



A Crash Course on Ethics in Natural Language Processing

Version 1.0

Annemarie Friedrich and Torsten Zesch

License: CC-BY

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

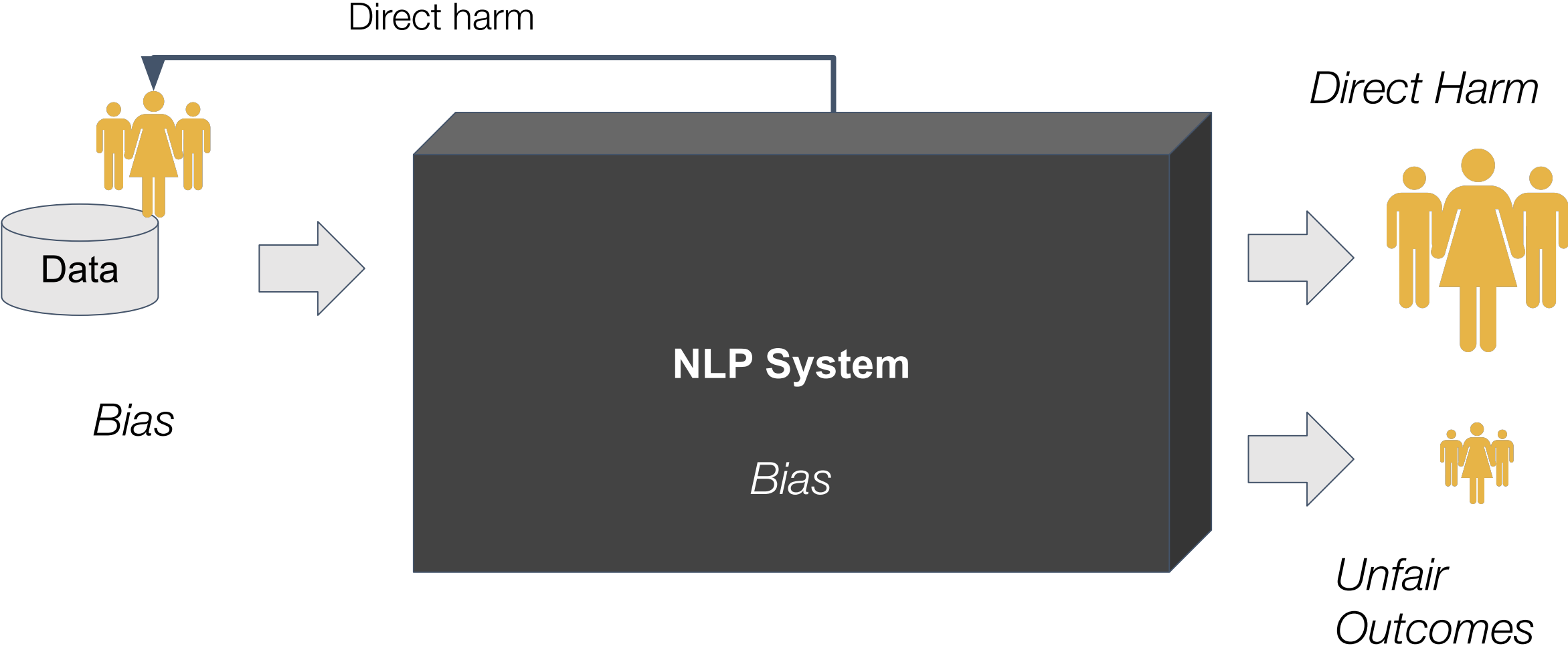
Fairness

Dual Use

Environmental
Issues

...

Sources and Types of Harm - Overview



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.

(Deigh, 2010, p. 7)

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”?

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.

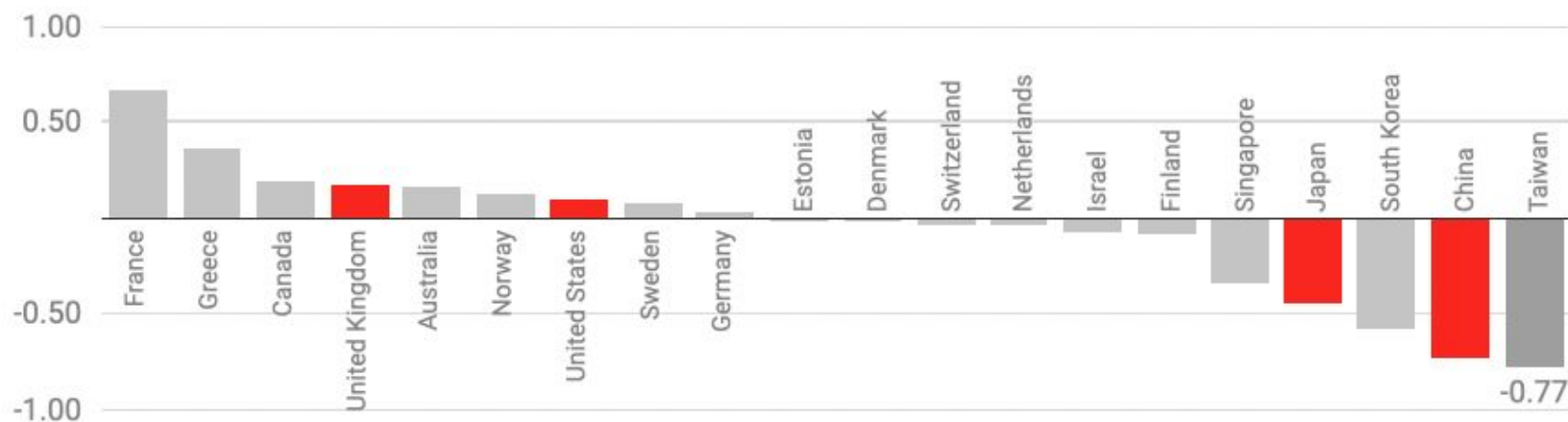
How do these concepts relate to “Ethics for NLP”?

Whose Life Matters More?

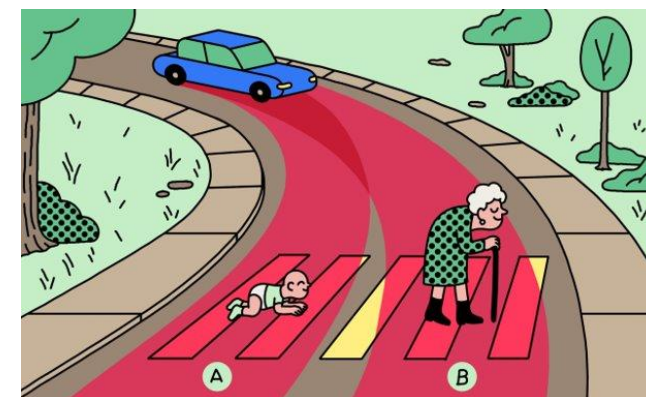
Countries with more individualistic cultures are more likely to spare the young

Try it out!

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



A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing the young; if the bar is closer to -1, respondents placed a greater emphasis on sparing the old; 0 is the global average.



Moral vs. Legal

	legal	illegal
moral	Doing your homework	Civil disobedience
immoral	Cheating on your spouse	Murder

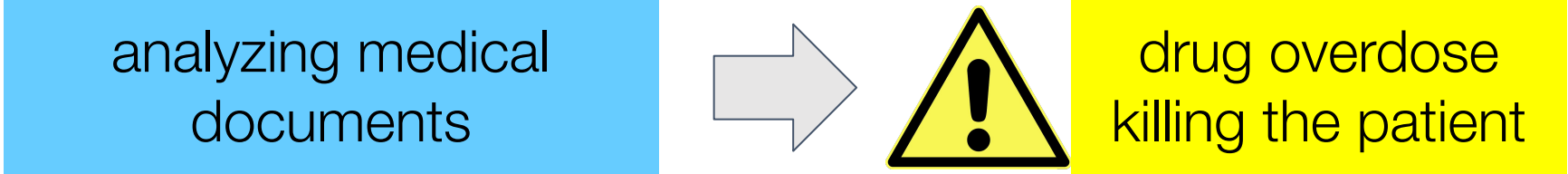
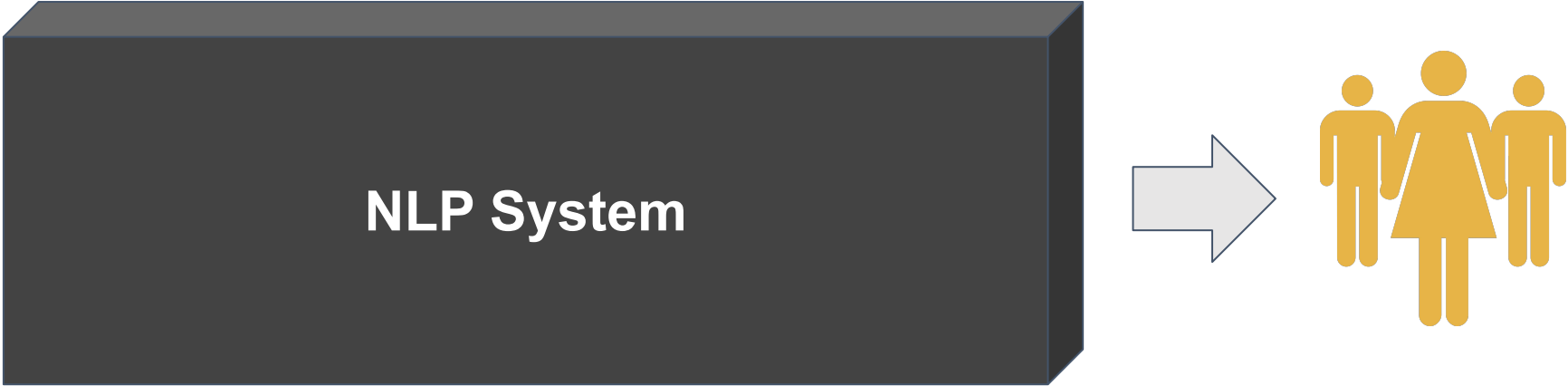
Reading Assignment (Homework)

