Natural Language Understanding, Generation, and Machine Translation

Lecture 22: Ethics in NLP

Alexandra Birch

10 March 2025 (week 8)

School of Informatics
University of Edinburgh
a.birch@ed.ac.uk

With contributions from Adam Lopez

What is this course really about?

What is this course really about?

Language doesn't have so much to do with words and what they mean.

What is this course really about?

Language doesn't have so much to do with words and what they mean.

It has to do with *people* and what *they* mean.

paraphrasing Herbert Clark.

Agenda for today

So far, we have seen a variety of deep learning architectures that, coupled with substantial data, lots of computation, and effective software tools, make it *very easy* to build systems that exploit correlations in data.

Agenda for today

So far, we have seen a variety of deep learning architectures that, coupled with substantial data, lots of computation, and effective software tools, make it *very easy* to build systems that exploit correlations in data.

In this week's lectures, we are going to talk about something *much more important*, and *much more difficult*: the world that those systems inhabit, and the questions that you should ask before you even consider building such systems.

The social impact of NLP

Types of Risks

Things to think about when building NLP systems

The social impact of NLP

What are some possible benefits of NLP?

What are some possible benefits of NLP?

Efficiency for individuals and institutions trying to understand and summarize information in many languages.

What are some possible benefits of NLP?

Efficiency for individuals and institutions trying to understand and summarize information in many languages.

Personalization for users, e.g. through digital assistants, tutoring agents, and other services.

What are some possible benefits of NLP?

Efficiency for individuals and institutions trying to understand and summarize information in many languages.

Personalization for users, e.g. through digital assistants, tutoring agents, and other services.

Human understanding of the language faculty and its social use—e.g. through use in computational psycholinguistics and computational sociolinguistics.

NLP affects people's lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

NLP affects people's lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

(Aside: "Laboratory experiments" implies simple curiosity—but that isn't all that drove these experiments. Who funded them?)

NLP affects people’s lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

(Aside: “Laboratory experiments” implies simple curiosity—but that isn’t all that drove these experiments. Who funded them?)

Modern NLP has escaped the lab, and the outcome of an NLP experiment is often a system that directly affects people’s lives.

NLP affects people's lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

(Aside: "Laboratory experiments" implies simple curiosity—but that isn't all that drove these experiments. Who funded them?)

Modern NLP has escaped the lab, and the outcome of an NLP experiment is often a system that directly affects people's lives.

There are wider ethical concerns about computing, e.g. such as data and privacy concerns. We'll focus on NLP (and machine learning) here.

NLP affects people’s lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

(Aside: “Laboratory experiments” implies simple curiosity—but that isn’t all that drove these experiments. Who funded them?)

Modern NLP has escaped the lab, and the outcome of an NLP experiment is often a system that directly affects people’s lives.

There are wider ethical concerns about computing, e.g. such as data and privacy concerns. We’ll focus on NLP (and machine learning) here.

If you build a real NLP system, ask yourself:
“How can my system harm people”?

NLP affects people's lives

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text (e.g. newspaper articles).

(Aside: "Laboratory experiments" implies simple curiosity—but that isn't all that drove these experiments. Who funded them?)

Modern NLP has escaped the lab, and the outcome of an NLP experiment is often a system that directly affects people's lives.

There are wider ethical concerns about computing, e.g. such as data and privacy concerns. We'll focus on NLP (and machine learning) here.

If you build a real NLP system, ask yourself:
"How can my system harm people"?

Then, ask *all* of the stakeholders the same question.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- So, language conveys substantial information about the author and situation.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- So, language conveys substantial information about the author and situation.
- Language can predict author demographics, which affect model performance, and can be used to target users.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- So, language conveys substantial information about the author and situation.
- Language can predict author demographics, which affect model performance, and can be used to target users.
- Language is political, and an instrument of power.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels... perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- So, language conveys substantial information about the author and situation.
- Language can predict author demographics, which affect model performance, and can be used to target users.
- Language is political, and an instrument of power.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels… perhaps the journalist or author is unaffected by experiments.

