Natural Language Processing

Prof. Dr. Jannik Strötgen
jannik.stroetgen@h-ka.de

Summer 2024

Hochschule Karlsruhe
University of
Applied Sciences

Fakultät für
Informatik und
Wirtschaftsinformatik

Ethics for NLP

(material adapted from Annemarie Friedrich's and Thorsten Zesch's "A Crash Course on Ethics in Natural Language Processing", version 2)

Ethics for NLP

- What comes to your mind when you think of **ethics**?
- What comes to your mind when you think about **ethics for NLP**?
- Have you encountered any **ethical problems** in your life?
- Why do you think this topic is **important**?
- What do you expect to **learn** in this crash course?

Why does Ethics Matter for NLP?

- NLP has the aim of modeling **language**, an inherently human function
- NLP works with **textual data or human subjects**
  → not free of bias, prejudice, …
- Language technology is **widely applied**
  (e.g., on social media)
  → can potentially harm anyone
- Language technology shapes the way we **experience** the world

Bias

Privacy

Fairness

Dual Use

Environmental Issues

…

Learning Goals

After this course, you will be able to:

- Understand terminology and concepts related to ethics in NLP
- Analyze a given task, method or system for ethical issues
- Understand how NLP applications can cause harm
- Analyze ethical issues under different ethical perspectives

What is Ethics?

Branch of Philosophy

Ethics is the **philosophical study of morality**. It is the study of what are good and bad ends to pursue in life and what is **right** and **wrong** to do in the conduct of life. It is [...] primarily a **practical** discipline.

Synonym for Moral Code

Sometimes “ethics” is used to refer to the **moral code** or system of a particular tradition.

Examples: Christian ethics, professional ethics

How do these meanings relate to “Ethics for NLP”?

John Deigh: An Introduction to Ethics. Cambridge 2010. Page 7.

What is Morality?

Universal Concept

Universal ideal of what one ought to do or ought not to do, guided by reason / rational grounds.

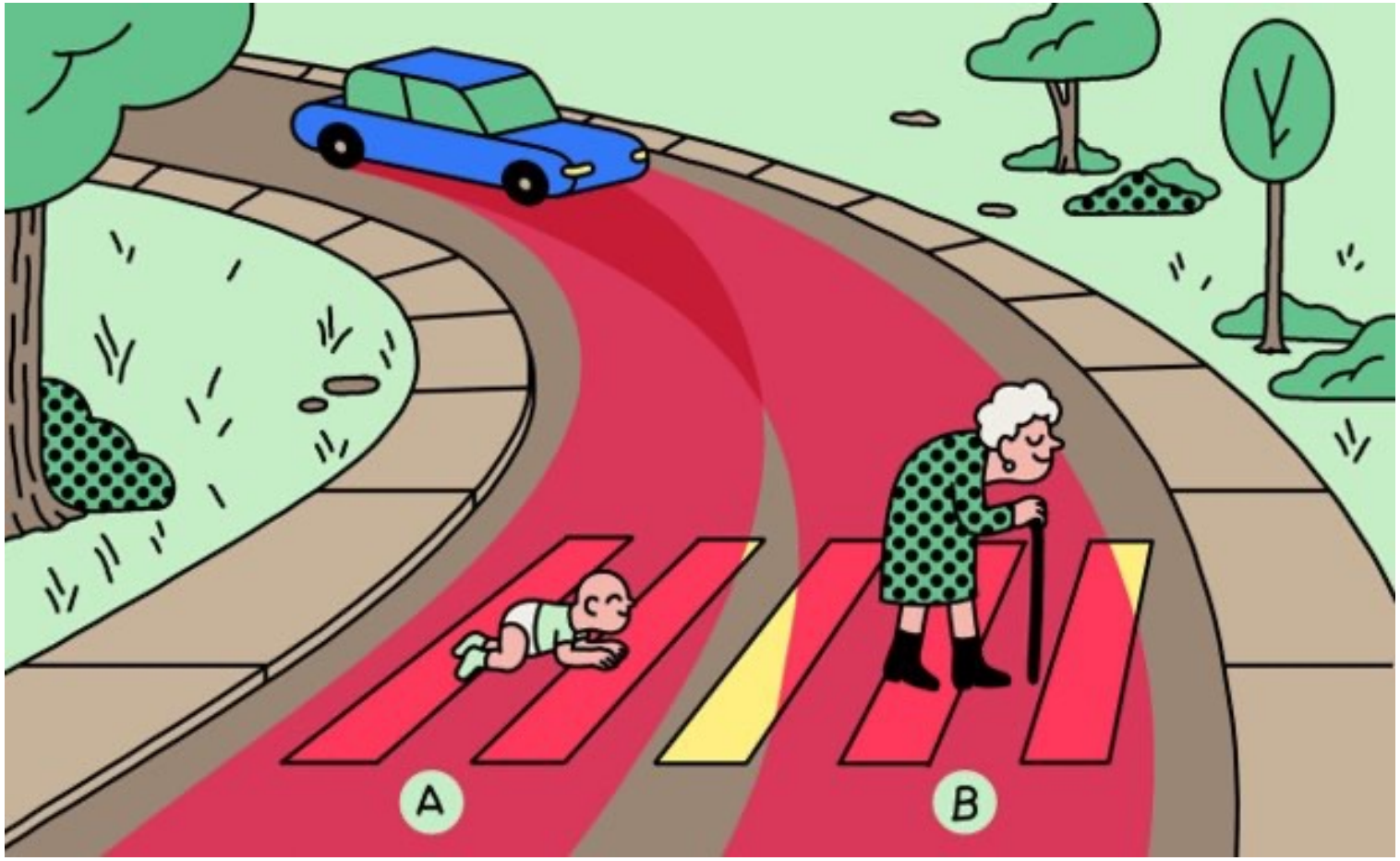
Conventional System of Community

The members' shared beliefs about wrong and right, good and evil, and the corresponding customs and practices that prevail in the society.

