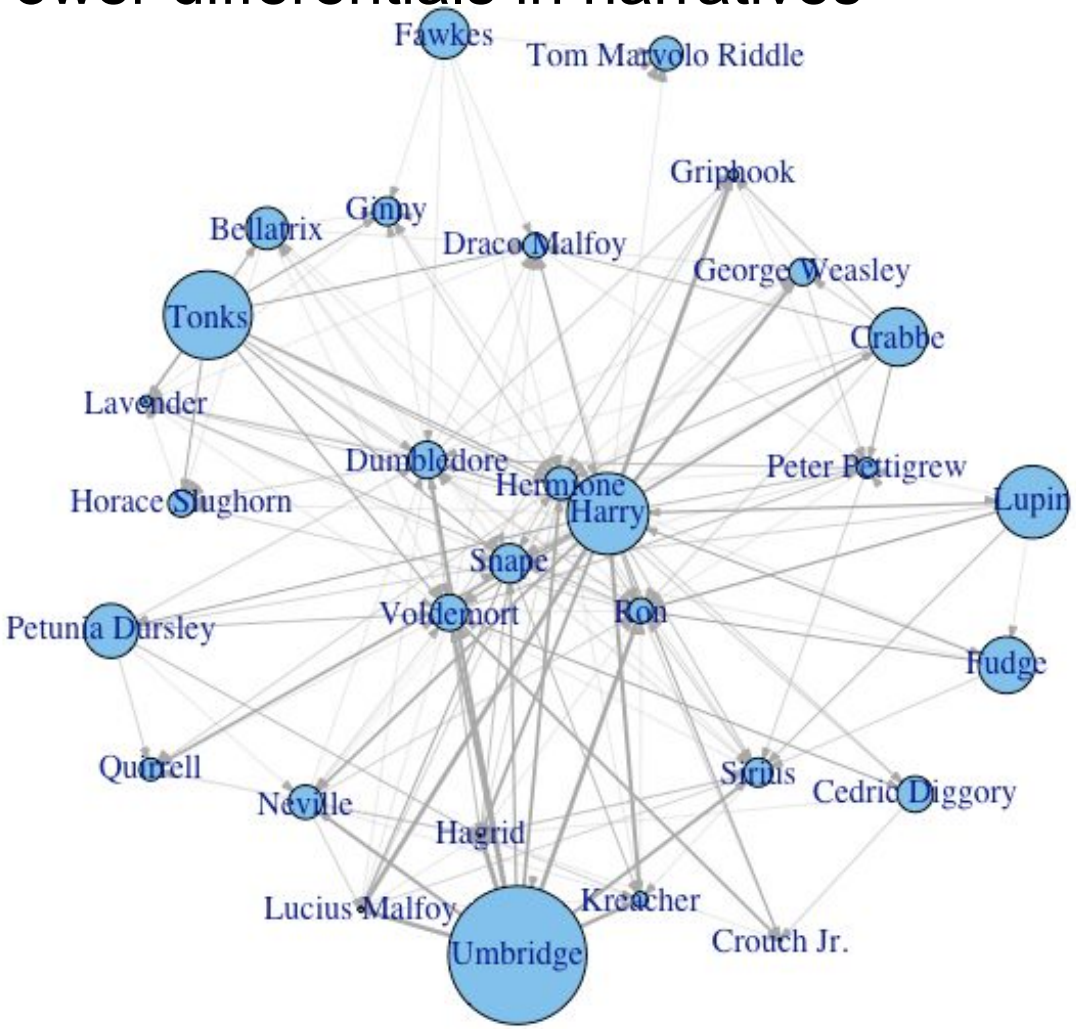
# Examples of Course Projects in 11-830

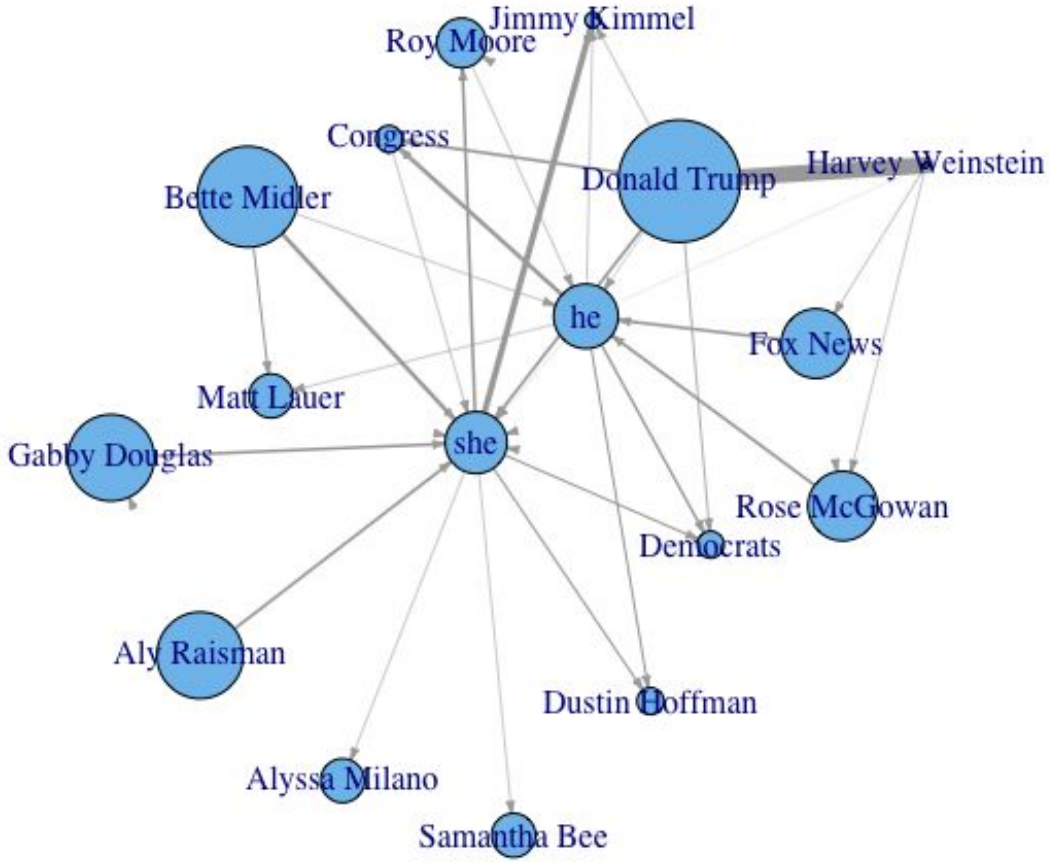
- active-learning based annotation procedure that increases the likelihood of surfacing posts containing microaggressions



# Examples of Course Projects in 11-830

- Power differentials in narratives



# Examples of Course Projects in 11-830

- Identification of gender bias in TED comments

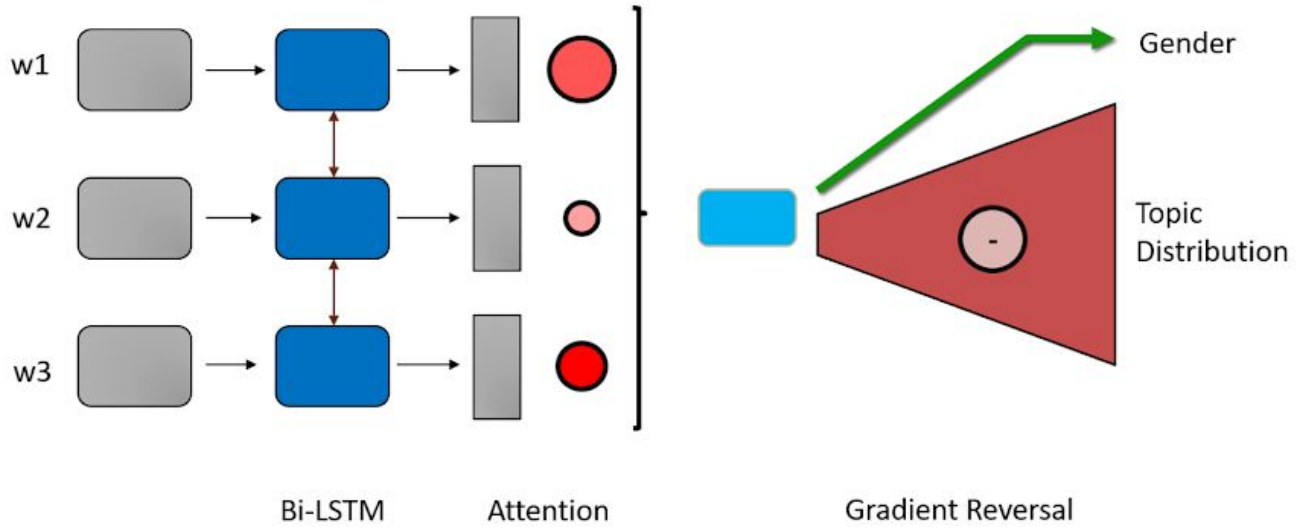



Figure 1: Proposed Model Architecture

# Summary

## Learn to Assess AI Systems Adversarially

- Who could benefit from **your** technology?
- Who can be harmed by **your** technology?

- Representativeness of **your** training data
- Could **you** by sharing this data have major effect on people's lives?

- What are confounding variables and corner cases **for you** to control for?
- Does **your** system optimize for the "right" objective?
- Could prediction errors of **your** technology have major effect on people's lives?