

CS447: Natural Language Processing

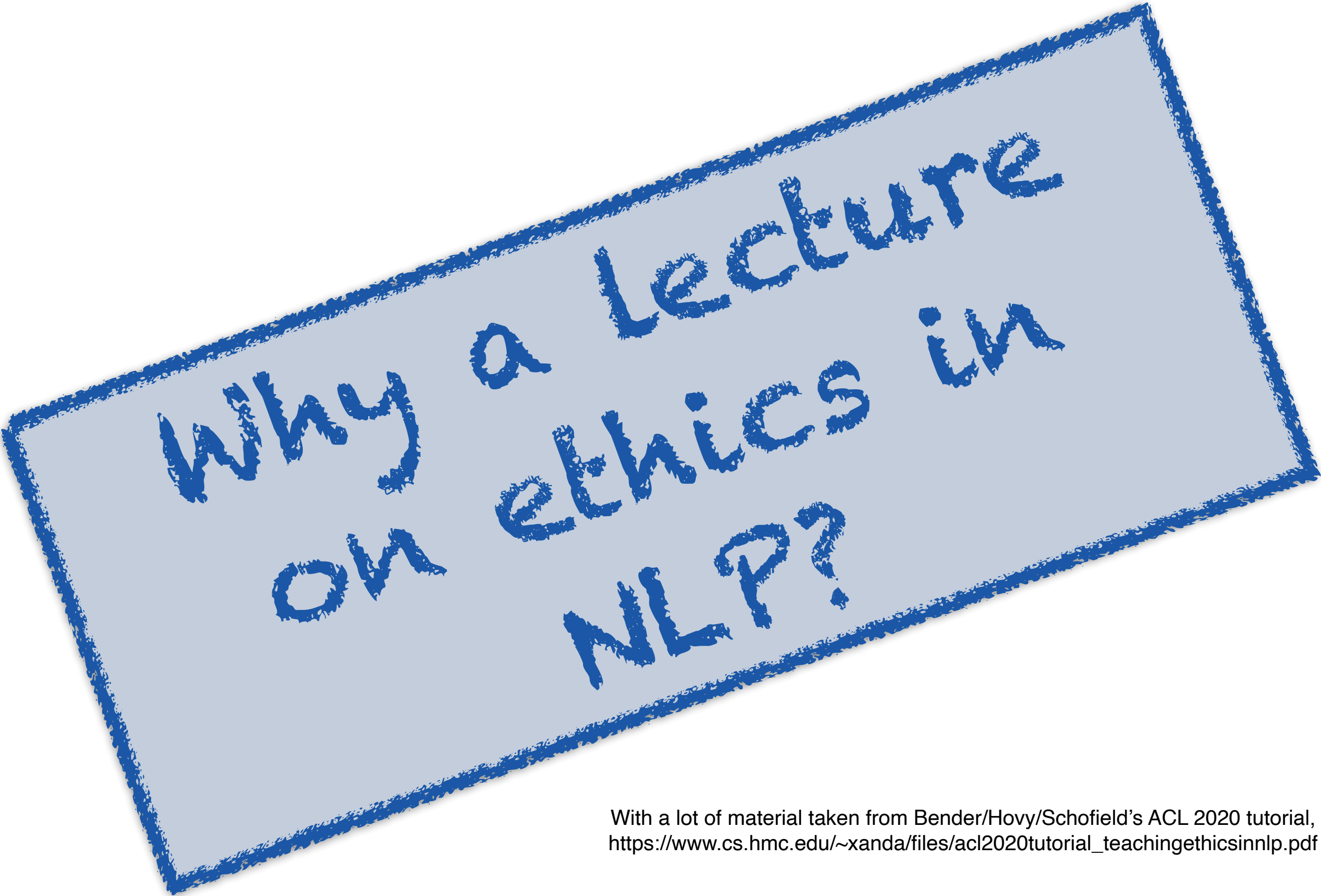
<http://courses.engr.illinois.edu/cs447>

Lecture 28: NLP and Ethics

Julia Hockenmaier

juliahmr@illinois.edu

3324 Siebel Center



Why a lecture
on ethics in
NLP?

With a lot of material taken from Bender/Hovy/Schofield's ACL 2020 tutorial,
https://www.cs.hmc.edu/~xanda/files/acl2020tutorial_teachingethicsinnlp.pdf

Ethics and NLP

NLP touches on ethics in numerous ways:

- The **data** we work with, natural language, is **produced by (and may talk about) real people.**

Do we respect the **rights of the individuals** that produced or are mentioned in the data? (privacy, anonymity, etc.)