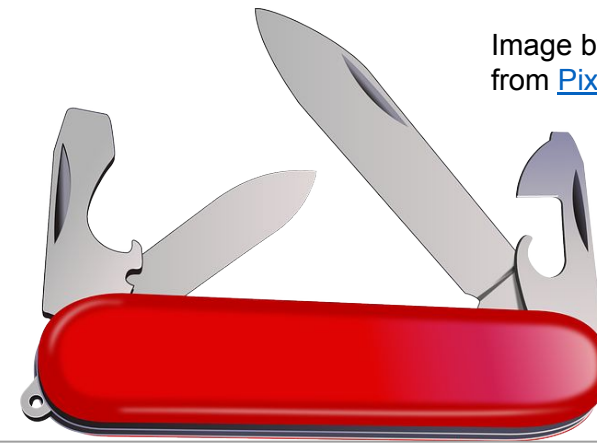
Hovy & Spruit: [The Social Impact of Natural Language Processing](#).
(ACL 2016)

TODO: add questions / instructions regarding the paper

Political correctness classifier?

Source of Harm - Direct





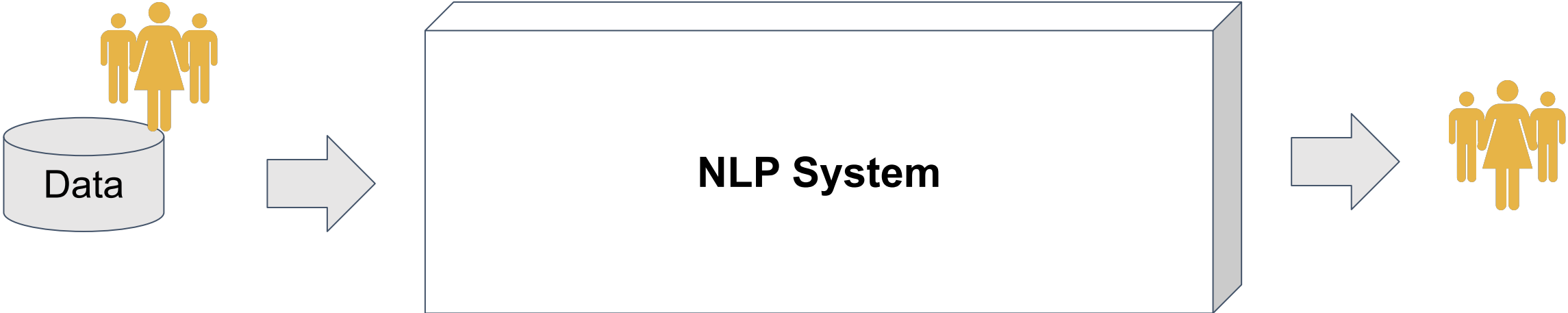
Dual Use

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?

Source of Harm - Bias



Doctor vs. 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



Translate

Useful or harmful?

Turn off instant translation

Bengali English Hungarian Detect language ▾



English Spanish Hungarian ▾

Translate

ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.



110/5000

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.



Image source: <https://arxiv.org/pdf/1809.02208.pdf>

Bias in Machine Translation

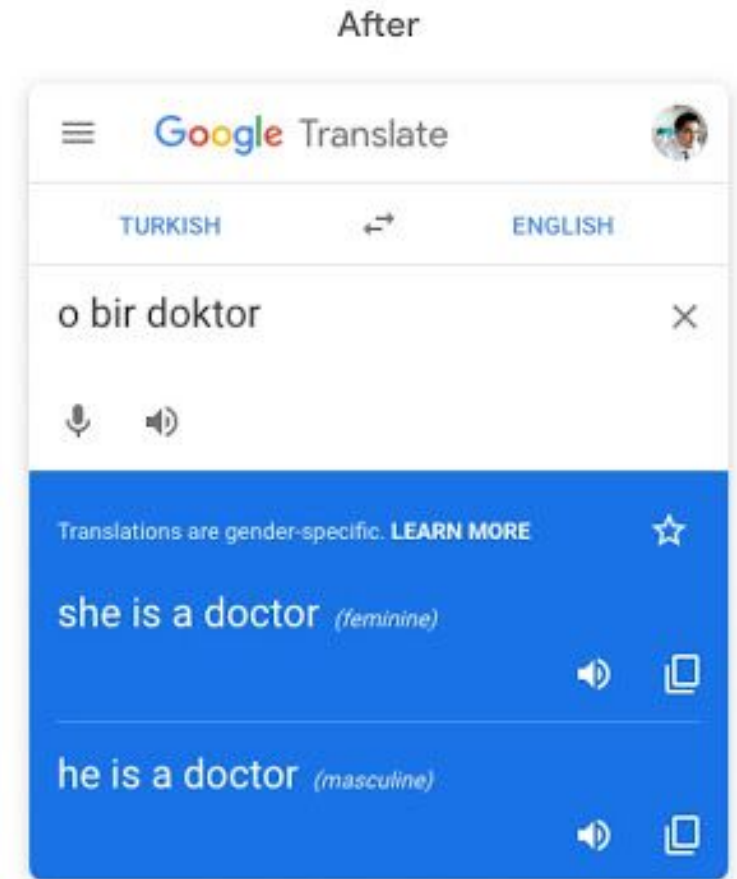
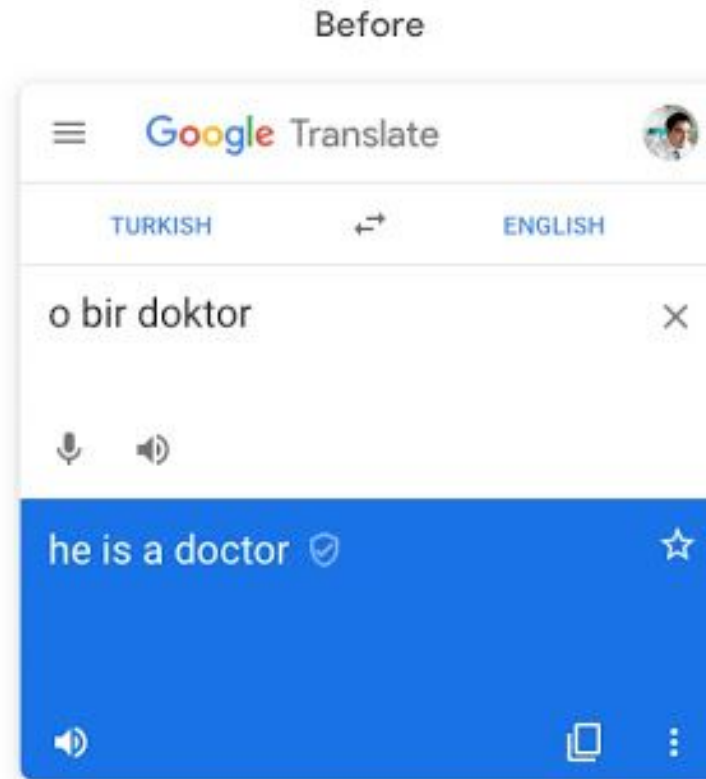
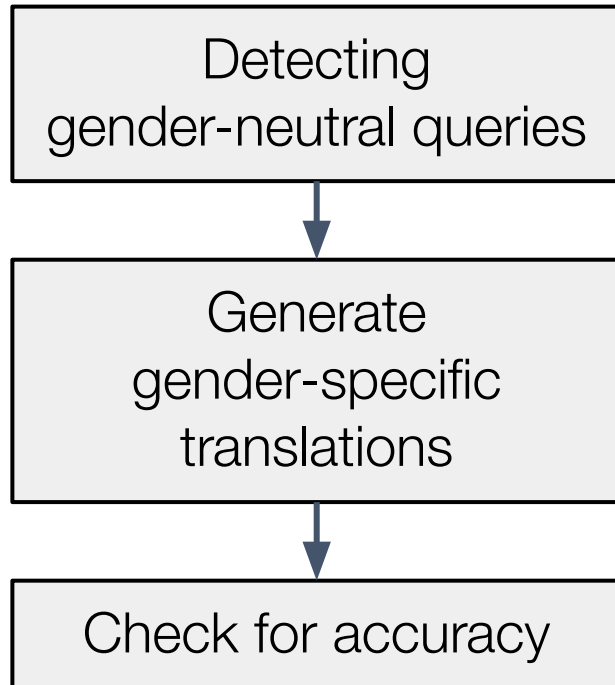


Image Source:

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

What is Bias?

