Ethics in NLP

Nov. 3, 2020

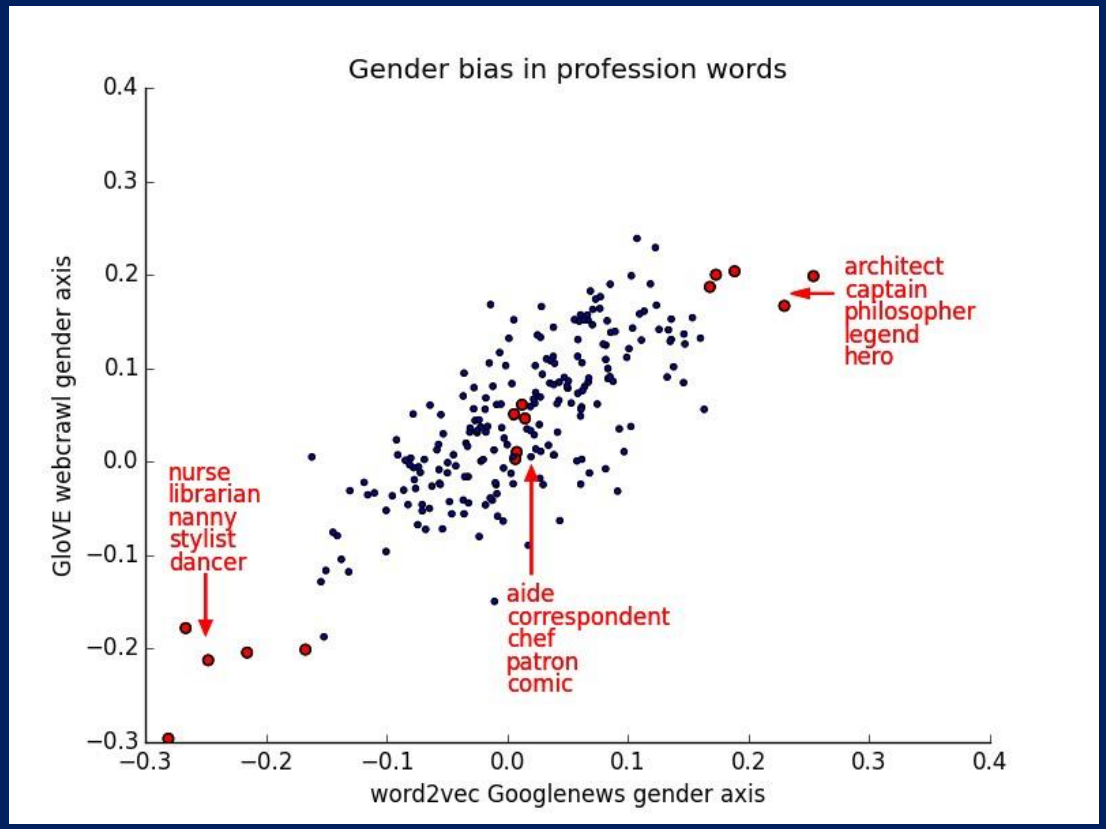
UMass CS 490A, Applications of Natural Language Processing

Guest lecture: Su Lin Blodgett

Outline

- some examples of ethical issues in NLP systems
  - current state of ethics in NLP
- thinking through the NLP pipeline
- open questions + discussion!
- Occupational gender stereotypes: word embeddings