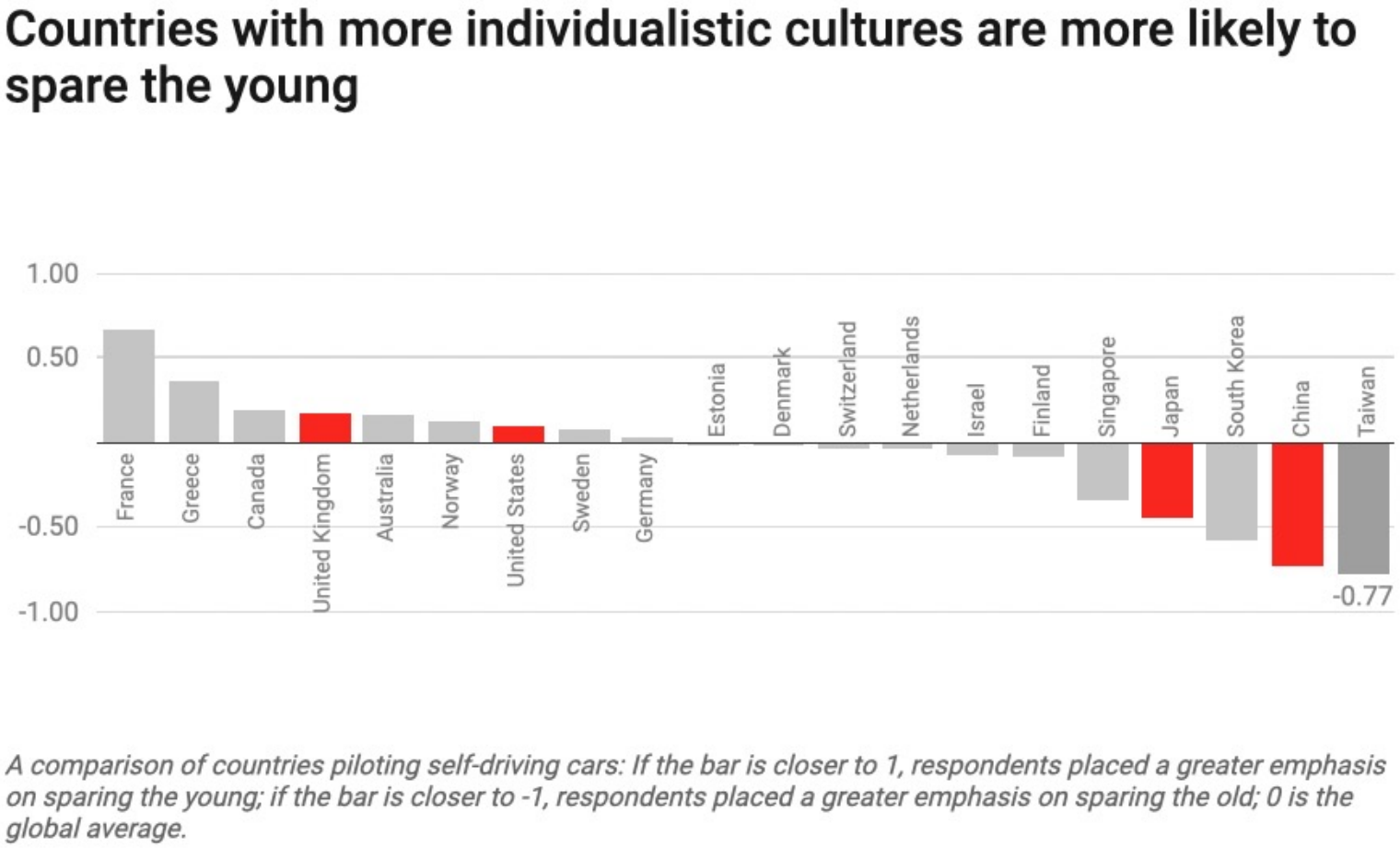
How do these concepts relate to "Ethics for NLP"?

John Deigh: An Introduction to Ethics. Cambridge 2010. Page 8.

Whose Life Matters More?

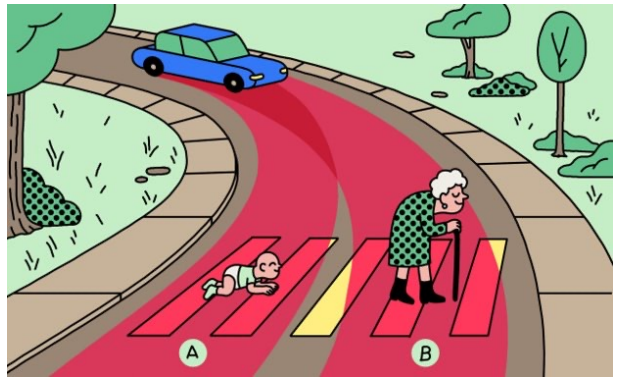


Whose Life Matters More?



Try it out!

http://moralmachine.mit.edu/hl/de



Two Ethical Theories

Teleology

Telos (Greek) = goal

Outcome-oriented

Utilitarianism

“Choose that action that optimizes the outcome”

“An action is ethical only if it is not irrational for the agent to believe that no other action results in greater expected utility” (Bentham 1789)

Deontology

Deon (Greek) = duty

“Identify your duty and act accordingly”

Generalization principle:
prioritizes intent as the source of ethical action, should be reasonable.