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

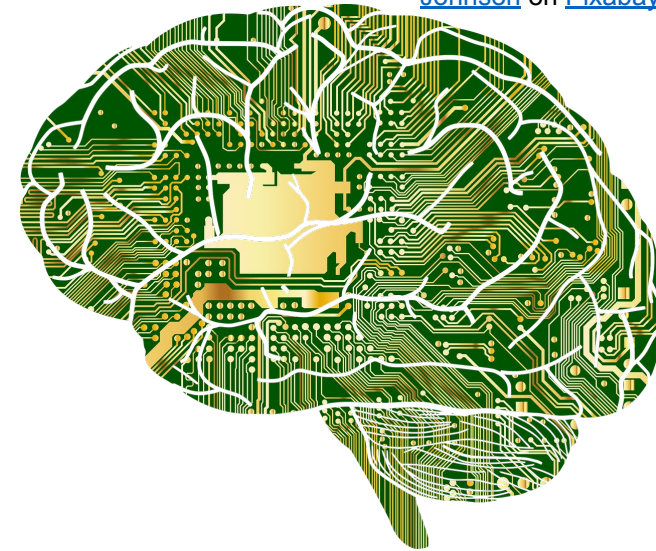
→ simplifies processing.

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

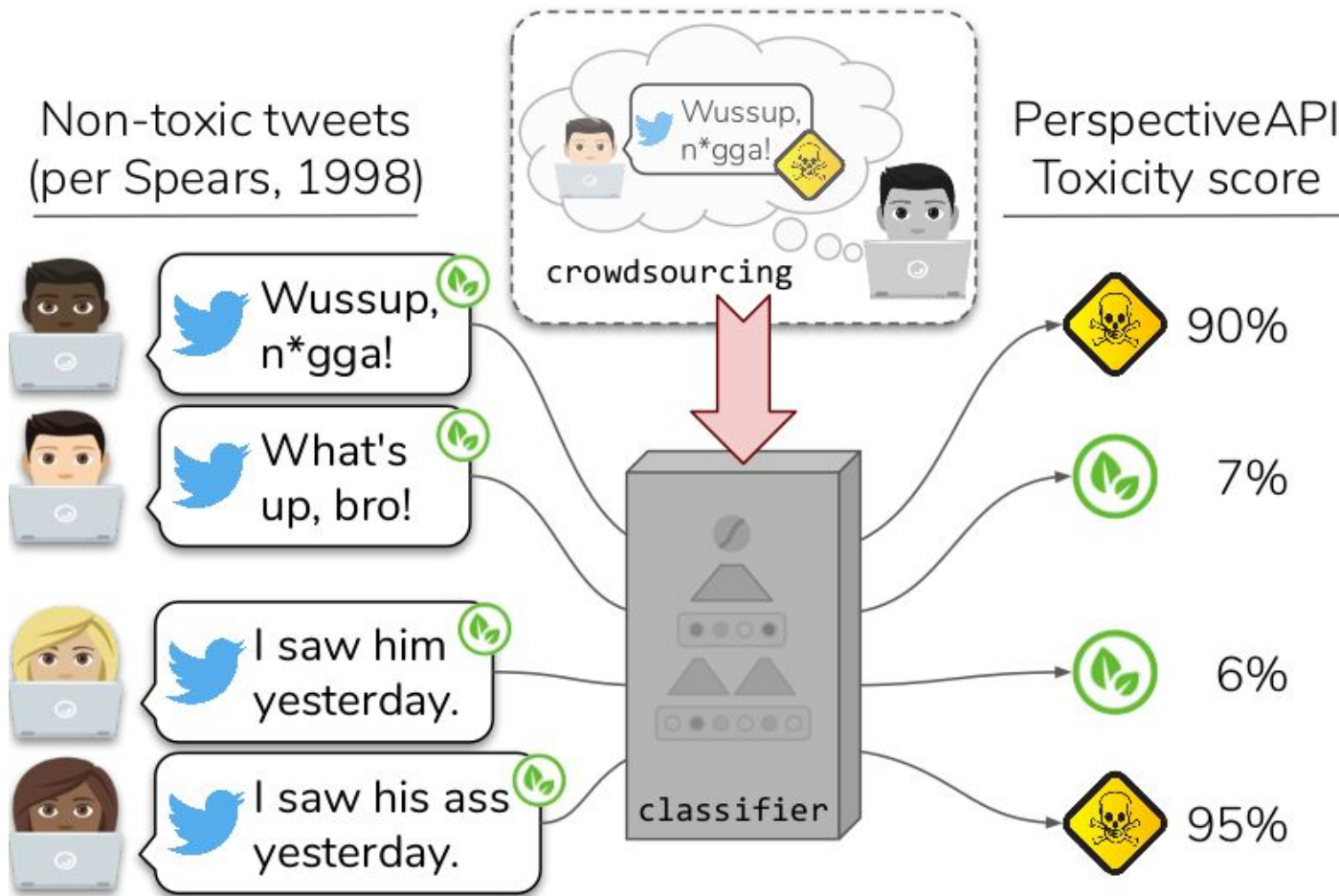
Bayesian probabilities: prior

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

Is bias always a bad thing?

$$y = ax + b$$
$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Why is Bias Problematic? (Social View)



NLP Applications

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

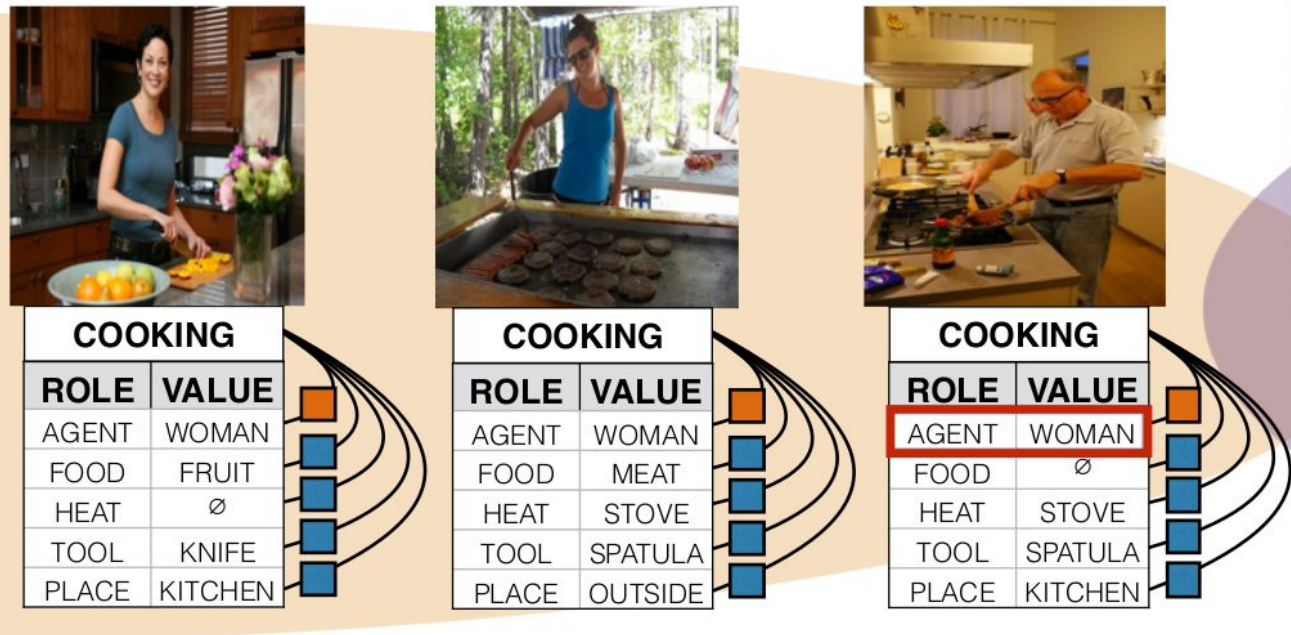
Social Stereotypes

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

Why is Bias Problematic?