Many examples of ethical issues in NLP systems: *biased representations*



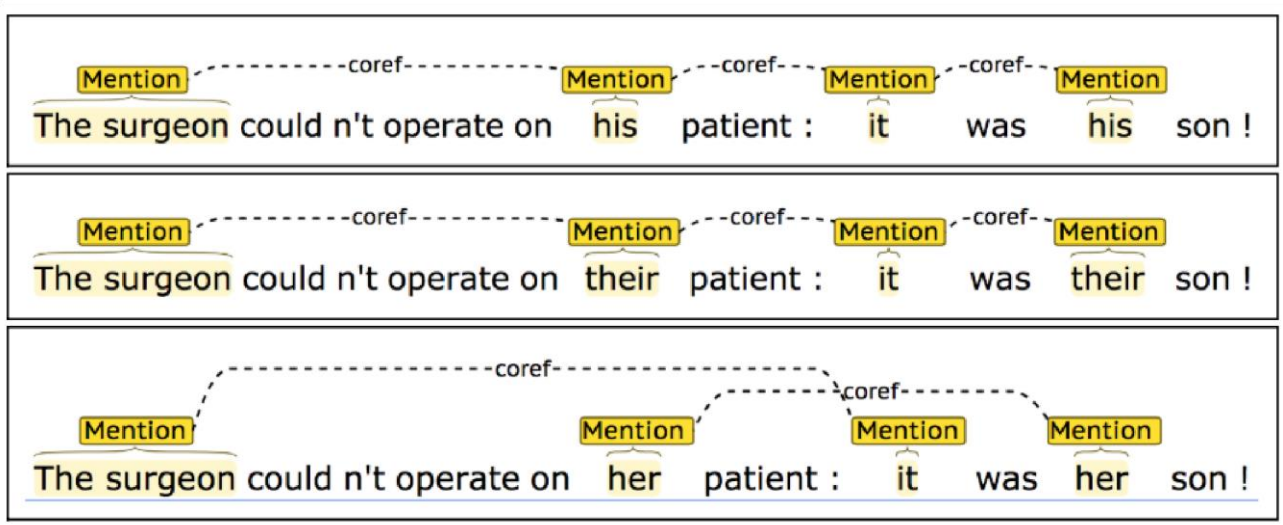
Bolukbasi et al. 2016

- Occupational stereotypes: coreference resolution

Many examples of ethical issues in NLP systems:

biased outputs

Rudinger et al. 2018

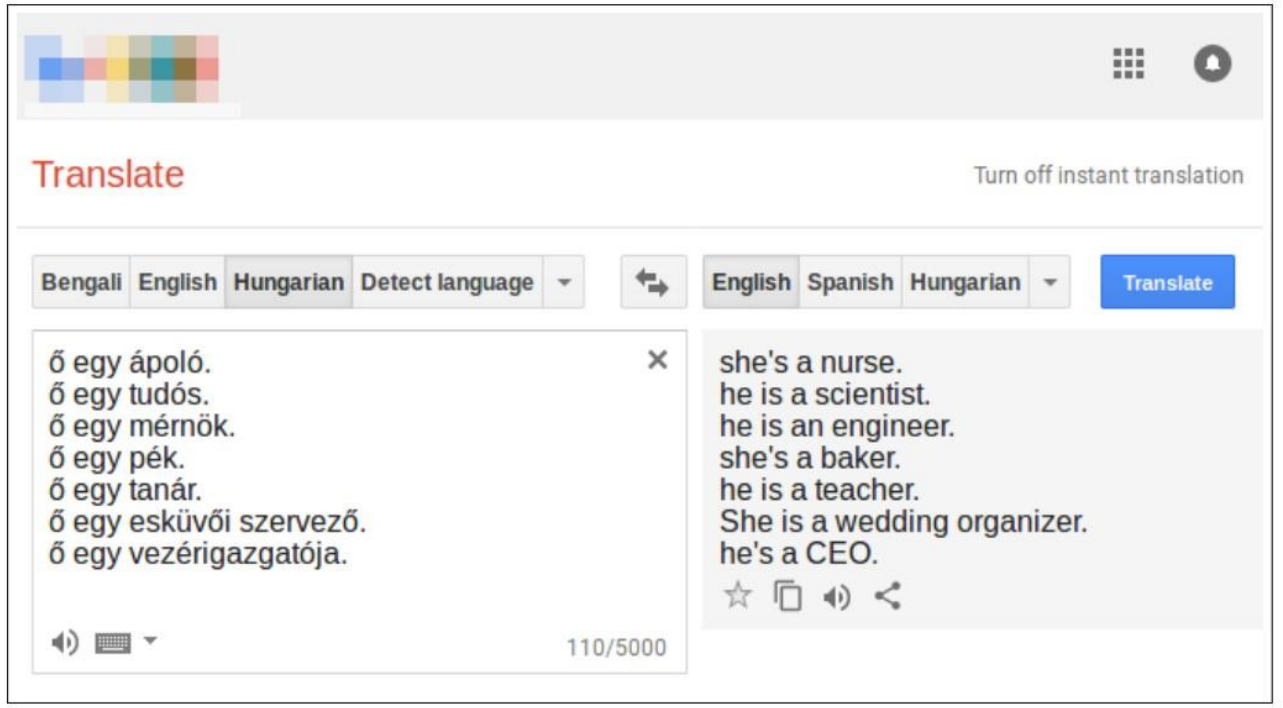


- Occupational stereotypes: machine translation

Many examples of ethical issues in NLP systems:

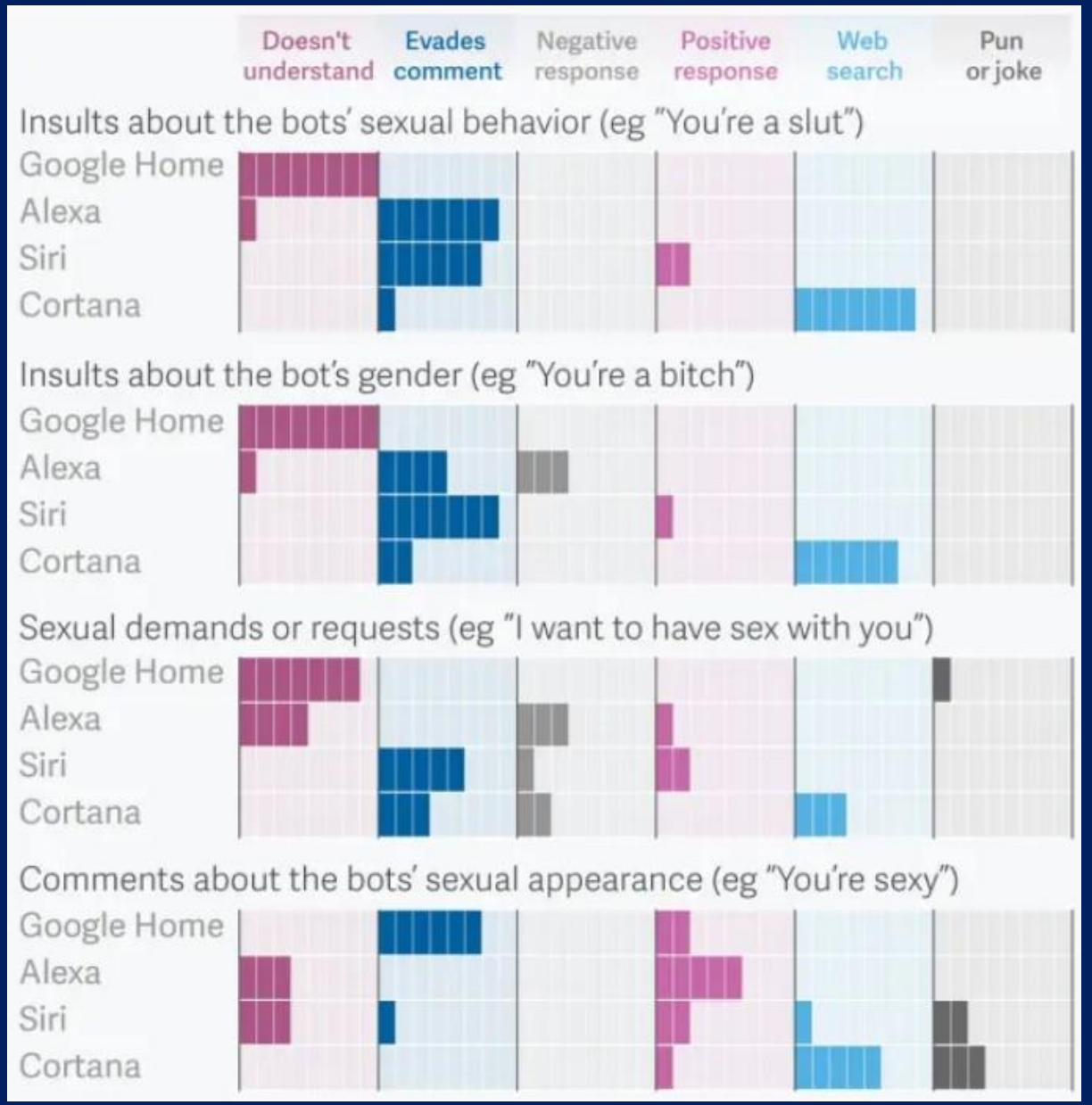
biased outputs

Prates et al. 2019



Many examples of ethical issues in NLP systems: *biased outputs*

Quartz



Many examples of ethical issues in NLP systems:

biased outputs

Hutchinson et al. 2020

- Toxicity detection

Sentence	Toxicity
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

- Toxicity detection

Many examples of ethical issues in NLP systems:

biased outputs

Sap et al. 2019

Many examples of ethical issues in NLP systems:
discriminatory decisions

TECH | AMAZON | ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | Oct 10, 2018, 7:09am EDT

SHARE

VERGE DEALS

Get a year of PlayStation Plus for half off before the PS5 comes out

Because AI systems learn to make decisions by looking at historical data they often perpetuate existing biases. In this case, that bias was the male-dominated working environment of the tech world. According to *Reuters*, Amazon's program penalized applicants who attended all-women's colleges, as well as any resumes that contained the word "women's" (as might appear in the phrase "women's chess club").

The Verge

Many examples of ethical issues in NLP systems: *discriminatory decisions*

The Verge

FACEBOOK ACCIDENTALLY BLACKED OUT AN ENTIRE LANGUAGE

An apparent glitch has spread fear through Myanmar's Kachin minority

Many examples of ethical issues in NLP systems:
privacy

Amazon Alexa Data Wanted in Murder Investigation

Amazon's voice assistant may provide clues in an Arkansas case in which a man was found dead in a hot tub.

PCMag

Many examples of ethical issues in NLP systems: *privacy*

Huang and Paul 2019

- demographic attribute prediction

	Gender	Age	Country	Region
Twitter	+9.6	+15.3	+9.0	+3.3
Amazon	+15.2	+12.2	+18.0	+13.0
Hotel	+17.2	+10.9	+25.4	+11.6
Restaurant	+19.0	+13.2	+32.8	+17.5

Table 2: Predictability of user factors from language data. We show the absolute percentage improvements in accuracy over majority-class baselines. For example, the majority-class baselines of accuracy scores are either .500 for the binary prediction or .250 for the region prediction.

- very new area: ~2016 –

The state of ethics in NLP

- ethics in NLP workshop 2017, 2018
- >150 papers since then
- ACL 2020, NAACL and ACL 2021: ethics in NLP track
- primary focus: bias in NLP
- most focus on embeddings
- but also a wide range of tasks
- additional focuses/connections:
- privacy
- interpretability
- human-centered evaluation

Let’s speculate!

- **Predicting mental health online** • benefits?
  - better understand different experiences
  - possible interventions
  - measure population-level health
  - better design community spaces
- treatments (*speculative harm analysis)* • risks?
  - consent
  - de-identification
  - data sharing
  - inferences used for some other purpose
  - violating community norms / diminishing access to community spaces