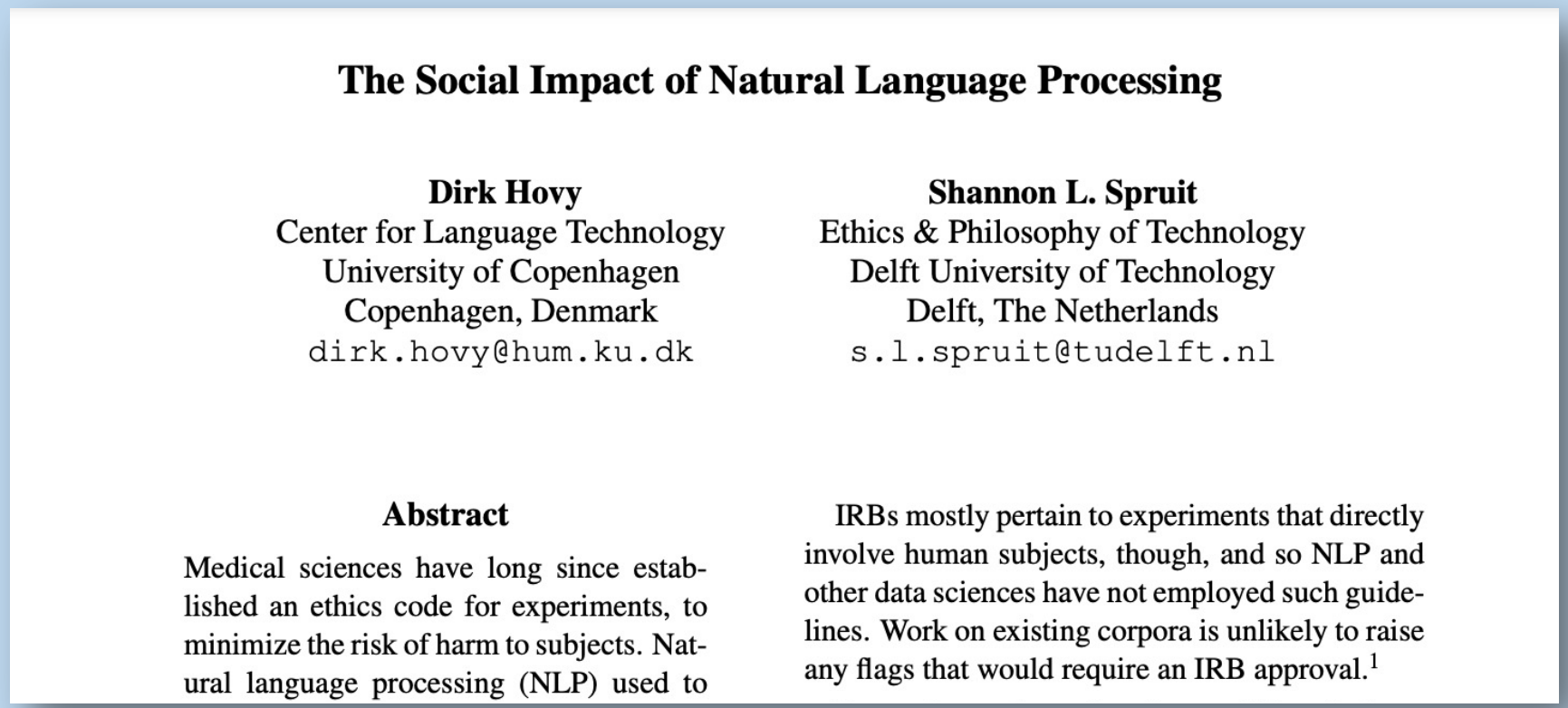
John Hooker. 2018. Taking Ethics Seriously: Why Ethics Is an Essential Tool for the Modern Workplace. Taylor and Francis.

Moral vs. Legal – or: Ethics ≠ Law

	legal	illegal
moral	doing your homework	civil disobedience
immoral	cheating on your spouse	murder

Homework

Hovy & Spruit: The Social Impact of Natural Language Processing. (ACL 2016)



The Social Impact of Natural Language Processing

Dirk Hovy
Center for Language Technology
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Shannon L. Spruit
Ethics & Philosophy of Technology
Delft University of Technology
Delft, The Netherlands
s.l.spruit@tudelft.nl

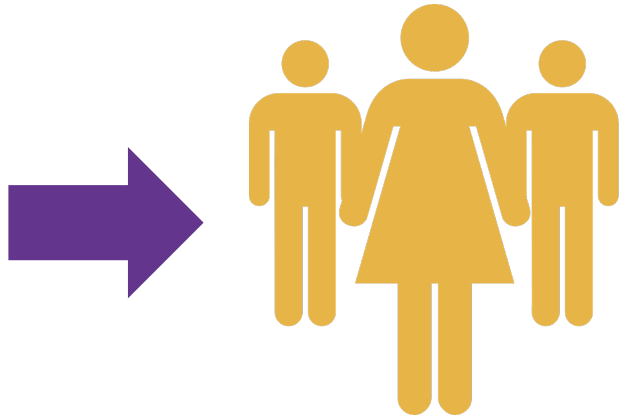
Abstract

Medical sciences have long since established an ethics code for experiments, to minimize the risk of harm to subjects. Natural language processing (NLP) used to

IRBs mostly pertain to experiments that directly involve human subjects, though, and so NLP and other data sciences have not employed such guidelines. Work on existing corpora is unlikely to raise any flags that would require an IRB approval.[1]

Source of Harm – Direct

NLP System



Analyzing medical documents

Drug overdose killing the patient

Image by CIker-Free-Vector-Images from Pixabay

Dual Use

NLP Task	Beneficial Use	Malicious Use
Hate Speech Detection	fighting hate crimes	censorship of free speech
Detection of fake news / reviews	fighting misinformation	generation of fake news / reviews
…		

Can you think of other NLP tasks that have beneficial but also potentially malicious uses?

Assume you are publishing a piece of software on GitHub. Should you mention potential malicious uses in the corresponding readme?

Doctor or Nurse

The doctor recommended to perform an X-ray.

