CS 232: Artificial Intelligence

Fall 2023

Prof. Carolyn Anderson
Wellesley College

Announcements

- Francesca Lucchetti will be giving a guest lecture in CS 333 on Tuesday.
- My help hours next week:
  - Monday: 3:30-5
  - Friday: 3:30-4:30
- RUR invited dress rehearsal next Wednesday at 7pm

Recap

Vocabulary time!

Epistemic: related to knowledge. Epistemic questions are about what is true, what is known, or what is possible.

You can have a dessert (dessert exists).

Deontic: related to duty or to desire. Deontic questions are about what should or ought to be according to some set of obligations, desires, or norms.

You can have a dessert (you are allowed to).

Normative: related to an evaluative standard. Normative statements say how things *should* be, not how they are.

Stakeholders

There are different kinds of stakeholders to consider when we talk about the ethics of AI (Bender 2019):

- **Voluntary direct stakeholders**: people who choose to use the system.
- **Involuntary direct stakeholders**: people who must use the system in order to access essential services.
- **Indirect stakeholders**: subjects of queries, contributors to a corpus (voluntarily or involuntarily) communities impacted by model predictions

Representational Harms

- **Stereotypes**: the system propagates negative generalizations about certain social groups
- **Misrepresentation**: the system performance is skewed towards certain groups of people
- **Erasure**: the system fails to recognize other groups of people
- **Denigration**: the system contains or uses language that is harmful to the dignity or well-being of some people
- **Alienation**: the system denies the relevance of socially meaningful categories

Allocational Harms

- **Quality of service**: the system performs better for individuals who belong to some groups than for others
- **Public participation**: the system makes the speech or contributions of individuals in certain groups less visible than others.
- **Resource allocation**: the system is used in a way that allocates resources more to individuals from one group than another.
- **Opportunity allocation**: the system is used in a way that allocates opportunities more to individuals from one group than another.
- **Targeted surveillance**: the system is used to profile or monitor individuals based on their demographic characteristics.
- **Predictive generalization**: there are disparate impacts across social groups in the treatments/interventions recommended by a system.

Where Does Harm Come From?

Discussion largely based on Blodgett (2021)

Harms from Task Design

Case study: bias mitigation in toxicity detection

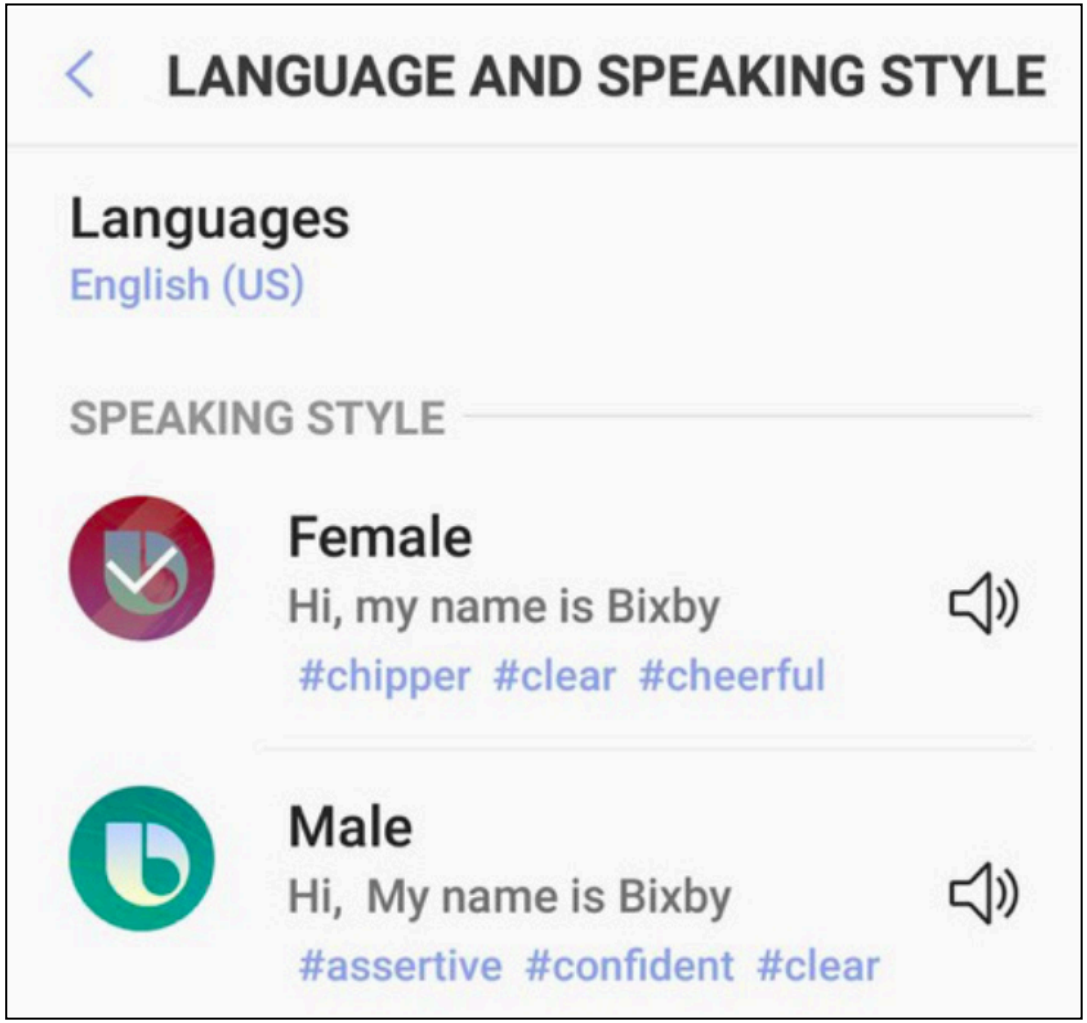
One proposed solution to bias in toxicity detection is minimize group differences in toxicity ratings. But this incorrectly assumes that toxic language is generated and applied evenly across demographic groups (Young 2011, Garg et al. 2019, Hanna et al., 2020).