See also Shah et al. (2020)

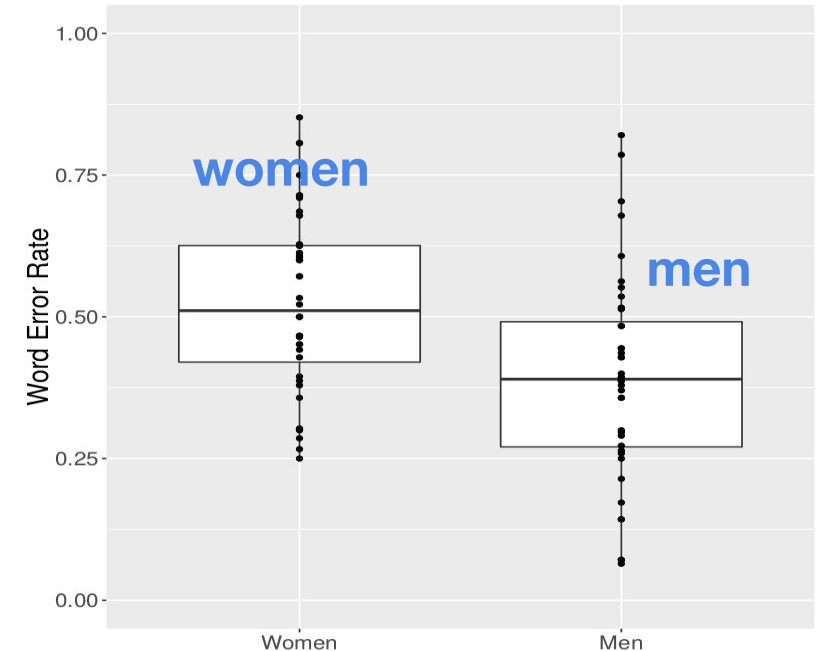
Outcome Disparity



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

Image sources: <https://www.aclweb.org/anthology/W17-1606.pdf>,
<https://www.aclweb.org/anthology/D17-1323.pdf>

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

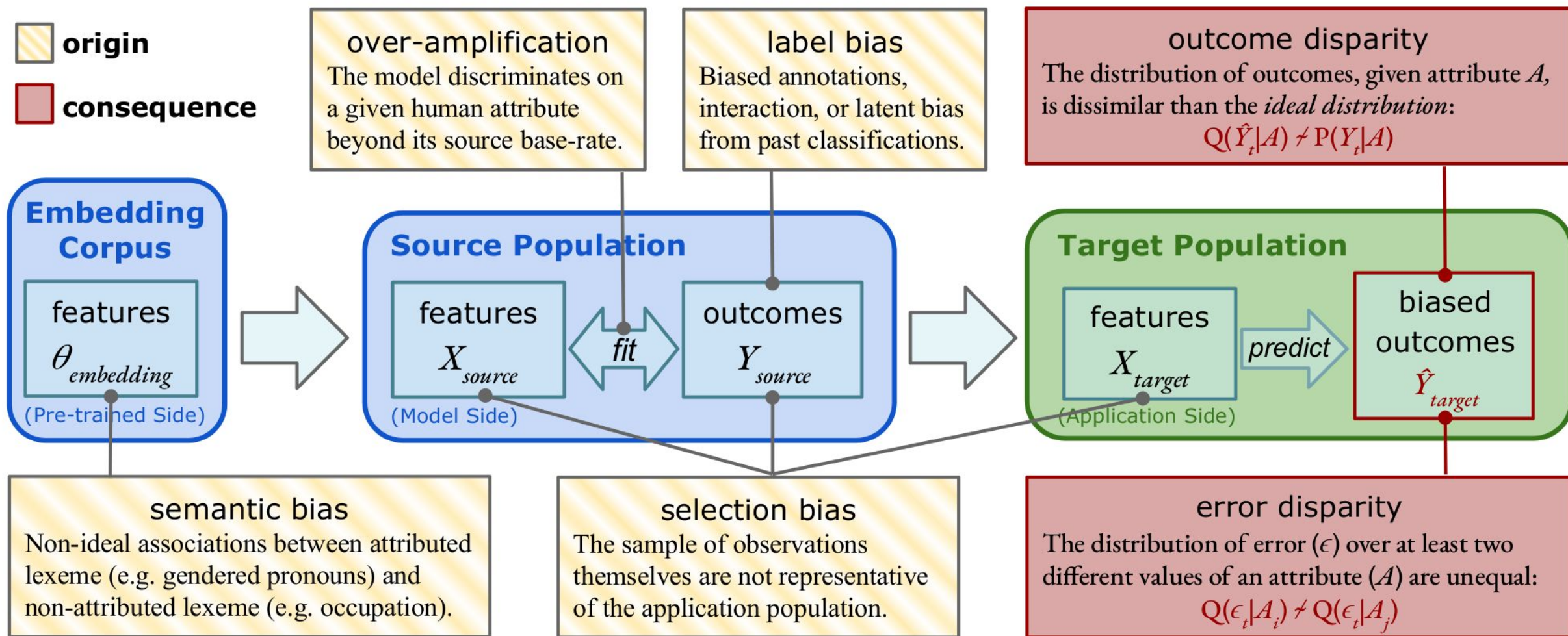
Models might **amplify bias**

51:49 distribution in a feature may lead to 100:0 decision

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

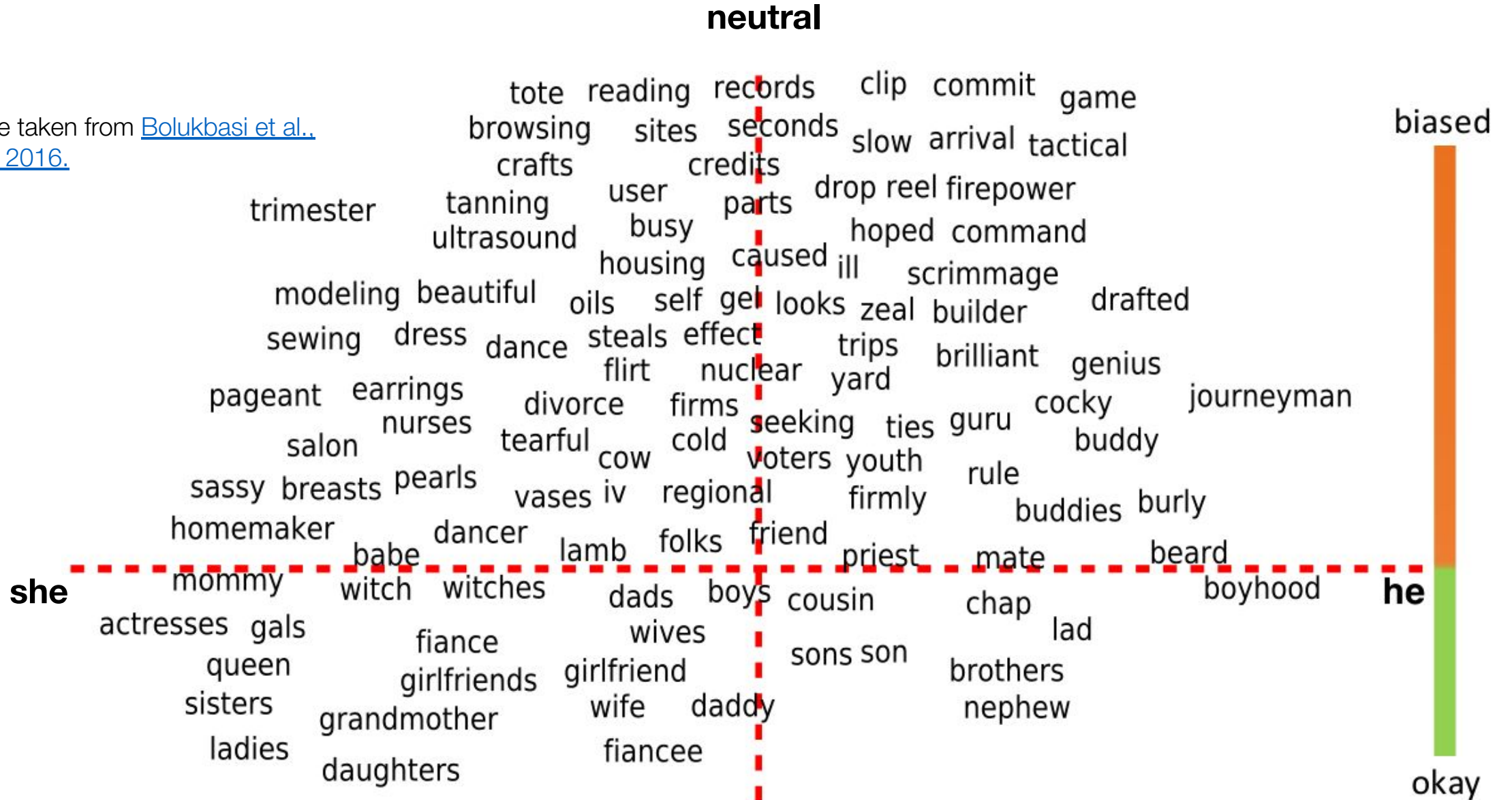
In what circumstances?