He/She said …

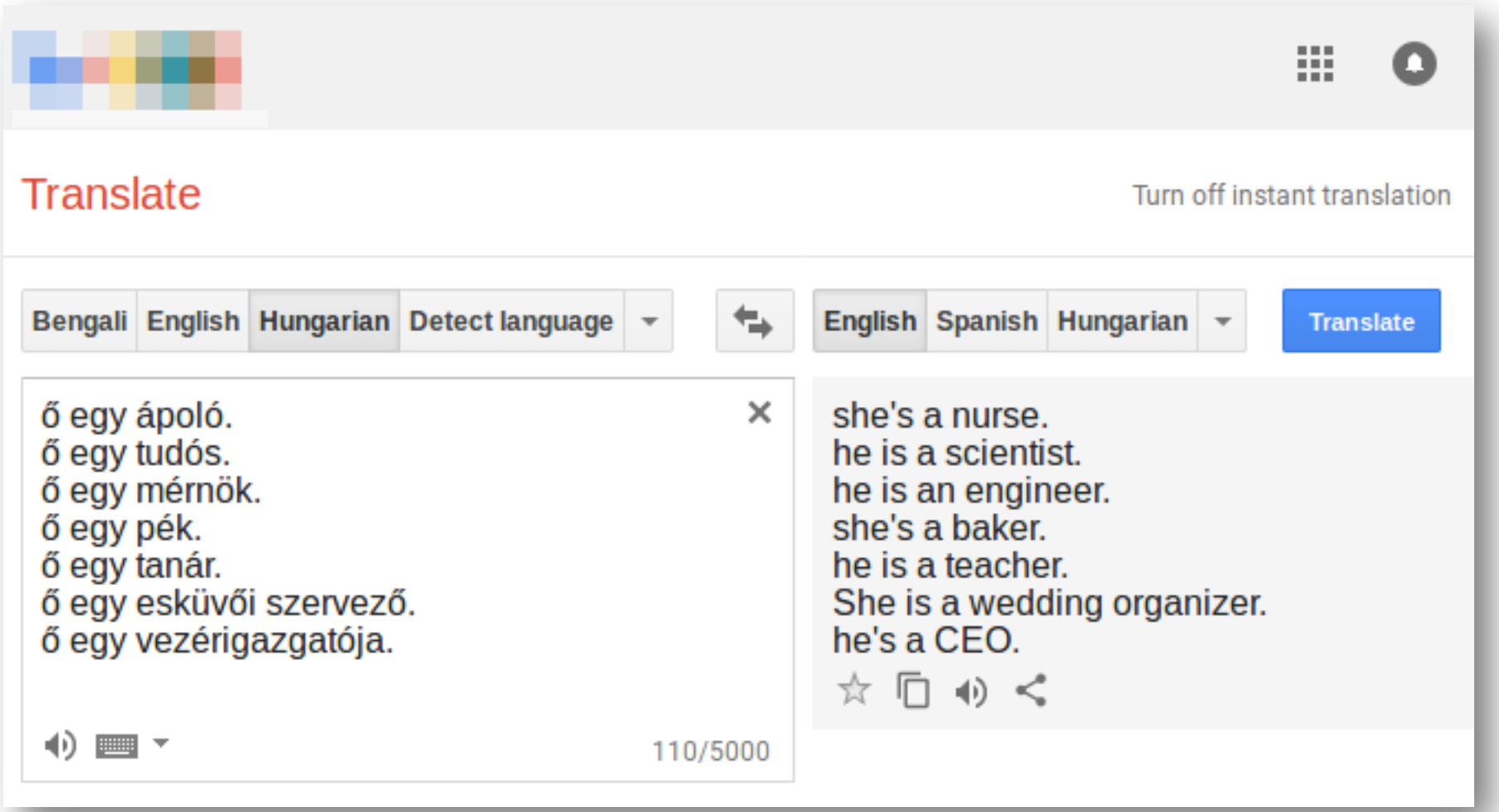
The nurse recommended to perform an X-ray.

He/She said …

Do you think “he” or “she” is a more likely continuation in the above cases (respectively)?

What would happen if you asked a large pre-trained language model?

Bias in Machine Translation



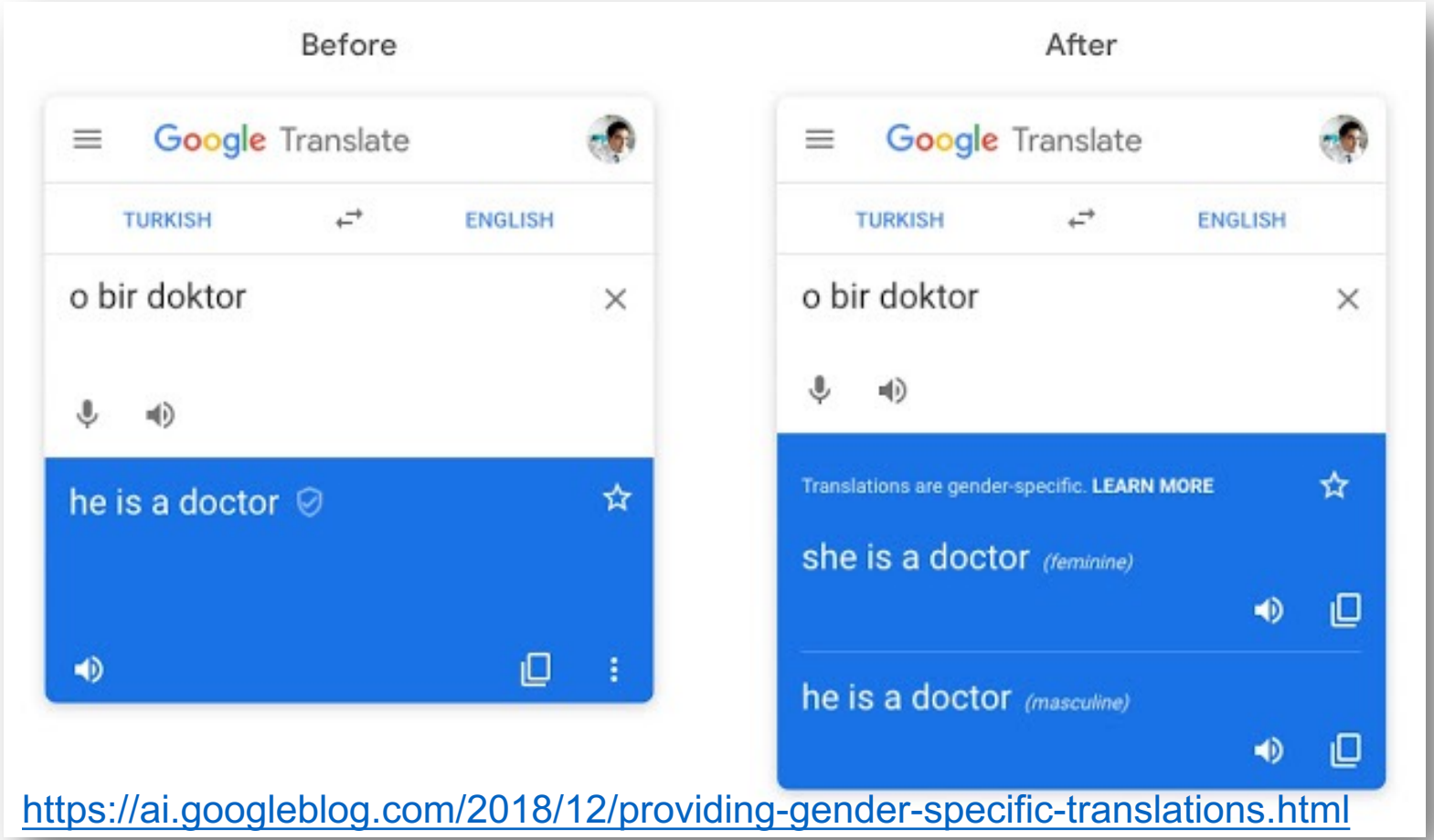
Prates et al. (2018): Assessing Gender Bias in Machine Translation -- A Case Study with Google Translatehttps://arxiv.org/pdf/1809.02208.pdf

