

# Ethical thinking

# Outline

- *Dual-use* technology
- *Exclusion* and *overgeneralization*
- The ethics of *massive multilingualism*, and *decolonizing* SLP
- The *ecological costs* associated with modern SLP

# Disclaimer

I am not an ethicist, I am merely a concerned practitioner and educator.

AI ethics, as a topic of research, is at present quite primitive, with whole contours of fruitful discussion not yet probed.

At the same time, it deserves more time than I can give it today.

(We're considering developing a course on AI ethics, in collaboration with the CS program at the Graduate Center.)

# Dual-use technology

*Dual-use* refers to technology useful for both peaceful and military applications. Nearly all SLP technologies are dual-use and their development for peaceful use may also contribute to military aims.

Note that this notion should not imply

- the boundary between peaceful and military use is always clear-cut, or
- that all peaceful uses of technology are ethical.

For example, it is known that multilingual *keyword spotting* is used by the NSA to spy on civilians (Froomkin 2015), and that *sentiment analysis* tools are used by various law enforcement agencies to surveill BLM supporters (Biddle 2020).

# HOW THE NSA CONVERTS SPOKEN WORDS INTO SEARCHABLE TEXT

Top-secret documents show the NSA can automatically recognize the content within phone calls, generating easily searched transcriptions.



Dan Froomkin

May 5 2015, 10:08 a.m.

# **POLICE SURVEILLED GEORGE FLOYD PROTESTS WITH HELP FROM TWITTER- AFFILIATED STARTUP DATAMINR**

Artificial intelligence firm Dataminr is making questionable use of its Twitter firehose privileges — and not for the first time.



Sam Biddle

July 9 2020, 2:00 p.m.

# Dual-use: an alternative framing

What are possible *prosocial* and *antisocial* uses of a technology?

E.g., stylometric analysis can be used to study authorship of historical texts (e.g., who wrote Shakespeare?) or to deanonymize dissenters (Hovy & Spruitt 2016).

What are *your* values? What do *you* think is pro- or anti-social? How do they align with legal principles in your community?

# Exclusion

Any linguistic data set bears *demographic bias*, information about the demographics of the speakers it represents.

*Demographic misrepresentation*, when the data used to build systems does not match the data it is used to model, results in what Hovy & Spruitt (2016) refer to as *exclusion*.

Most severely, speakers are excluded if systems do not support the languages, dialects, or registers they use.



# Overgeneralization

Exclusion is a side-effect of the data. *Overgeneralization* is a modeling side-effect. (Hovy & Spruitt 2016: 593)

Models may amplify biases present in the data.

E.g., Miller (2021) finds that models which predict gender from text systematically misgender trans people.

Models which incorrectly generalize may also lead to *bias confirmation*.

# Biased representations and debiasing

Bolukbasi et al. (2016), and much subsequent work, find that certain NLP models recapitulate, or even amplify, societal biases and stereotypes in the texts they are trained on. This has consequences for "downstream" models.

E.g., how would you translate the following Turkish sentence into English?

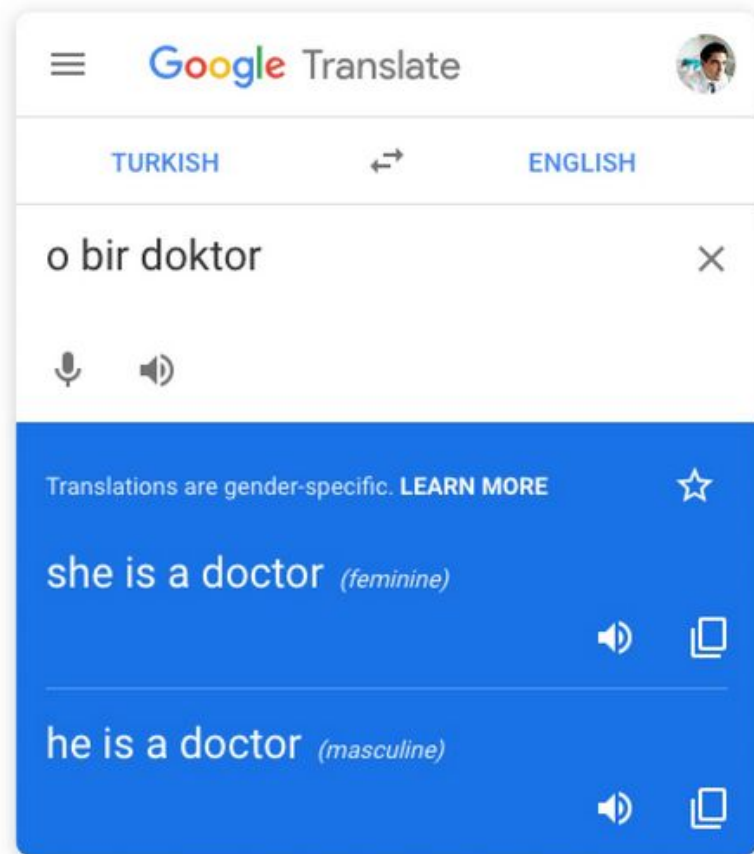
<i>O</i>	<i>bir</i>	<i>doktor.</i>
3sg.	is	doctor