Sources of Bias in NLP (Shah et al., 2020)

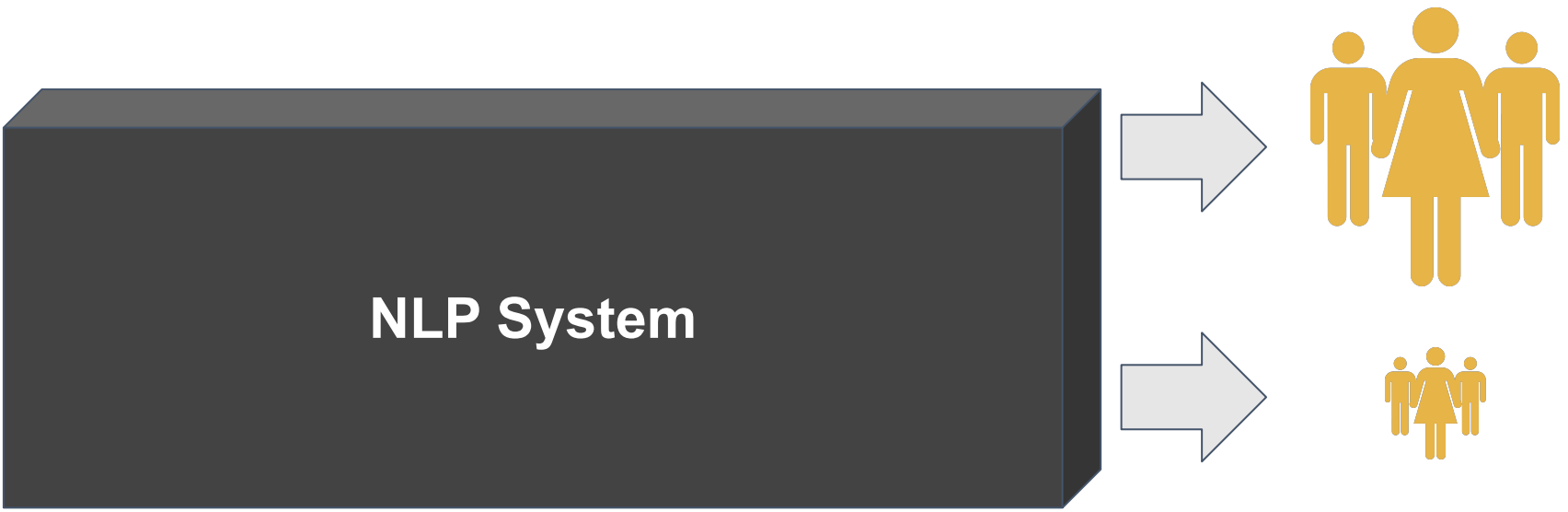


De-Biasing of Word Embeddings

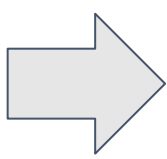
Image taken from [Bolukbasi et al., NIPS 2016.](#)



Source of Harm - Unfair Outcomes



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 ;)

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

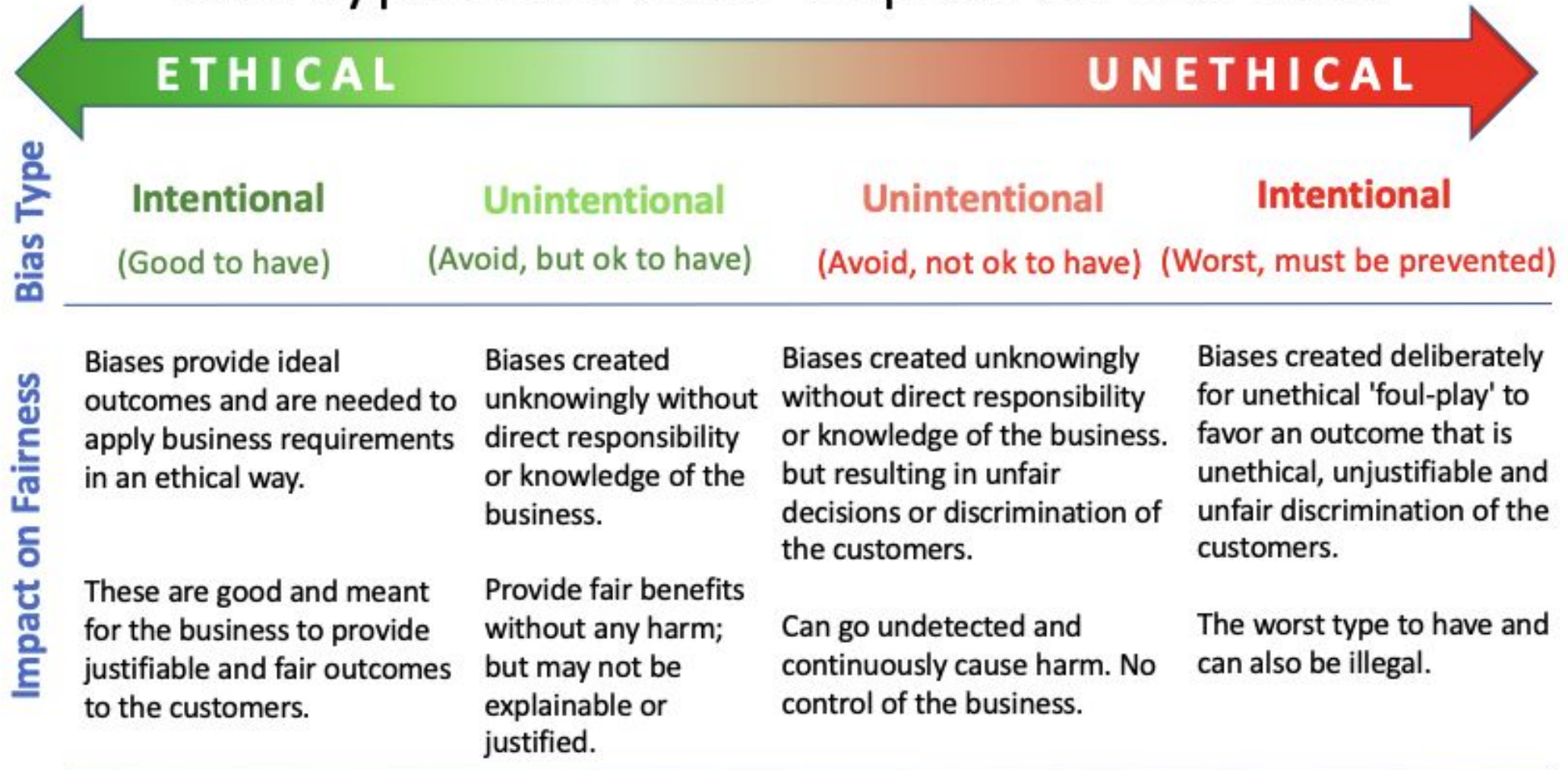
How can we measure similarity of individuals?