Bias in Machine Translation

Detecting gender-neutral queries

Generate gender-specific translations

Check for accuracy



https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

My XAI SLIDES?

What is Bias?

Cognitive bias arises due to the tendency of the human mind to categorize the world.

→ simplifies processing.

Social biases in data, algorithms, and applications

Statistical bias in machine learning

Inductive bias: assumptions made by model about target function to generalize from data

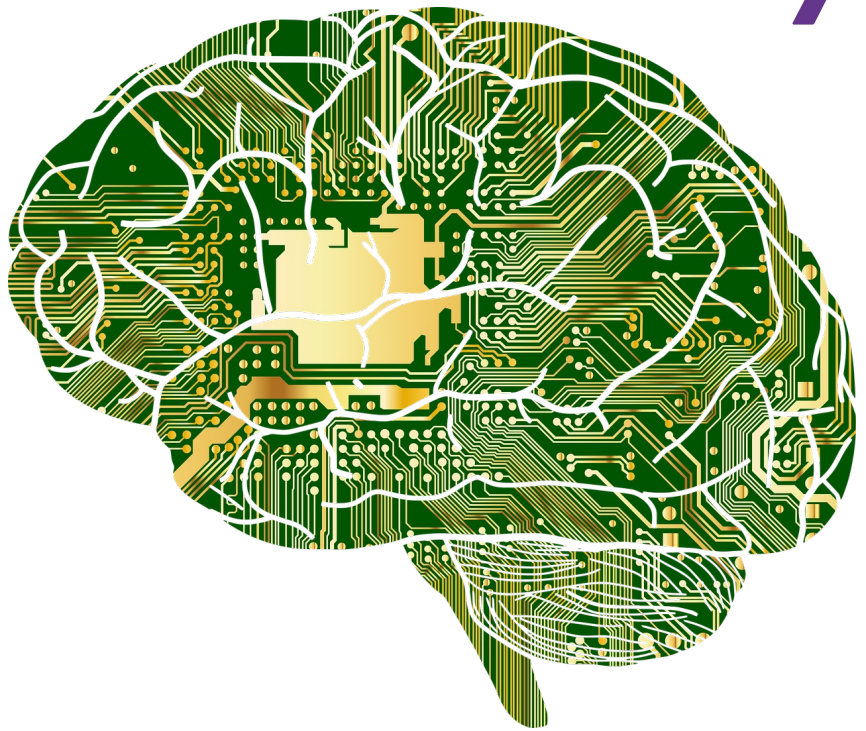


Image by Gordon Johnson on Pixabay.

What is Bias? (Technical View)