Before



The screenshot shows the Google Translate interface. At the top, there is a hamburger menu icon, the Google Translate logo, and a profile picture. Below this, the source language is set to 'TURKISH' and the target language is 'ENGLISH', with a bidirectional arrow between them. The input text is 'o bir doktor' with a close button (X) on the right. Below the input, there are microphone and speaker icons. The output area is a solid blue box containing the text 'he is a doctor' followed by a shield icon and a star icon. At the bottom of the blue box, there are speaker, copy, and more options icons.

After



The screenshot shows the Google Translate interface with an updated output. The top navigation and input fields are identical to the 'Before' state. The output area is a solid blue box. At the top of this box, it says 'Translations are gender-specific. [LEARN MORE](#)' followed by a star icon. Below this, there are two lines of text: 'she is a doctor (feminine)' and 'he is a doctor (masculine)'. Each line has a speaker icon and a copy icon to its right.

# Data statements

Some initiatives towards better ethical practices have focused on procedures to encourage researchers to report on the provenance, intended use, and ethical issues of the data and/or models they provide. Examples of this include

- *data statements for natural language processing* (Bender & Friedman 2018) and
- *model cards* (Mitchell et al. 2019).

# "Multilingualism" in speech & language technology

An overwhelming majority of the speech and language processing research conducted over the last few decades can be described as *monolingual*.

Some technologies are *monolingual by design*: engineered to exploit the vagaries of English, or one of a few other richly resourced, globally hegemonic languages.

But far more often, researchers fail to demonstrate generalizability beyond a single language, once again, usually English (Bender 2009). Such research can be described as *monolingual by evaluation*.

Monolingualism persists **despite** the recent emergence of large, freely-available, high-quality data sets spanning dozens or even hundreds of languages.

# Massive multilingualism

One response to this has been to encourage the development of *massively multilingual* models, shared tasks, and data sets.

However, most research is conducted by small teams of engineers and linguists who will not be able to perform linguistically-informed quality assurance on dozens or hundreds of languages. In other words, one cannot simply look at the data.

For example, Gorman et al. (2019) carefully inspect the data for 12 of the 52 languages in the CoNLL-SIGMORPHON 2017 Shared Task on Morphological Reinflection and find many trivial, systematic errors that could have been caught by cursory inspection by speakers of those languages.

# Massive multilingualism: *cui bono*?

Scientists are interested in testing SLP across many languages: i.e., they may be interested in developing systems that are effective universally, across all languages.

Furthermore, the *ethic of inclusion* discussed by Hovy & Spruitt suggests that SLP should be available in all languages, though this needs to be balanced with the actual desires of "stakeholders".

Corporations also wish for their products (ad copy, etc.) to reach as many speakers as possible, though fiduciary and scientific motives may be misaligned.

State actors have been particularly interested in SLP technologies for the languages used in competitor states.

# Decolonizing speech and language technology

Bird (2020) argues that the "gotta catch 'em all" approach to endangered languages recapitulates colonialism:

*After generations of exploitation, Indigenous people often respond negatively to the idea that their languages are data ready for the taking. By treating Indigenous knowledge as a commodity, speech and language technologists risk disenfranchising local knowledge authorities, reenacting the causes of language endangerment. (p. 3504)*

Bird continues to suggest ways in which SLP researchers might begin take a postcolonial approach to revitalization of Indigenous languages.



# Ecological costs

Strubell et al. (2019) attempt to estimate the carbon footprint (i.e., CO<sub>2</sub> emissions) associated with certain newer SLP methods (acknowledging that most power generation uses a mixture of carbon-emitting and non-carbon-emitting sources). They find that certain deep learning techniques,

- particularly those associated with from-scratch training of massive neural network language models like ELMo, BERT, and GPT-2, and
- particularly those associated with applying a hyperparameter tuning method known as *neural architecture search*

may account for a large fraction of a user's yearly CO<sub>2</sub> emissions.

<b>Consumption</b>	<b>CO<sub>2</sub>e (lbs)</b>
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
<b>Training one model (GPU)</b>	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO<sub>2</sub> emissions from training common NLP models, compared to familiar consumption.<sup>1</sup>

# Other ethical issues

- Privacy
- Legal compliance
- Human subjects protection (e.g., when annotators are also "subjects")

Questions?