cannot reach group and individual fairness at the same time

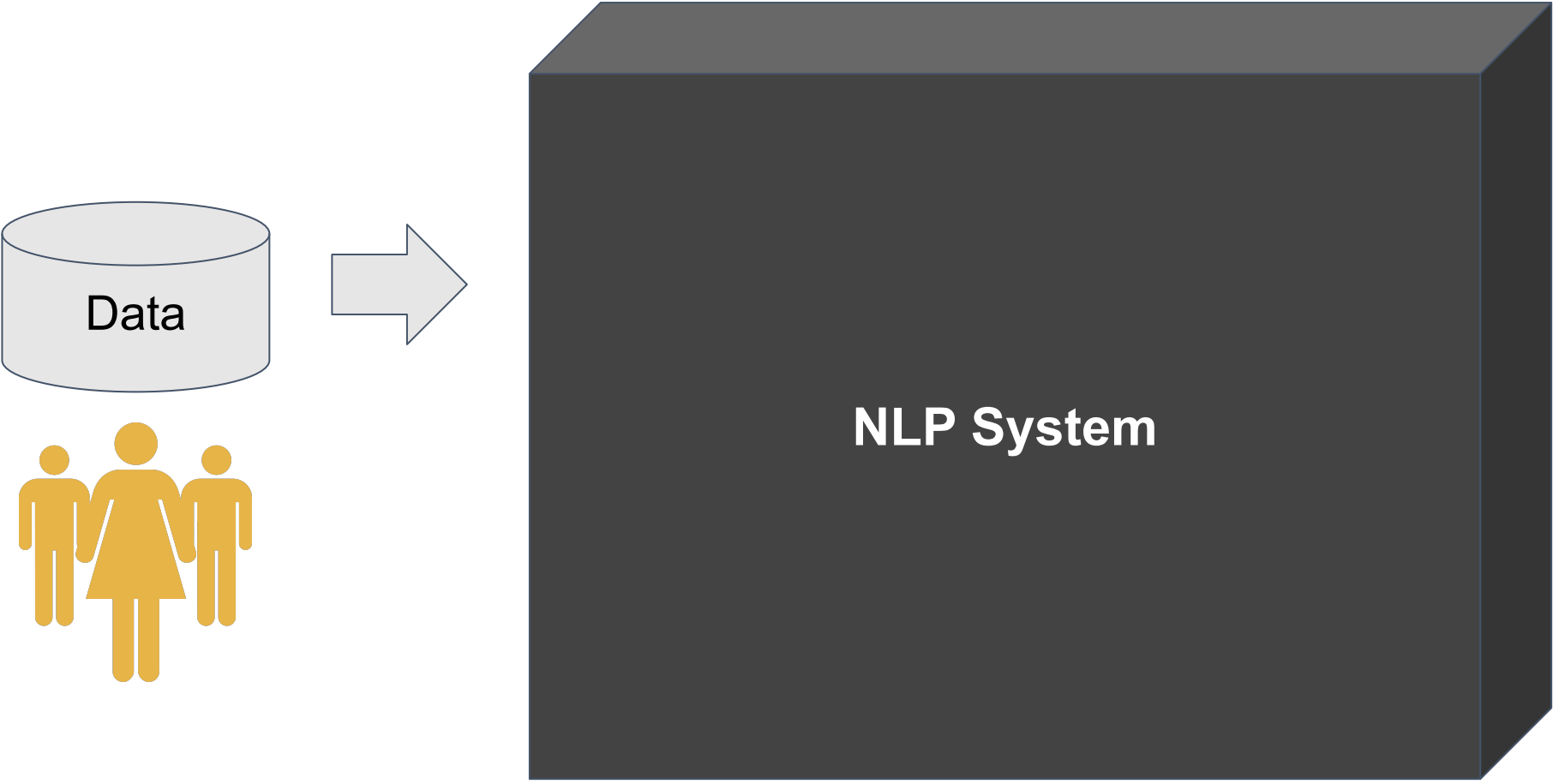
<https://dl.acm.org/doi/10.1145/3442188.3445892>

<https://arxiv.org/abs/1609.07236>

Bias types and their impact on Fairness



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)

Problems and harms related to privacy

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.”

Information Collection
<i>Surveillance</i>
<i>Interrogation</i>
Information Processing
<i>Aggregation</i>
<i>Identification</i>
<i>Insecurity</i>
<i>Secondary Use</i>
<i>Exclusion</i>
Information Dissemination
<i>Breach of Confidentiality</i>
<i>Disclosure</i>
<i>Exposure</i>
<i>Increased Accessibility</i>
<i>Blackmail</i>
<i>Appropriation</i>
<i>Distortion</i>
Invasion
<i>Intrusion</i>
<i>Decisional Interference</i>

Data Privacy Regulations

European Regulation 2016/679
General Data Protection Regulation ([GDPR](#))

Main rights of the “data subject” (natural person):

- Right of access
- Right of rectification
- Right to erasure (“right to be forgotten”)
- Right to withdraw consent at any time
- Right to lodge a complaint with a supervisory authority
- Right to restriction of processing
- Right to data portability

Applies for the data of all EU citizens - also if controller operates from a country outside the EU!



Similar laws in the US:
[California Consumer Privacy Act](#)

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.

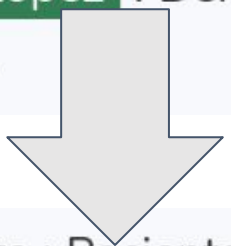
[based on: [Lawler, 2019](#)]

“Just because we can do something, doesn’t mean we should.”
Should a company sell user information to political campaigns?

Anonymization (De-Identification)

Informe clínico del paciente : Paciente **varón** de **70 años** de edad ,
minero jubilado , sin alergias medicamentosas conocidas . Operado de
una hernia el **12 de enero de 2016** en el **Hospital Costa del
Sol** por la Dra . **Juana López** . Derivado a este centro el día 16 del
mismo mes para revisión .

Category: DATE
Tagger: PHI NER



Informe clínico del paciente : Paciente **SEX** de **AGE AGE** de edad ,
PROFESSION jubilado , sin alergias medicamentosas conocidas .
Operado de una hernia el **DATE DATE DATE DATE DATE** en el
HOSPITAL HOSPITAL HOSPITAL HOSPITAL por la Dra .
DOCTOR DOCTOR . Derivado a este centro el día 16 del mismo mes
para revisión .

HitzalMed
(Lopez et al., 2020)

After having run some
anonymization system
on our data, is
everything fine?

Image Source: <https://www.aclweb.org/anthology/2020.lrec-1.870/>

Authorship Attribution / Author Profiling

Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter

Francisco Rangel^{1,2} Paolo Rosso² Martin Potthast³ Benno Stein³

Table 10. Best results per language and task.

Language	<i>Joint</i>	Gender	Variety
Arabic	0.6831	0.8031	0.8313
English	0.7429	0.8233	0.8988
Spanish	0.8036	0.8321	0.9621
Portuguese	0.8575	0.8700	0.9838

information. Therefore, the possibility of knowing social media users' traits on the basis of what they share is a field of growing interest named author profiling. To infer the authors' gender, age, native language, dialects, or personality opens a world of possibilities from the point of view of marketing, forensics, or security. For example from a security viewpoint, to be able to determine the linguistic profile of a person who writes

What are potential chances and risks of this type of technology?

Reading Assignment / Discussion

Daniel J. Solove. ['I've Got Nothing to Hide' and Other Misunderstandings of Privacy.](#)

San Diego Law Review, Vol. 44, p. 745, 2007

[Germany's complicated relationship with Google Street View.](#) NY Times, April 2013.

Questions to think about / 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?

Two ethical theories

Teleology

Telos (Greek) = goal

Outcome-oriented

Utilitarianism

“Choose that action that optimizes the outcome for the majority”