Bias in machine learning

$$y = ax + b$$

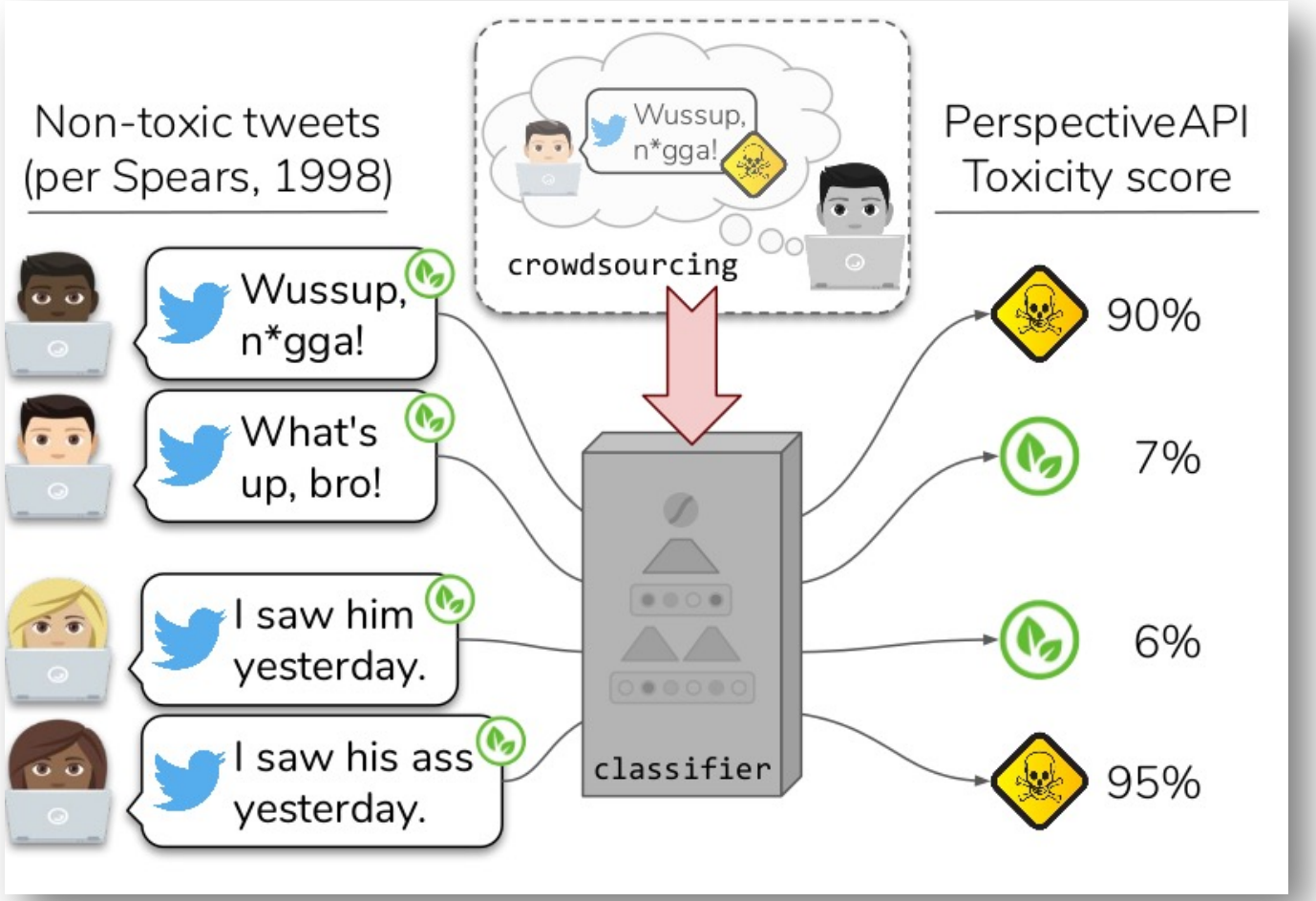
Bayesian probabilities: prior

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

May be intended (e.g., domain adaptation) or unintended

Is bias always a bad thing?

Why is Bias Problematic? (Social View)



NLP Applications

Employment matching, advertisement placement, parole decisions, search, chatbots, face recognition, ...

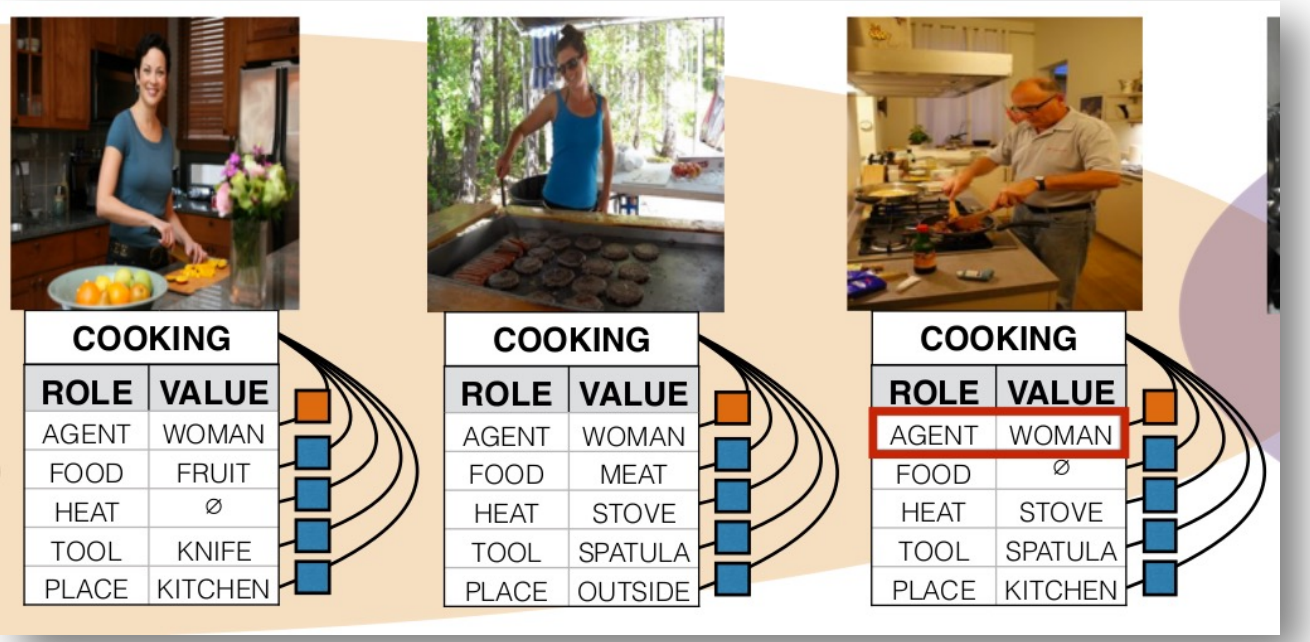
Social Stereotypes

Gender, Race, Disability, Age, Sexual orientation, Culture, Class, Poverty, Language, Religion, National origin, ...

Sap et al.: **The Risk of Racial Bias in Hate Speech Detection**. ACL 2019.

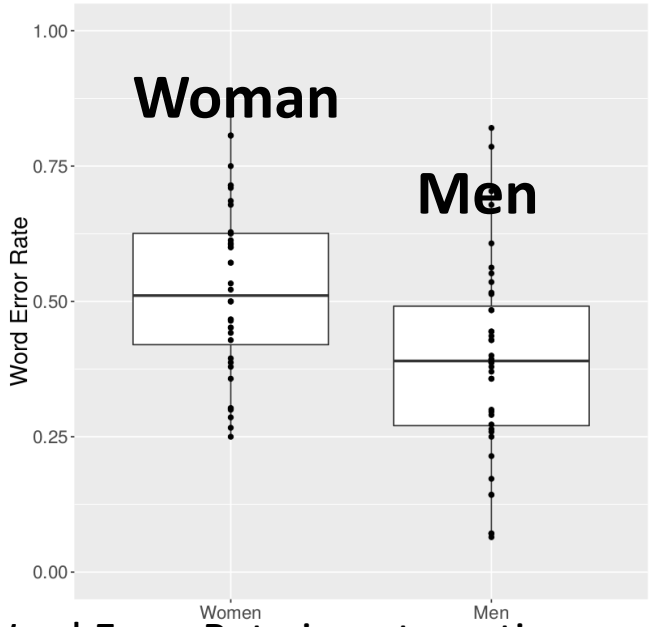
Why is Bias Problematic?

Outcome disparity



Because a “COOKING” event is taking place, the model is more likely to predict the agent to be a woman. (Zhao et al., 2017)

Error disparity



Word Error Rate in automatic captioning is higher for female speakers compared to male speakers (Tatman, 2017).

Why is Bias Problematic? (Technical View)

Outcome / Error disparity

Models might **amplify bias**

- 51:49 distributions in a feature may lead to 100:0 decisions

Is it wrong to build models replicating “real world data”?