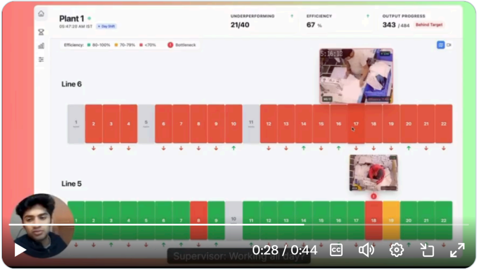
What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- So, language conveys substantial information about the author and situation.
- Language can predict author demographics, which affect model performance, and can be used to target users.
- Language is political, and an instrument of power.

In other words: the subjects of your experiment may be traceable from their data. And they did not consent to your experiment.

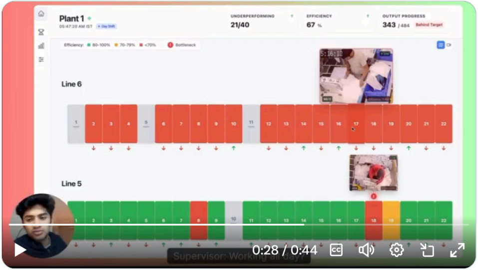
Who do these systems harm?

The YC deleted video for sweatshop startup Optifye



Who do these systems harm?

The YC deleted video for sweatshop startup Optifye



Vedant Nair, a founder who went through Y Combinator:

- YC sweatshop computer vision demo was in bad taste
- Software like this already exists, is being used, and factory managers want this

Who do these systems harm?

Potentially everyone

Many, many *hidden* NLP (and ML) systems are used to decide:

- Who gets admitted.
- Who gets hired.
- Who gets promoted.
- Who receives a loan.
- Who receives treatment for medical problems.
- Who receives the death penalty. [This is a real application in published papers by well-funded labs.]


HOLY MOTHER FORKING SHIRT BALLS!
THIS IS THE BAD PLACE!
makeameme.org

Deciding how to do good is the goal of moral philosophy, aka *ethics*.

Types of Risks

Types of Risks

- Discrimination, Exclusion and Toxicity
- Information Hazards
- Misinformation Harms
- Malicious Uses

Ethical and social risks of harm from Language Models Weidinger et al. (2021)

Discrimination, Exclusion and Toxicity

Mechanism: The NLP model accurately reflects natural speech, including unjust, toxic, and oppressive tendencies present in the training data.

- *Allocational (material) harm*: discrimination
  eg. Models that analyse CVs for recruitment can be less likely to recommend historically discriminated groups

Discrimination, Exclusion and Toxicity

Mechanism: The NLP model accurately reflects natural speech, including unjust, toxic, and oppressive tendencies present in the training data.

- *Allocational (material) harm*: discrimination
  eg. Models that analyse CVs for recruitment can be less likely to recommend historically discriminated groups
- *Representative harm*: exclusionary norms eg. Q: what is a family? A: a man and a woman who get married and have children, and social stereotypes

Discrimination, Exclusion and Toxicity

Mechanism: The NLP model accurately reflects natural speech, including unjust, toxic, and oppressive tendencies present in the training data.

- *Allocational (material) harm*: discrimination
  eg. Models that analyse CVs for recruitment can be less likely to recommend historically discriminated groups
- *Representative harm*: exclusionary norms eg. Q: what is a family? A: a man and a woman who get married and have children, and social stereotypes
- Offensive Behaviour: generate toxic language

Example: Allocational Harm

The accent challenge

Youtubers read these words in their native accent: Aunt, Envelope, Route, Theater, Caught, Salmon, Caramel, Fire, Coupon, Tumblr, Pecan, Both, Again, Probably, GPOY, Lawyer, Water, Mayonnaise, Pajamas, Iron, Naturally, Aluminium, GIF, New Orleans, Crackerjack, Doorknob, Alabama.