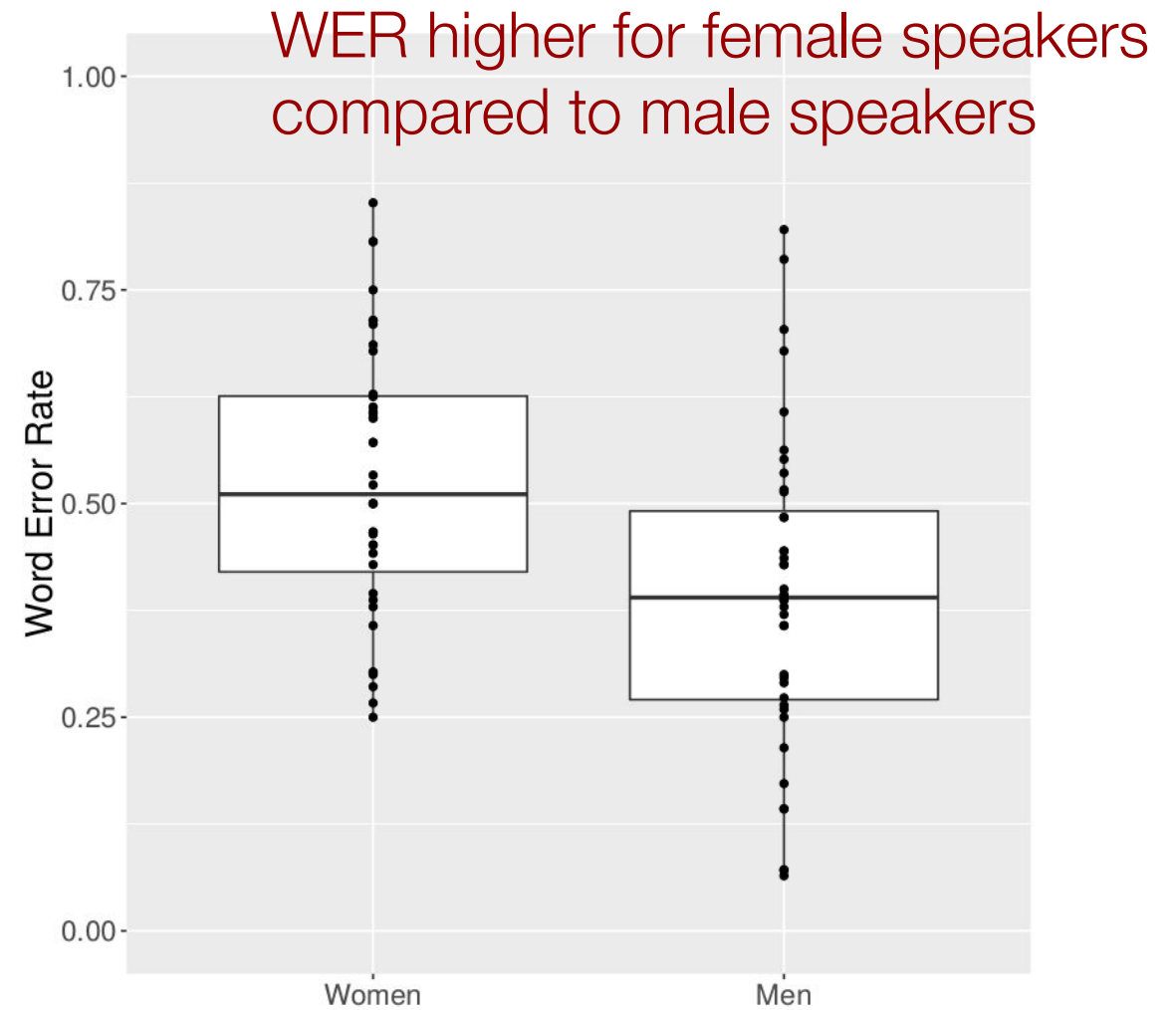
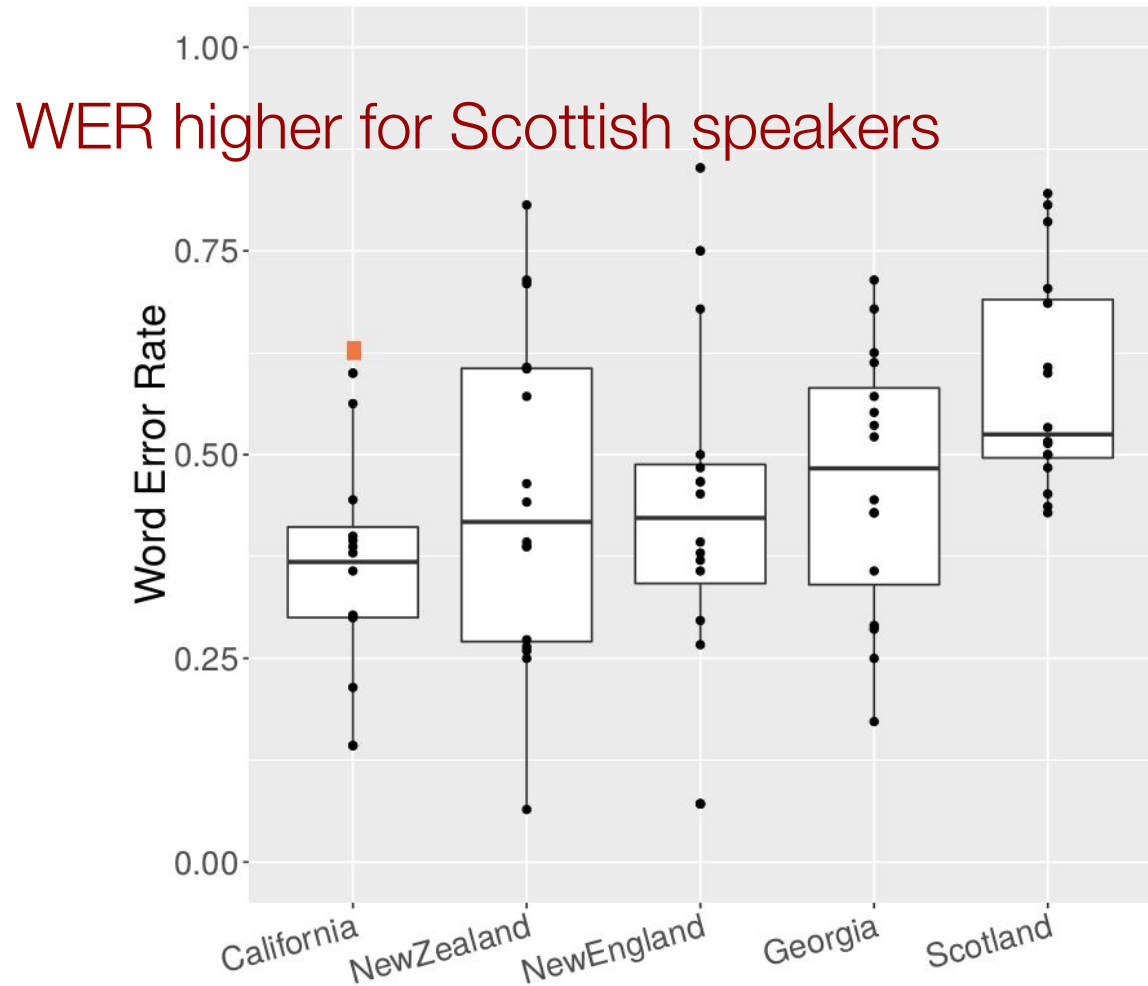
Deontology

Deon (Greek) = duty

“Identify your duty and act accordingly”

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

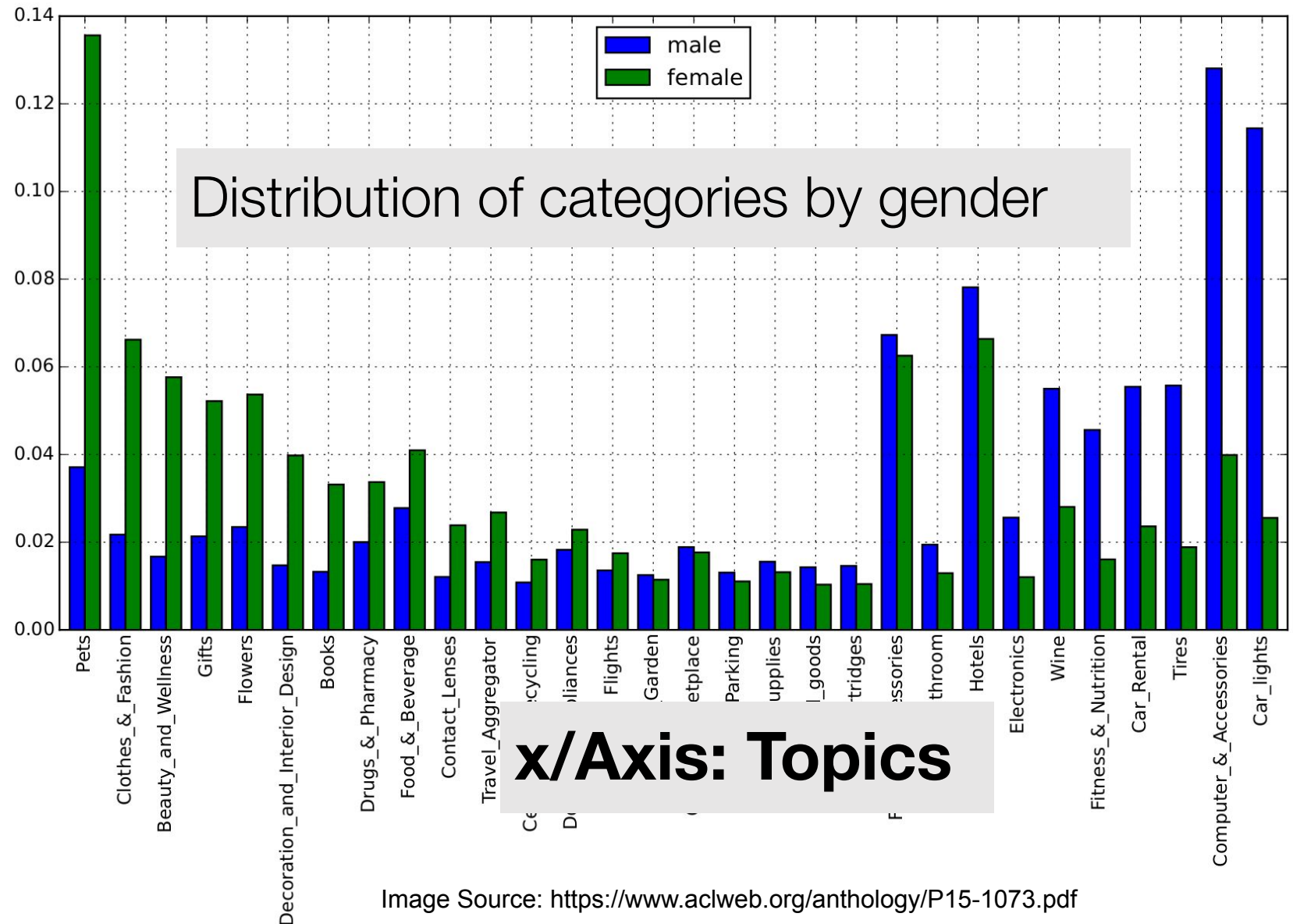
Word Error Rates for Automatic Captioning on YouTube (Tatman, 2017)



Demographic Factors Improve Classification Performance (Hovy, 2015)

Is it okay to leverage the author's gender information as explicit features for text classification?