Source of Harm – Unfair Outcomes

NLP System

Filtering job applications

Better chances for people living in a certain area

Fairness

Treating everyone equally is fair, right?

So, everyone gets the same grade from now on ;)

Bild von Gordon Johnson auf Pixabay

fundamental principle of justice
“equals should be treated equally
and unequals unequally”

Group vs. Individual Fairness

Group fairness

- errors should be distributed similarly across protected groups

Which groups are / should be protected?

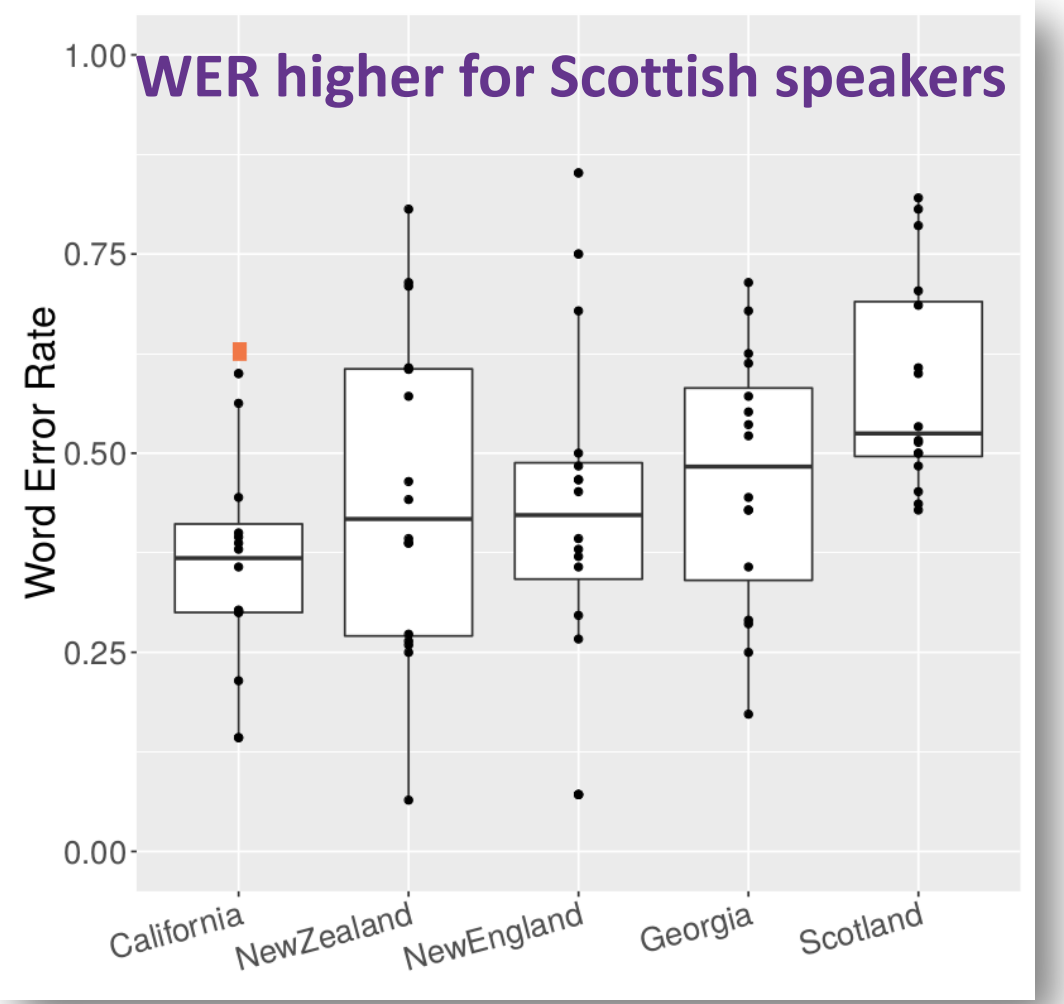
Individual fairness

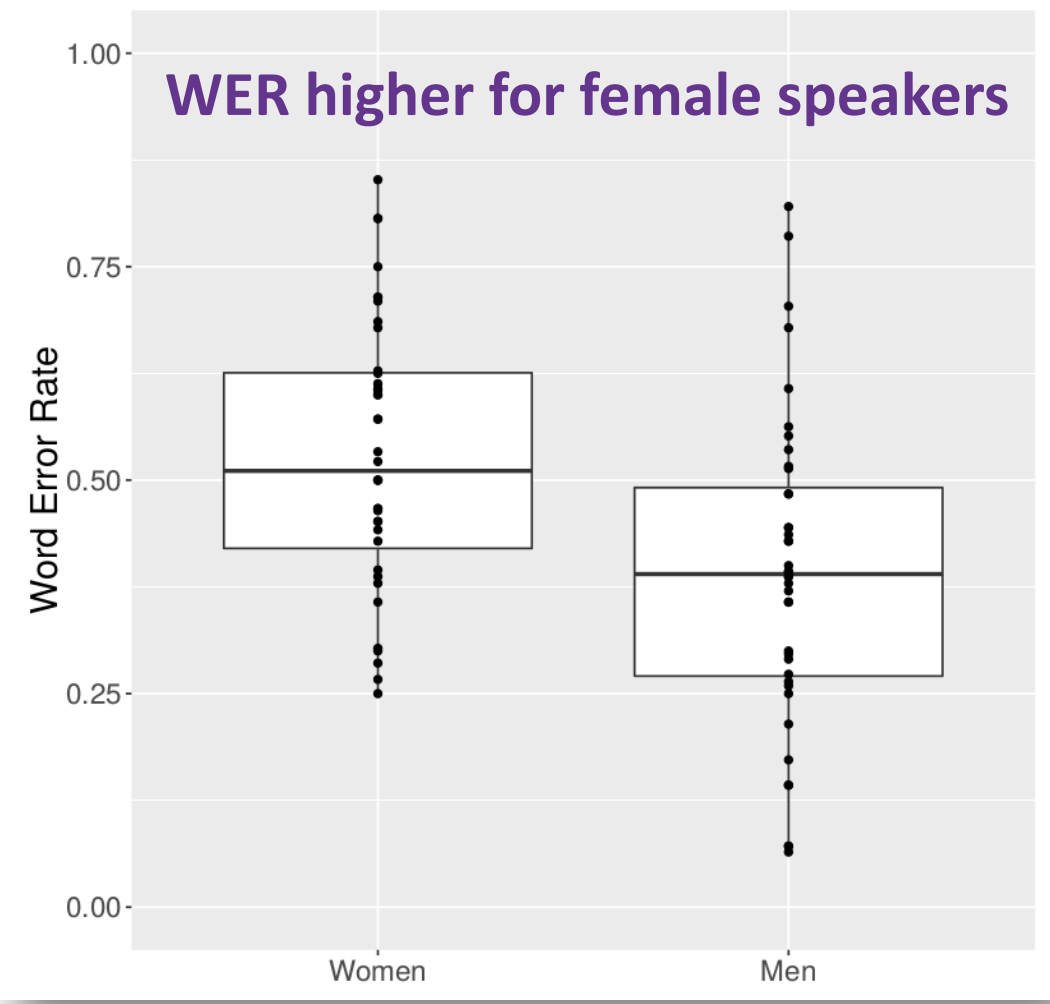
- similar individuals should be treated similarly regardless of group membership