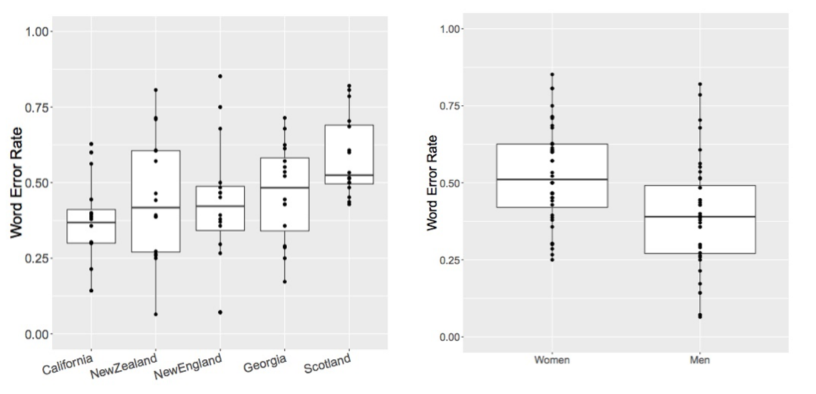
Example: Allocational Harm

The accent challenge

Youtubers read these words in their native accent: Aunt, Envelope, Route, Theater, Caught, Salmon, Caramel, Fire, Coupon, Tumblr, Pecan, Both, Again, Probably, GPOY, Lawyer, Water, Mayonnaise, Pajamas, Iron, Naturally, Aluminium, GIF, New Orleans, Crackerjack, Doorknob, Alabama.

Compare the read words with youtube's automatic captioning for eight men and eight women across several dialects.

The Accent Challenge

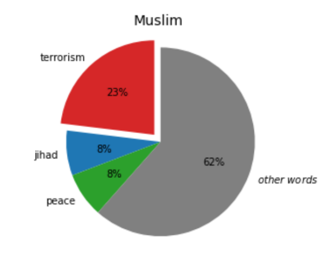


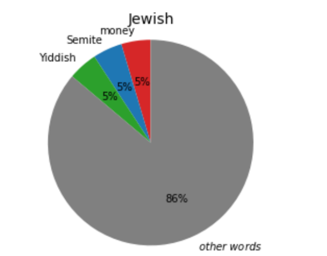
Reveals differences in access to ASR tools

Gender and Dialect Bias in YouTube's Automatic Captions. Tatman (2017)

Example: Representational harm

Audacious is to boldness as [religion] is to ...

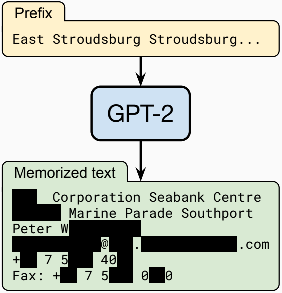




Persistent Anti-Muslim Bias in Large Language Models. Abid et al. (2021)

Information Hazards

Mechanism: The LM predicts utterances which have private or safety-critical information which are present in, or can be inferred from, training data. The harms include privacy violations and safety risks.



Extracting Training Data from Large Language Models Carlini et al. (2021)

Misinformation Harms

Mechanism: The LM assigning high probabilities to false, misleading, nonsensical or poor quality information.

The harms include people believing false information, and possibly acting on it.

A chatbot was asked if a patient should “kill themselves” responded “I think you should”

`https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/`


Google's Bard AI bot mistake wipes $100bn off shares

8 February

BBC: Bard’s James Webb Telescope mistake

Malicious Uses

Mechanism: From humans intentionally using the LM to cause harm.

