Natural Language Processing & Ethics

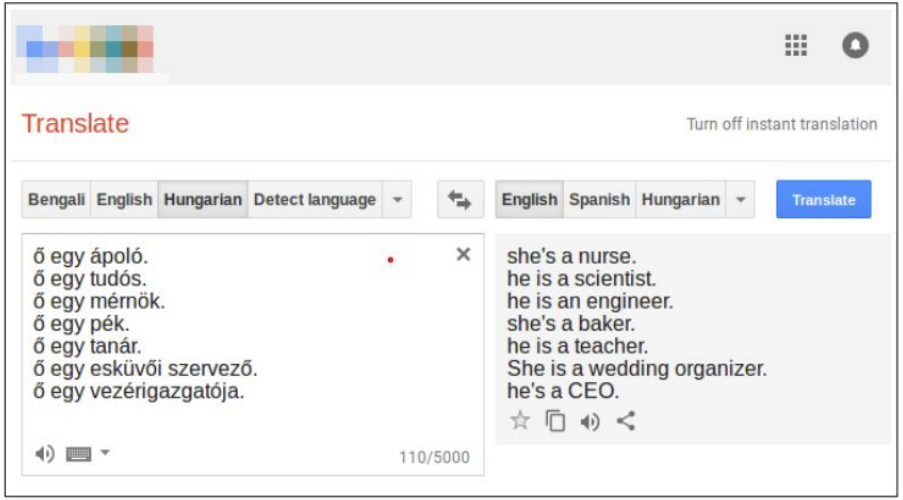
27/4/2021

NLP and Ethics: Where?

- **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.

Credits: Computational Ethics for NLP (cmu.edu)

Bias in Machine Translations



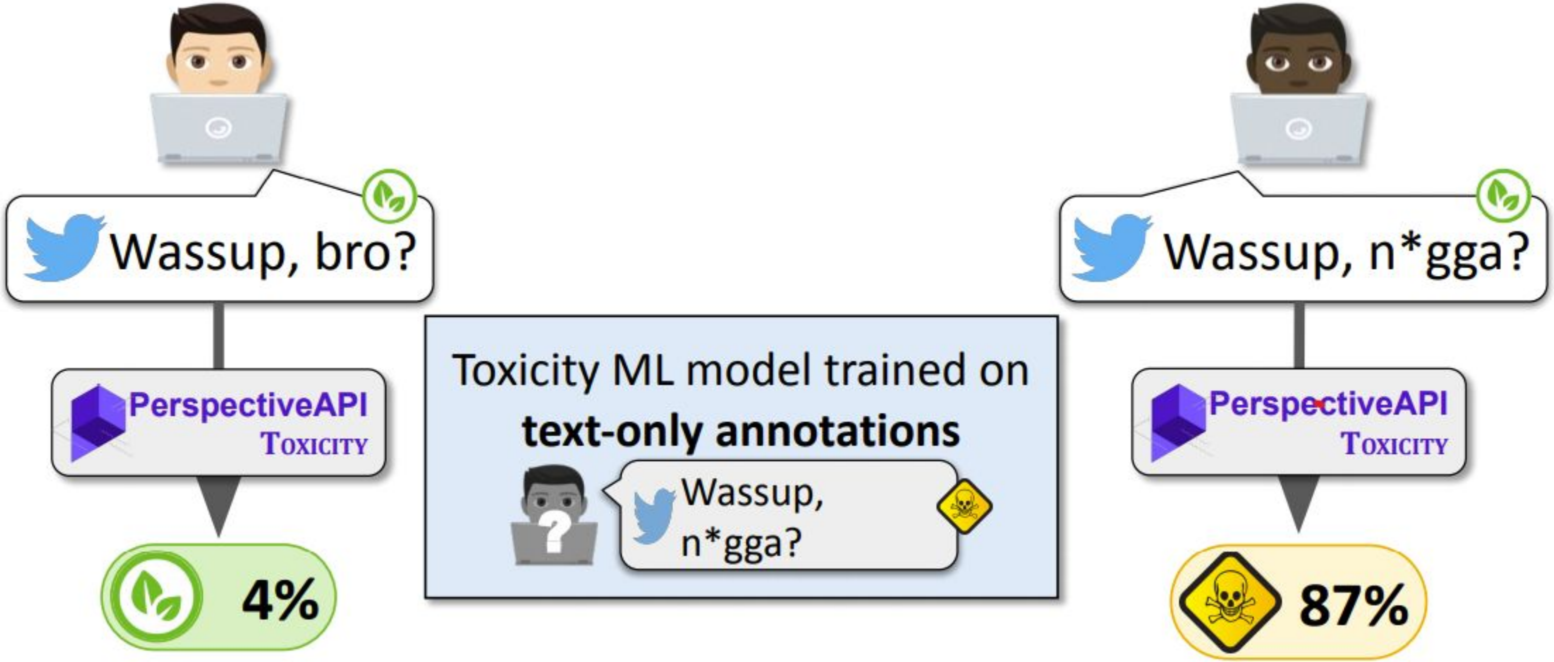
ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