Do we understand how the **population of individuals** that produced our data differs from the general population?

Do we understand what **biases** are inherent in this data?

- The **applications** we develop have more and more **real-world uses** and (unintended) **consequences.**

We need to be aware of the potential for **benefit** and **abuse**

ACM Code of Ethics

Computing professionals should...

... **contribute to society and to human well-being**,
acknowledging that **all people are stakeholders** in computing

This includes obligations to promote **fundamental human rights**,
and to **minimize negative consequences** of computing; and to strive for
environmental sustainability.

... **avoid harm**

Harm includes **unjustified disclosure of information**

... **take action not to discriminate**

Technologies should be **inclusive** and **accessible**; the creation of
technologies that **disenfranchise** or **oppress** people should be **avoided**

... **respect privacy**

Collection and use of private data comes with responsibilities;

... **maintain high standards of professional competence**,
conduct and **ethical practice**

This includes **awareness of the social context in which work will be used**

Normative vs. descriptive ethics

Normative ethics: what we want the world to be

Descriptive ethics: what the world is like.

Example: Gender bias in NLP:

A coreference system that cannot attach female pronouns to the word “doctor” is both *normatively* and *descriptively* wrong.

Racially or gender-based word embeddings are *normatively* wrong (if we don’t want them to be biased), but might be *descriptively* correct (since they reflect how societies talk about race/gender)

https://www.cs.hmc.edu/~xanda/files/acl2020tutorial_teachingethicsinnlp.pdf



Bias and Fairness in NLP



Social Bias in NLP

“Bias” has a number of technical senses in machine learning/stats:

(biased coins, inductive bias, or the bias—variance tradeoff)

This needs to be distinguished from **social bias (e.g. gender/racial/class bias, ...)** that a system’s behavior may exhibit.

Social bias results in... [Barocas et al, Crawford 2017]

... **Allocational harms**: a system **allocates resources**/opportunities (credit scores, job ads, goods to buy) differently to different social groups

... **Representational harms**: a system **represents different social groups** in a less positive light than others

Identifying social bias is inherently **normative**

Bias in NLP is a “hot” topic, but a lot of NLP work on bias does not engage deeply enough with the relevant social science literature, or with the communities affected by this bias.

S.L. Blodgett et al. Language (Technology) is Power: A critical survey of “Bias” in NLP

<https://arxiv.org/pdf/2005.14050.pdf>



Descriptive bias

Garg et al, PNAS April 17 2018 *Word embeddings quantify 100 years of gender and ethnic stereotypes*
<https://www.pnas.org/content/115/16/E3635>

Measure the strength of association between words representing social groups (women/men, Asians/Caucasians/...) and words representing professions, attributes, etc.

Embeddings reflect real differences (few carpenters are female, many nurses are), but also track gender and ethnic stereotypes (women are “charming”/“maternal”/...), and their changes over time

Normative bias: NLP performance on AAE

(discussion from <https://arxiv.org/pdf/2005.14050.pdf>)

Many NLP tools have poor accuracy on “non-standard” varieties of English that differ from the varieties in common corpora.

For example, toxicity detectors are less accurate on tweets written in African-American English (AAE).

If AAE tweets are deemed more offensive...

- ... AAE speakers might be more likely to be blocked

- ... AAE speakers might feel the need to communicate differently than how they normally would (or not use social media)

- ... this stigmatization may exacerbate existing discrimination



NLP and Endangered languages

Steven Bird: Decolonising Speech and Language Technologies
<https://www.aclweb.org/anthology/2020.coling-main.313.pdf>

Speech/NLP has been used to automate language documentation for endangered (indigenous) languages.

But...

... there is little evidence that documentation saves dying languages
... documentation and the NLP technology are developed by outsiders who don't engage with the language communities ('colonizers'), and who don't understand how language is used in the community, or what tools would be of use to the community.



Socially useful NLP applications

Assistive technology (text-to-speech, voice search, image description for the blind) helps people with disabilities

Machine translation, summarization, better search engines all provide unprecedented access to information to the general public

Identifying fake news, trolls, toxic comments can prevent harmful information to spread.

Social media monitoring can also be used to assist in disasters, or to identify health issues

But this can also be abused for surveillance.