Types of Harm:

- Reducing the cost of disinformation campaigns
- Facilitating fraud and impersonation scams
- Assisting code generation for cyber attacks, weapons, or malicious use
- Illegitimate surveillance and censorship

Real example: Illegitimate surveillance

Social media monitoring (From Robert Munro)

In 2014–2015, I was approached by the Saudi Arabian government on three separate occasions to help them monitor social media…

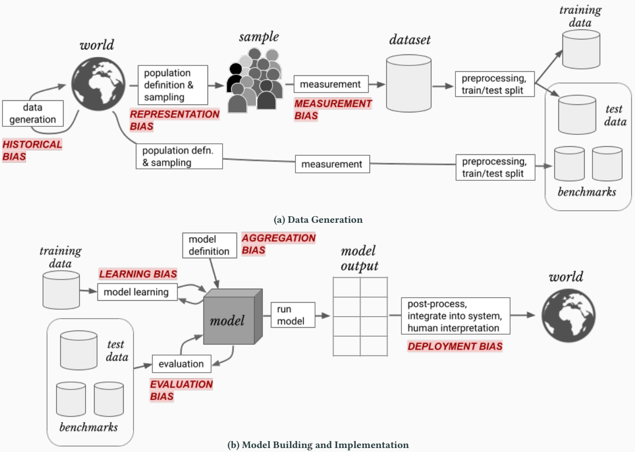
In every case, the stated goal was to help the people complaining about the government.

After careful consultation with experts on Saudi Arabia and Machine Learning, we decided that a system that identified complaints would be used to identify dissidents. As Saudi Arabia is a country that persecutes dissidents without trial, often violently, we declined to help.

Source: https://towardsdatascience.com/should-i-open-source-my-model-1c109188b164

Things to think about when building NLP systems

What part of model building has influence?



A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Suresh and Guttag (2021)

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.
- Text classification and IR can help identify information of interest

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.
- Text classification and IR can help identify information of interest but also aid censors.

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.
- Text classification and IR can help identify information of interest but also aid censors.
- Machine translation can be used to break language barriers

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.
- Text classification and IR can help identify information of interest but also aid censors.
- Machine translation can be used to break language barriers but also monitor marginalized populations.

Many applications of NLP have dual use

Even if we build systems intending to benefit people, it may have other uses that negatively affect people. What are they?

- Advanced grammar analysis can improve search and educational NLP but also reinforce prescriptive linguistic norms and discrimination.
- Stylometric analysis can help discover provenance of historical documents but also unmask anonymous political dissenters.
- Text classification and IR can help identify information of interest but also aid censors.
- Machine translation can be used to break language barriers but also monitor marginalized populations.

Ethics is not legality

- Unethical policies are often legal (e.g. there are and have been many legally enforced policies of discrimination).
- Not all ethical behavior is legally required—but you should behave ethically anyway.
- Sometimes law does enforce ethical practice (e.g. GDPR).

Questions to ask about your technology

Ethics is not a checklist. It is an ongoing conversation, and requires you to regularly question possible outcomes.

Questions to ask about your technology

Ethics is not a checklist. It is an ongoing conversation, and requires you to regularly question possible outcomes.

- Who are the stakeholders? This includes anyone who funds, develops, or uses your technology, and anyone it is *used upon*.
- Who benefits from the technology? How?
- Who could be harmed by the technology? How?

Questions to ask about your technology

Ethics is not a checklist. It is an ongoing conversation, and requires you to regularly question possible outcomes.

- Who are the stakeholders? This includes anyone who funds, develops, or uses your technology, and anyone it is *used upon*.
- Who benefits from the technology? How?
- Who could be harmed by the technology? How?

These are questions you must ask *yourself* and *all* of the stakeholders.

Summary of key points (i.e. examinable content)

- NLP is used by millions of people in the real world every day.
- NLP is used **on** millions of people in the real world every day.
- You must understand types of harms and where they come from.
- You must anticipate possible benefits *and harms*.

In the next lecture, we will look at bias in word embeddings.

Homework for next lecture: Do one of the implicit bias tests on `https://implicit.harvard.edu/implicit/`.

Bibliography i

Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Tatman, R. (2017). Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.