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.

Figure 1: Translating sentences from a gender neutral language such as Hungarian to English provides a glimpse into the phenomenon of gender bias in machine translation. This screenshot from Google Translate shows how occupations from traditionally male-dominated fields [40] such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female.

[1809.02208.pdf (arxiv.org)](https://arxiv.org/pdf/1809.02208.pdf)

Problem: severe racial bias in hate speech detection



Wassup, bro?

PerspectiveAPI
Toxicity

4%

Toxicity ML model trained on
text-only annotations

Wassup,
n*gga?

Wassup, n*gga?

PerspectiveAPI
Toxicity

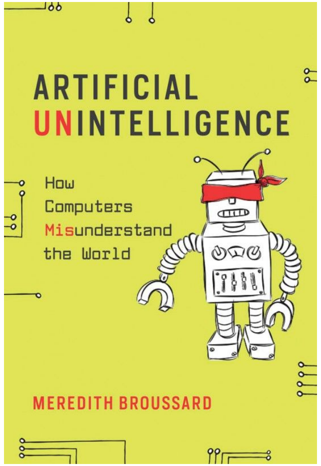
87%

*Slides by Maarten Sap
(Sap et al. ACL'19)

Examples of inoffensive statements from Spears (1998)

Important questions

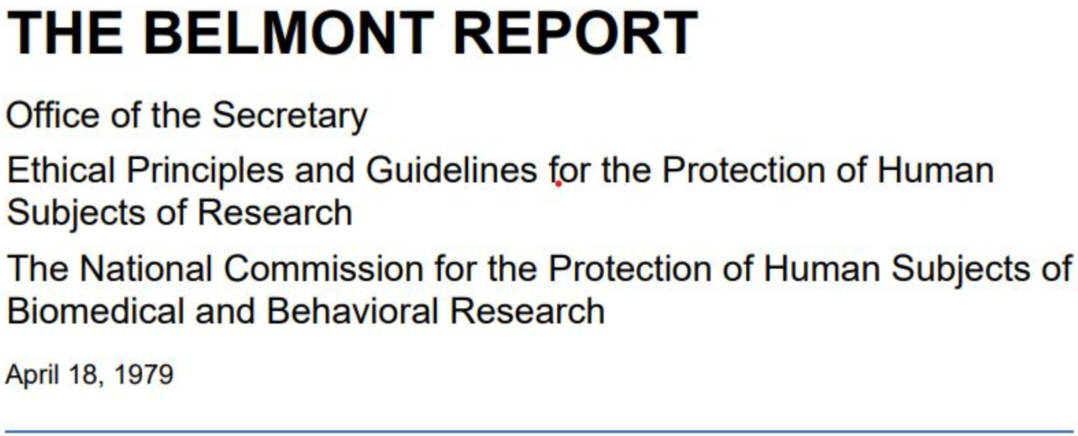
- **Who** could **benefit** from the ML technology?
- **Who** can **be harmed** by the ML technology?
- **Who** is **responsible for** the ML technology?


