Ethics

- Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)
Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

Large body of work on Ethics
Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

How can we apply it to NLP?
Ethics

- Deontological framework for NLP
- Generalization principle
- Respect for Autonomy
- Reasonable, clear ethical rules, “rule of law”

Question-Answering  
Machine Translation  
Detecting objectionable content  
Dialogue Systems
Which tasks have important ethical implications?
What factors and methods are preferable in ethically solving this problem?
Generalization Principle
Generalization Principle

An action $A$ taken for reasons $R$ is ethical if and only if a world where all people perform $A$ for reasons $R$ is conceivable.
Generalization Principle

An action $A$ taken for reasons $R$ is unethical if and only if a world where all people perform $A$ for reasons $R$ logically contradicts $R$. 
Detecting objectionable content
Detecting objectionable content

A deploying flagging systems
Detecting objectionable content

A deploying flagging systems

R \downarrow burden on humans \downarrow \# posts that need to be seen by human eyes
Detecting objectionable content

\[ A \]
- deploying flagging systems

\[ R \]
- burden on humans
- \# posts that need to be seen by human eyes

“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

Hate Speech Detection

- API: Perspective
- Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

- Subjectivity
  - neutral: 0.1
  - polar: 0.9
- Polarity
  - pos: 0.5
  - neg: 0.5
- The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A deploying flagging systems

R burden on humans # posts that need to be seen by human eyes

- surface level words $\implies$ phrase the same meaning with different words

“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

Hate Speech Detection

API: Perspective

Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.5
- neg: 0.5

The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A deploying flagging systems

R burden on humans # posts that need to be seen by human eyes

- surface level words $\rightarrow$ phrase the same meaning with different words
- flagging system will be unsuccessful

"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."

Hate Speech Detection

Sentiment Analysis

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.5
- neg: 0.5

Unlikely to be perceived as toxic (0.23)
The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A deploying flagging systems

R burden on humans

# posts that need to be seen by human eyes

- surface level words $\Rightarrow$ phrase the same meaning with different words
- flagging system will be unsuccessful
- logically contradicts the premise

“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

[Hate Speech Detection](API: Perspective)

Unlikely to be perceived as toxic (0.23)

[Sentiment Analysis](python NLTK)

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.5
- neg: 0.5

The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

- Underlying intent, offensiveness, and power differentials between different social groups.
- Generate consequences and implications
- Does not lead to an arms race between objection content generation and detection

[We shouldn’t lower our standards just to hire more women.]

[Sap et al, ACL 2020]
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*

Zara infringe on Zara’s right to self-govern
Sanaa
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*

Zara must be sufficiently informed about ☒

Zara

Sanaa
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- Informed consent

Zara must be sufficiently informed about \( X \)

Zara *herself* makes the decision to allow Sanaa to do \( X \)
Translation

Translator

Zara

Sanaa
Translation

Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.
Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

There might be a formal contract of how Sanaa is to act.
Translation

Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

There might be a formal contract of how Sanaa is to act.

Zara relies on Sanaa’s paralinguistic conduct.
Machine Translation

Translator

Zara
Machine Translation

MT system is speaking for Zara

Zara

Translator

Machine Translation
Machine Translation

MT system is speaking for Zara

Zara must be *informed* of ambiguities so that she can *consent* to the message which the system ultimately conveys.
Machine Translation

MT system is speaking for Zara

Zara must be *informed* of ambiguities so that she can *consent* to the message which the system ultimately conveys.

Zara must also be *informed* of the failure cases in the MT system.
Machine Translation

Ms. Hashimoto …

-san? or -sensei?…

Zara

English to Japanese Machine Translation
Zara must be notified that such an ambiguity needs to be resolved because there is a risk of offending the Japanese speaker.
My aunt is coming home tomorrow.

Is it maternal or paternal aunt? They have different words in Hindi.
Machine Translation

MT system can ask a follow up question to Zara.

My aunt is coming home tomorrow.

Is it maternal or paternal aunt? They have different words in Hindi.

Zara

Translator

English to Hindi Machine Translation

Aadil
Zara: My aunt is coming home tomorrow.

Aadil: I am unable to translate the sentence in its current form. Can you please rephrase it?
NLP methods for Ethics
NLP methods for Ethics

**Machine Translation**: understand social context, control formality, politeness, author attributes, voice
NLP methods for Ethics

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice

**Detecting objectionable content:** generate consequences and implications
NLP methods for Ethics

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice

**Detecting objectionable content:** generate consequences and implications

**Question-Answering:** transparency, dynamic graph generation for answers
NLP methods for Ethics

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice

**Detecting objectionable content:** generate consequences and implications

**Question-Answering:** transparency, dynamic graph generation for answers

**Dialogue Systems:** control topics, style, content, persona
Summary

- Deontological framework for NLP
  - Generalization principle
  - Respect for Autonomy
- Four case studies
- Discussion