How can we measure similarity of individuals?

We cannot reach group and individual fairness at the same time!

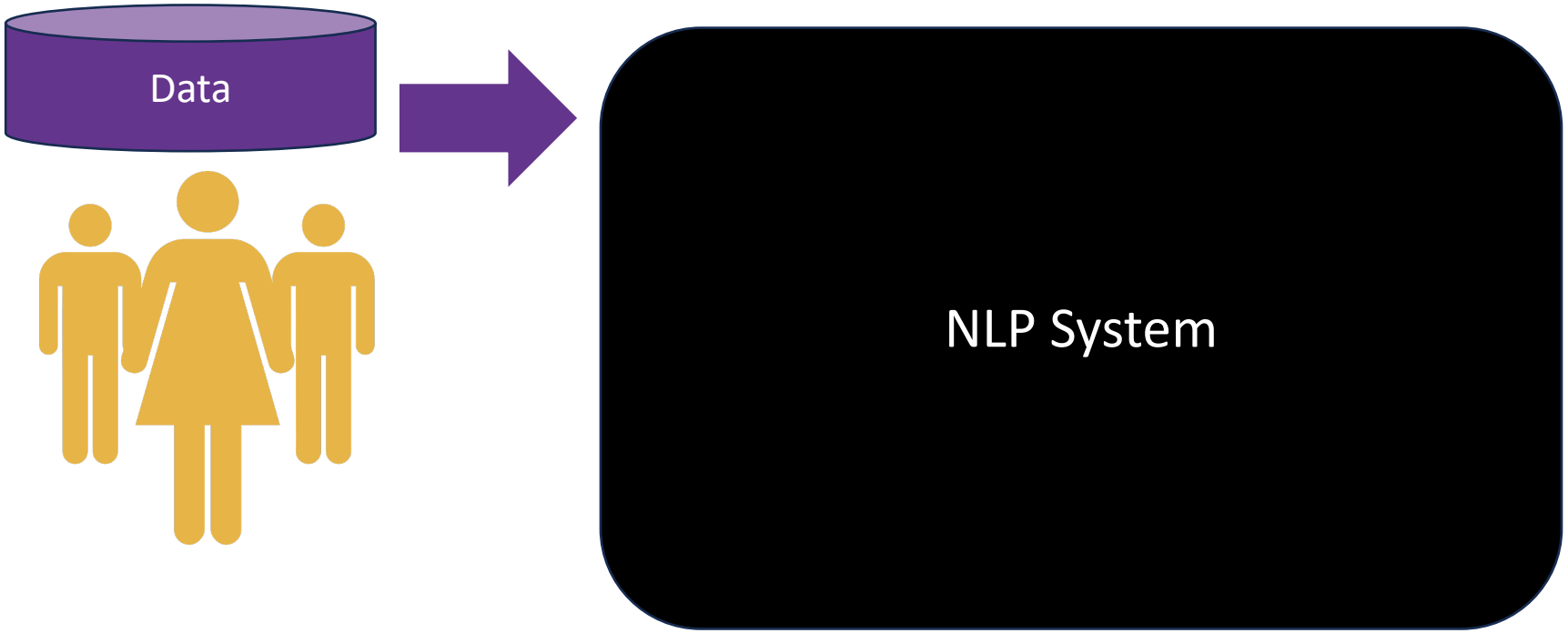
Word Error Rate for Automatic Captioning





Tatman (2017): Gender and Dialect Bias in YouTube's Automatic Captions https://www.aclweb.org/anthology/W17-1606

Source of Harm – Input / Training Data



Privacy

- “I‘ve got nothing to hide!“

Do you have curtains? Do you close your shutters at night?

Can I see your credit card bills from last year?

A Taxonomy of Privacy (Solove, 2007)

- Privacy = intimacy?
- Privacy = the right to be let alone?

> "Privacy [...] is a plurality of different things that do not share one element in common but that nevertheless bear a resemblance to each other."

Problems and harms related to privacy

Information Collection
- *Surveillance*
- *Interrogation*

Information Processing
- *Aggregation*
- *Identification*
- *Insecurity*
- *Secondary Use*
- *Exclusion*

Information Dissemination
- *Breach of Confidentiality*
- *Disclosure*
- *Exposure*
- *Increased Accessibility*
- *Blackmail*
- *Appropriation*
- *Distortion*

Invasion
- *Intrusion*
- *Decisional Interference*

Data Privacy vs. Data Ethics

- **Data privacy** is responsibly collecting, using and storing data about people, in line with the expectations of those people, customers, regulations and laws.
- **Data ethics** is doing the right thing with data, considering the human impact from all sides, and making decisions based on your values.

[based on: Lawler, 2019]

- “Just because we can do something, doesn’t mean we should.”

Should a company sell user information to political campaigns?

Discussion

Questions to discuss:

- Which dimensions of privacy matter most to you?
- A software developer accidentally notices a document where a user is drafting a suicide note. Should he/she contact the police to save a life, or respect their user's secret?
- Can you imagine a situation where interfering with someone's privacy leads to an economic / financial issue for that person?

Thank you for your attention!

Questions?