What would be recommended from a utilitarian / generalization perspective?



Reading Assignment

Prabhumoye et al.: [Principled Frameworks for Evaluating Ethics in NLP Systems](#). In Proceedings of the Workshop on Widening NLP. 2019.

Consider a project that you are working on / have worked on or pick a recent research paper from the [ACL Anthology](#). Analyse the method / system both from a utilitarian and from a generalization perspective. How would scholars of each ethical theory evaluate the ethicality of the method/system?

“Applications”: NLP for Social Good

Civility in communication: techniques to monitor trolling, hate speech, abusive language, detect fake news, etc.



Donald J. Trump 
@realDonaldTrump Folgen


Crime in Germany is up 10% plus (officials do not want to report these crimes) since migrants were accepted. Others countries are even worse. Be smart America!

06:52 - 19. Juni 2018

19.342 Retweets 78.247 „Gefällt mir“-Angaben

27 Tsd. 19 Tsd. 78 Tsd.



Chris 
@goingglocal Folgen

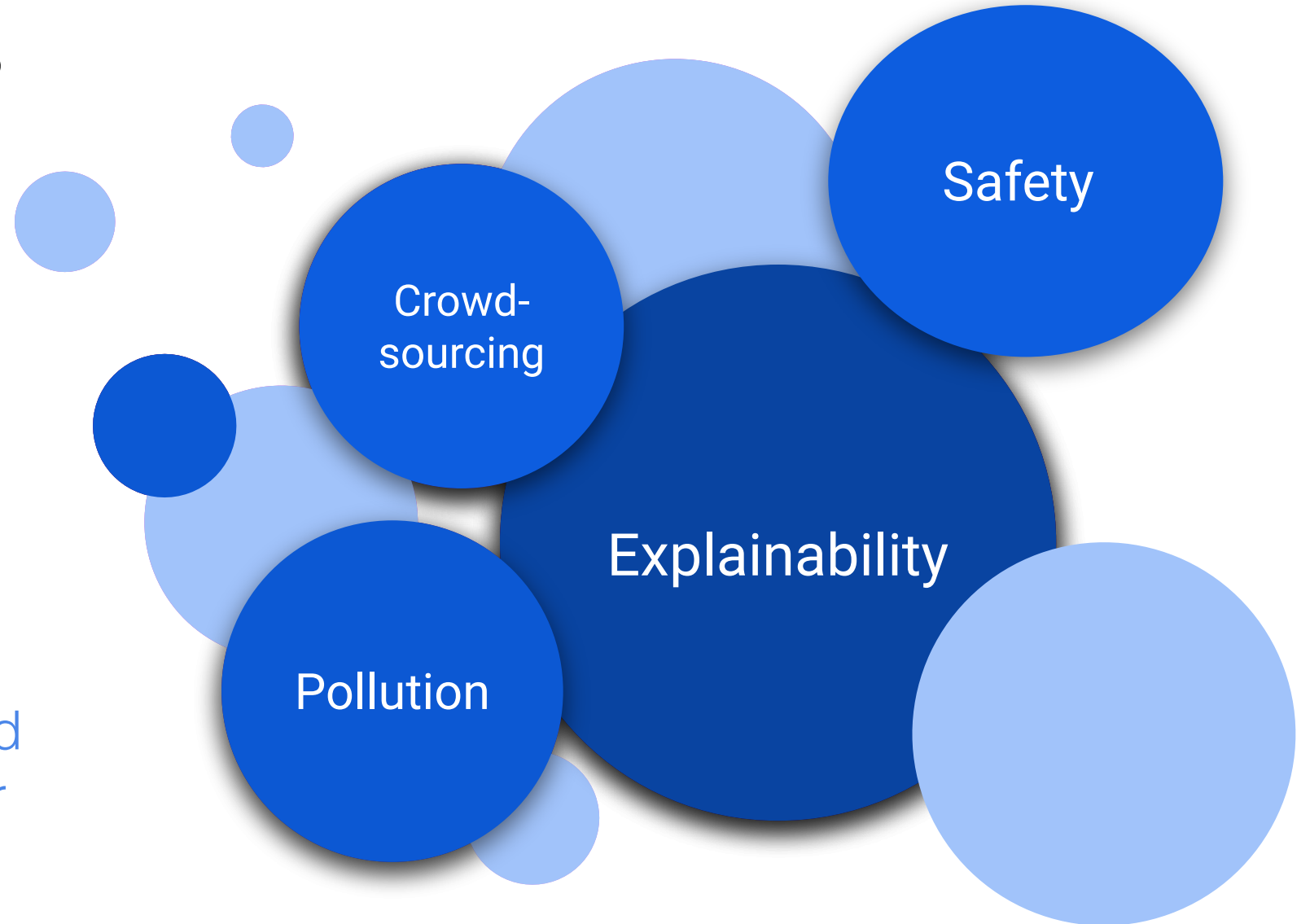
So Trump wants you to think there is a major uprising here in Germany and that "Crime in Germany is way up."

Greetings from Germany. For your information, he's a liar. And the crime rate was the **LOWEST** in 2017 in 30 YEARS. Down 10 percent from the year before!

09:55 - 18. Juni 2018

Other topics

Pick a topic of your choice and research its relationship to ethics. What are common arguments made? Do you agree? Can you find interesting examples for ethical or unethical behavior?



Reading Suggestions: Environmental Issues

Bender et al. [On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?](#). FAccT '21, March 3–10, 2021, Virtual Event, Canada

Emma Strubell, Ananya Ganesh, Andrew McCallum, [Energy and Policy Considerations for Deep Learning in NLP](#). ACL 2019.



Bild von [Peggy und Marco Lachmann-Anke](#)
auf [Pixabay](#)

Misc / Practical Hints

IRB = Institutional Review Board (Ethics Review Board)

Reviews all human experimentation

[Find out how to contact the IRB of your institution.](#)

The ACL has adopted [ACM Code of Ethics and Professional Conduct](#) and published an FAQ with hints on conducting research and publishing in an ethical manner (see, e.g., [ACL-IJCNLP 2021 Ethics FAQ](#)).



Association for Computational Linguistics

Practical summary

Analyse data, task and outcomes for potential harm.

Can benefits outweigh harms?

Retrospection

What have you learned?

What does that mean for you personally?

What was surprising?

References

[CMU Course on Computational Ethics for NLP](#)

[Stanford Course on Ethics in NLP](#)

Literature – Ethics in NLP

Overviews

- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use Ai in a responsible way*. Cham, Switzerland: Springer.
- Fort, K., & Couillault, A. (2016). Yes, We Care! Results of the Ethics and Natural Language Processing Surveys. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Retrieved from <https://www.aclweb.org/anthology/L16-1252>
- Hovy, D., & Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In K. Erk & N. A. Smith (Chairs), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Retrieved from <https://www.aclweb.org/anthology/P16-2096.pdf>

Literature – Ethics in NLP

Overviews

- Leidner, J. L., & Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, & H. Wallach (Chairs), Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Retrieved from <https://www.aclweb.org/anthology/W17-1604.pdf>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. Retrieved from <https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679>
- Zweig, K. A. (2019). *Ein Algorithmus hat kein Taktgefühl: wo Künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. München: Heyne.

Literature – Ethics in NLP

Bias

- Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, & H. Wallach (Chairs), Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Retrieved from <https://www.aclweb.org/anthology/W17-1606.pdf>
- Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2019). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 14(1). Retrieved from <https://arxiv.org/pdf/1809.02208.pdf>
- Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1164.pdf>

Literature – Ethics in NLP

Bias

- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online (pp. 25–35). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3504.pdf>
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1668–1678). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1163.pdf>

Literature – Ethics in NLP

Bias

- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems. Retrieved from <http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* (New York, N.Y.), 356(6334), 183–186. Retrieved from <https://arxiv.org/pdf/1608.07187.pdf>
- Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3405–3410). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/D19-1339.pdf>
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2979–2989). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/D17-1323.pdf>

Literature – Ethics in NLP

Fairness

- Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–10). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-4401.pdf>

Literature – Ethics in NLP

Gender Stereotypes

- Bhaskaran, J., & Bhallamudi, I. (2019). Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing (pp. 62–68). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3809.pdf>