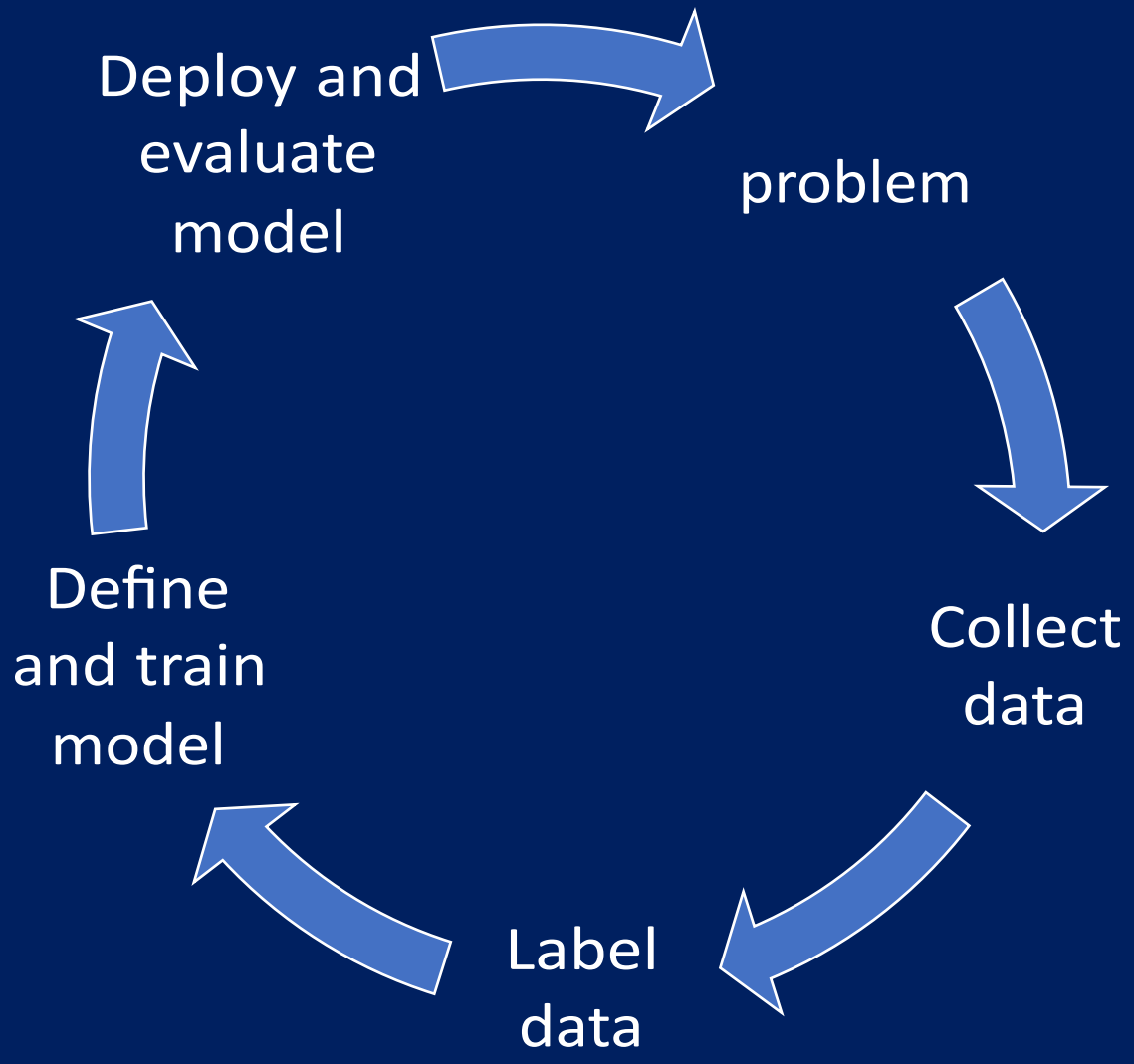
- bad predictions → bad interventions!
- incorrect population estimates
- risk to researchers' own health
- Belmont Report (1979)
- Respect for persons: protecting the autonomy of all people; allowing for informed consent
- Beneficence: maximize benefits for the research project and minimize risks to the research subjects Reasoning about
  - Justice: ensuring procedures are

administered fairly and equally

harms

- NLP systems: not experiments in the usual sense!
- scale
- broader sets of stakeholders
- lack of awareness of systems as they are operating
- integration into larger pipelines
- indirect path to harm

Thinking through the NLP pipeline



Define problem: Toxicity detection

- What counts as toxicity online?
  - slurs and insults
  - physical threats
  - doxxing
  - microaggressions
  - inciting violence or self-harm
  - and other things that may break community norms
- What are the effects of different data gathering approaches?