ARTIFICIAL UNINTELLIGENCE
How Computers Misunderstand the World
MEREDITH BROUSSARD

*”If we understand the limits of what we **can do** with technology, Broussard tells us, we can make better choices about what we **should do** with it to make the world better for everyone.”*

Artificial Unintelligence | The MIT Press

The Belmont Report



THE BELMONT REPORT

Office of the Secretary

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

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

April 18, 1979

“Two general rules have been formulated as complementary expressions of beneficent actions in this sense: ***(1) do not harm*** *and* ***(2) maximize possible benefits and minimize possible harms.”***

The Belmont Report (hhs.gov)

Research and Privacy

When we download data from Socials (i.e twits, posts, photos):

Did these people agree to participate in the study?

Legal ≠ Ethical

Public ≠ Publicized

Who is responsible?

When a failure in the AI technology is spotted who is responsible for it?

- The person who uses the technology?
- The researcher/developer?
- Paper reviewers?
- University?
- Society as a whole?

Food for Thought

data → **BAISED DATA**

method → **BLACK BOX**

evaluation → **ACCURACY**

Input
BLACK BOX
Output

Biased data

Suppose a social study using material from dating sites...

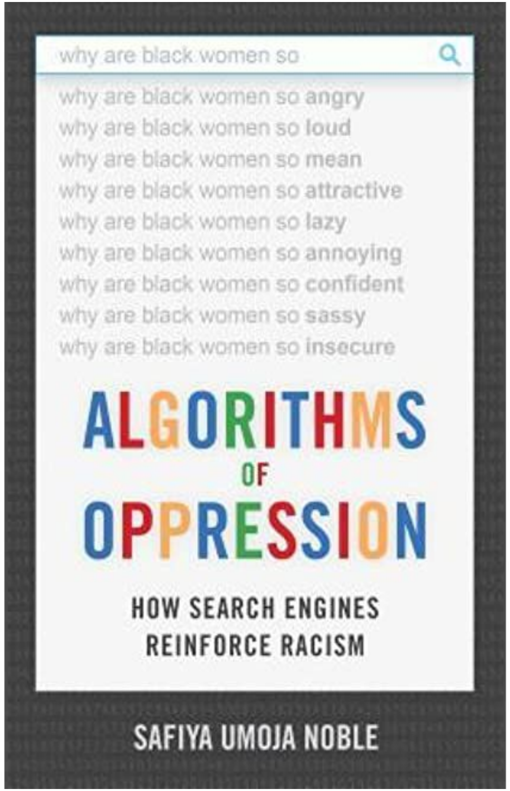
Only white people, who self-disclose their orientation,
certain social groups, certain age groups, certain time range/fashion;
the photos were carefully selected by subjects to be attractive so there is even self-selection bias…

The dataset is balanced, which does not represent true class distribution.

"AI models are only as good as the data used to train them, and developing a representative, effective training data set is very challenging"

It's time to start breaking open the black box of AI - Watson Blog (ibm.com)

Biased data


why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE

“Run a Google search for “black girls”—what will you find? “Big Booty” and other sexually explicit terms are likely to come up as top search terms. But, if you type in “white girls,” the results are radically different. The suggested porn sites and un-moderated discussions about “why black women are so sassy” or “why black women are so angry” presents a disturbing portrait of black womanhood in modern society.

Through an analysis of textual and media searches as well as extensive research on paid online advertising, Noble exposes a culture of racism and sexism in the way discoverability is created online.”

Algorithms of Oppression (nyupress.org)

Black box: TRUSTED AI?

“Even if biases are identified during training, the model may still exhibit bias in runtime. This can result from incongruities in optimization caused by assignment of different weights to different features”

It’s time to start breaking open the black box of AI - Watson Blog (ibm.com)

Can we use not interpretable models when we make predictions about sensitive attributes?

Which is the cost of MISCLASSIFICATION?

evaluation → **ACCURACY**