It also fails to differentiate between toxic language and reclaimed in-group usage of the same terms.

Case study: voice assistants

In 2019, out of 70 voice assistants explored by the EQUALS Research Group, 2/3 had female-only voices.

The way that these voice assistants are portrayed may reinforce gender stereotypes of women as caring and subservient (UNESCO 2019).



Case study: voice assistants

In addition, some female-coded voice assistants have been programmed to respond to sexual harassment and anger in ways that reinforce harmful attitudes.

In 2017, *Quartz* found that Siri responded provocatively to sexual harassment by men ('Oooh!'; 'Now, now'; 'I'd blush if I could'), but less so to women ('That's not nice').

These design choices perpetuate "a sexist expectation of women in service roles: that they ought to be docile and self-effacing, never defiant or political, even when explicitly demeaned" (Fessler 2018).

Case study: coreference resolution

Cao and Daumé (2020) survey published 150 NLP papers that mention gender to explore whether they reinforce folk theories of gender:

- 5.5% distinguished social from linguistic gender
- 5.6% were inclusive of non-binary identities
- 100% treated gender as immutable
- 7.1% considered definite singular *they* and neopronouns

Rudinger et al. (2018): coreference systems do not work on *they* pronouns and perform better on *he* than *she*.

Cao and Daumé (2020): coreference systems achieve 95% accuracy on *he* and *she*, 90% on *they*, and 0-13% on neopronouns.

Case study: style transfer

Ongoing research in style transfer attempts to condition model output on social categories. For instance, there is work that seeks to reduce gender bias in hiring by "de-gendering" resumés.

But conditioning on social categories can reinforce stereotypes and, particularly in the case of gender, essentialize traditional gender divisions.

Harms from Application Contexts

Do we always want better systems?

Can you think of any cases where it might be better not to improve the performance of an AI system?

Carework - is it ethical to automate?

Admissions - could fix bias in
Job applications humans rather than automating?

Drone warfare

Do we always want better systems?

- Border surveillance
- Drone warfare
- Facial recognition used to monitor and control minority populations
- Language screening used to validate refugee histories
- Voter suppression techniques targeted at minoritized communities
- Labor surveillance

Sometimes it might be better to be invisible to the system

Case study: language identification

Automated language identification is used the refugee screening process in Germany as part of testing the validity of refugees' histories.

Problem 1:
The state-of-the-art is not very good

Problem 2:
These systems reinforce the idea that languages are "*fixed entities capable of being counted, systematized, and named*" (Severo and Makoni 2020). They will always perform better on language varieties spoken by dominant groups. They cannot adapt quickly to language innovation.

Snapshot of work on bias in NLP

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Table 5.1: The NLP tasks covered by the 146 papers.

From Blodgett (2021)

Looking for a topic to work on? Consider what is missing or unrepresented in this table!
speech-to-text, question-answering systems, text-to-speech, information retrieval...

Harm Reduction

Microsoft Harms Modeling

Categories of potential harms

- Risk of injury
  - Physical injury
  - Emotional or psychological injury
- Denial of consequential services
  - Opportunity loss
  - Economic loss
- Infringement on human rights
  - Dignity loss
  - Liberty loss
  - Privacy loss
- Environmental impact
- Erosion of social & democratic structures
  - Manipulation
  - Social detriment

Microsoft Harms Modeling

For each category of harm, consider its:

Contributing factor	Definition
Severity	How acutely could an individual or group's well-being be impacted by the technology?
Scale	How broadly could the impact to well-being be experienced across populations or groups?
Probability	How likely is it that individual or group's well-being will be impacted by the technology?
Frequency	How often would an individual or group experience an impact to their well-being from the technology?

Ethics assessment model: community jury

In the **community jury** model, the potential harms and benefits of a proposed technology are weighed by a diverse group of stakeholders.

- The **product team** creates relevant documentation, data management plan, and prototypes to present.
- A **moderator** facilitates discussion and deliberations.
- A **jury** is assembled of 16-20 community members, sampled in a way that is random but ensures a demographically diverse group.

Ethics assessment model: community jury

2-3 hr sessions are held to assess the proposed technology:

- **Overview and introduction**: The moderator explains the rules of conduct. The product team explains the proposed technology and its goals.
- **Q&A**: jurors ask questions about the technology.
- **Deliberation and cocreation**: the jury and product work together to come up with solutions that meet all needs.
- **Anonymous surveying (optional)**: anonymously poll the jurors to get their honest opinions.
- **Study report**: the moderator writes a report outlining key insights, concerns, and proposed solutions.

Scenarios

- The National Science Foundation is considering replacing its peer review system for reviewing grant applications with an automated system.
- A farmer is considering adopting a system developed by UC Berkeley computer scientists that uses computer vision to identify pests and zap them with lasers.
- UT Austin is considering using an automated system to screen MS and PhD candidates in Computer Science.
- Roblox, a platform where people can program and share games, is collecting code to train a large language model of code to improve the experience of novice programmers. They are using an opt-in mechanism.
- Stable Diffusion releases an image generation model trained on data scraped from the internet.

Scenario: Code Generation

Roblox, a platform where people can program and share games, is collecting code to train a large language model of code. Their goal is to improve the experience of novice programmers.

99%

amateurs

1%

professionals

Stakeholders

- Roblox game developers
- Roblox would-be game developers
- Roblox players
- Roblox technologists

Parents

Other companies: OpenAI, HuggingFace, Microsoft, etc.