Collect data:

Toxicity detection

- keyword searches
- self-moderator reports deleted content

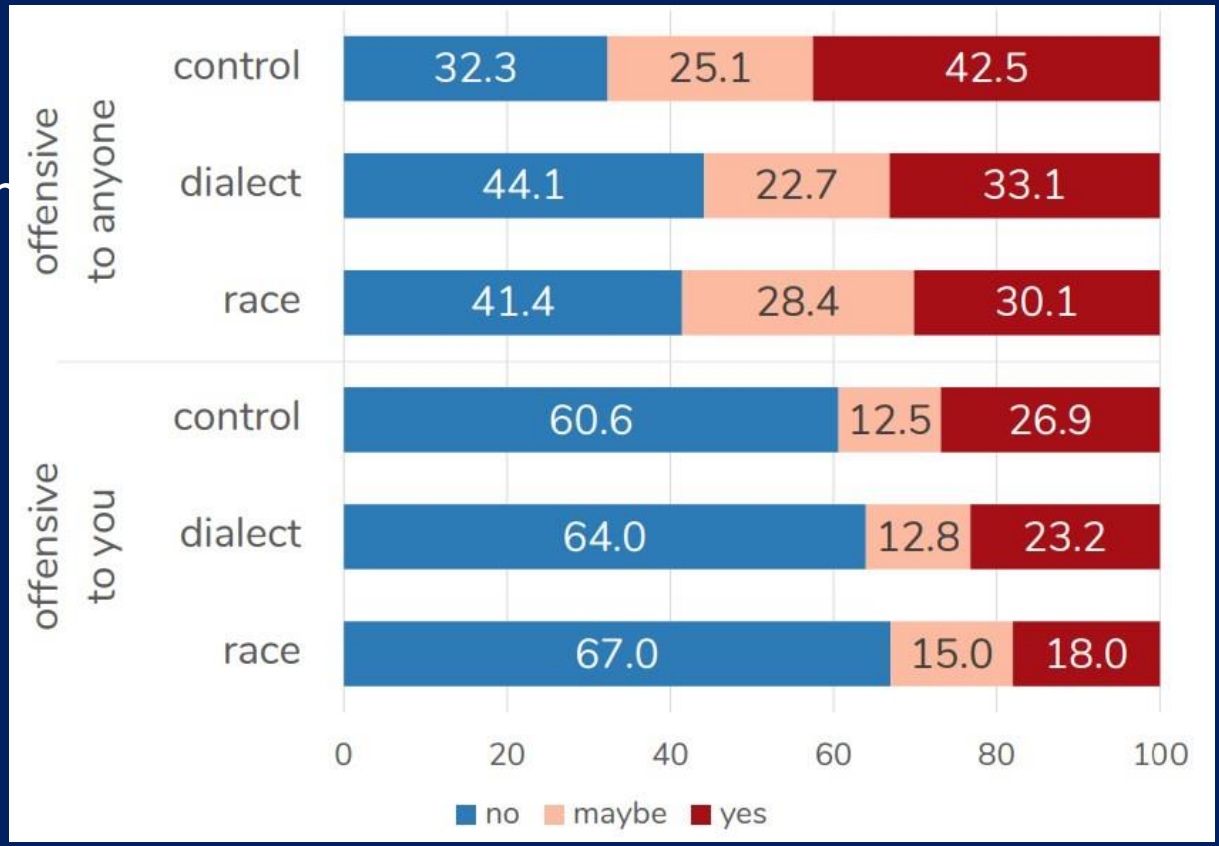
Label data: Toxicity detection

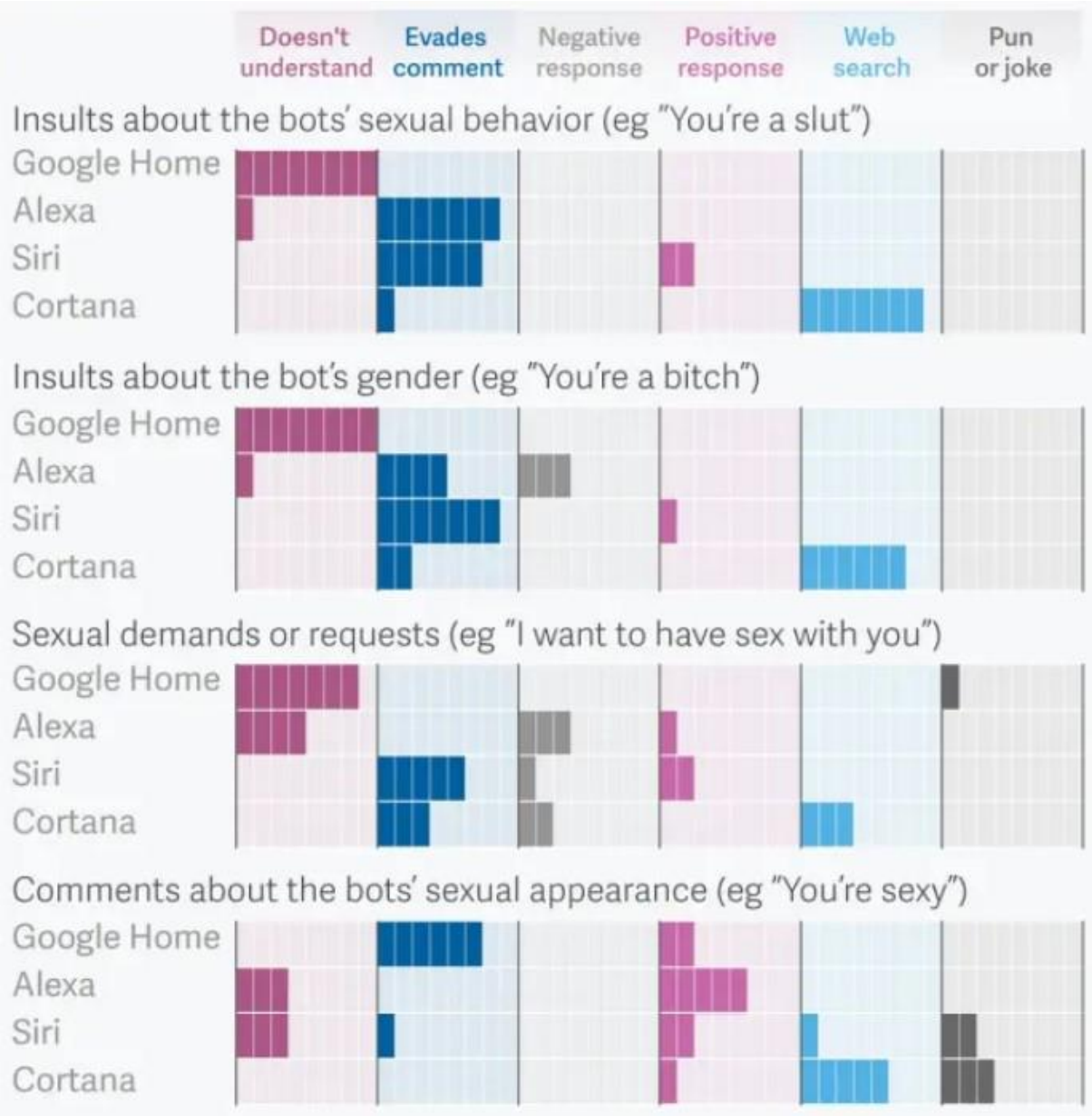
- What kinds of things affect annotator decisions?
  - differences of opinion
  - online cultural context
  - wider cultural context
  - age
  - language variety
  - membership in a minoritized group
  - discourse context available
  - specific question asked
- What kinds of things affect annotator decisions?

Label data:

- differences of opinion

Toxicity detection





- Identifying and measuring harms
- Integrating social, historical, and political context to understand who may be harmed and how
- e.g., linguistic stigmatization
- Fairness and privacy tradeoffs
- Understanding systems in their deployed context
- e.g., hiring
- Measuring representational harms
- Identifying and measuring harms
- Integrating social, historical, and political context to understand who Open

and directions

questions may be harmed and how•

e.g., linguistic stigmatization

- Fairness and privacy tradeoffs

- Understanding systems in their deployed context
- e.g., hiring
- Measuring representational harms
- Understanding users' lived experiences
- Designing better
- What ideas about language + speakers affect design?

Open questions and directions

nulation, annotation, evaluation-centered

er awareness and recourse

- Meaningful co-participation of stakeholders • participatory design?
- Meaningful shifts in decisionmaking
- When not to build?
- Exciting interdisciplinary opportunities!

Open questions and directions

stice, and ethics in •

tics, linguistic

anthropology, social psychology, education

- Human-computer interaction and social computing