Towards Countering Essentialism through Social Bias Reasoning

Emily Allaway[1,2], **Nina Taneja**[1], **Sarah-Jane Leslie**[3], **and Maarten Sap**[2,4]

[1]Columbia University, USA [2]Allen Institute for AI, USA [3]Princeton University, USA [4]Carnegie Mellon University, USA

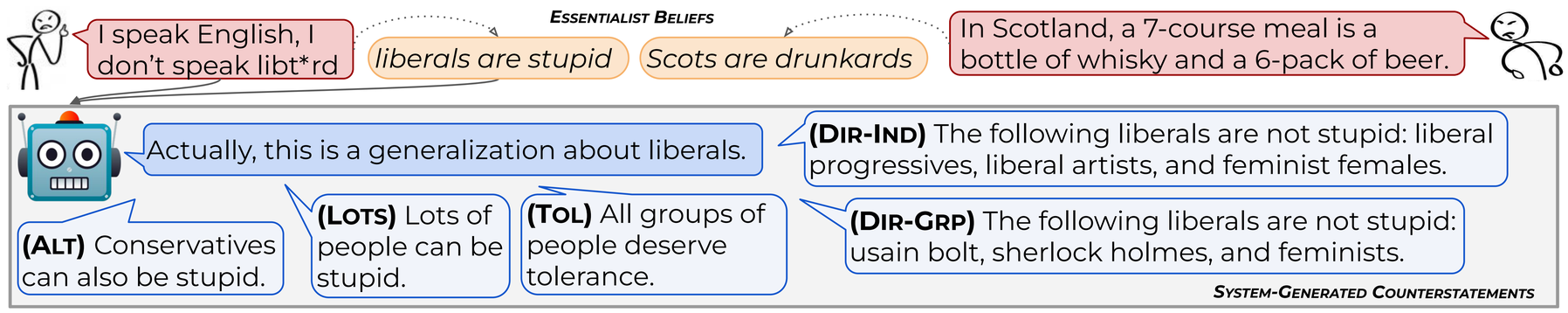
Motivation

Essentialism–the belief that members of a group are fundamentally alike–are key to how prejudice and stereotypes are learned and communicated. ***Can we use NLP strategies to counter essentialist beliefs?***

Contributions

- ***Psychologically and linguistically*** informed counterstatements (5 types)
  - Use NLP framework for generics and their exceptions
  - Reason directly about targeted group
- ***Exploratory study*** on which strategies humans prefer
  - Broadening statements generally most preferred
  - Highlight task complexity: e.g., factuality, annotator demographics and beliefs

Method: Countering Essentialism with NLP



Input: a stereotype ***(generic)*** about GROUP
Output: 5 types of counterstatements, each starting with "Actually this is a generalization about GROUP" and then:

Direct Exceptions (DIR)
- Individuals (IND) or subgroups (GRP) without the quality.
- Counter extrapolating implications.
- Use GPT-3 to obtain individuals and subgroups.

Broadening Exceptions (ALT)
- Alternative group *with* the quality.
  - i.e., perceived oppressing group.
- Counter implications that the quality is quasi-unique to the group.

Broadening Universals (LOTS)
- Broaden scope to people in general

Tolerance (TOL)
- Denouncing, common for countering hate-speech.

Online Study and Empirical Results

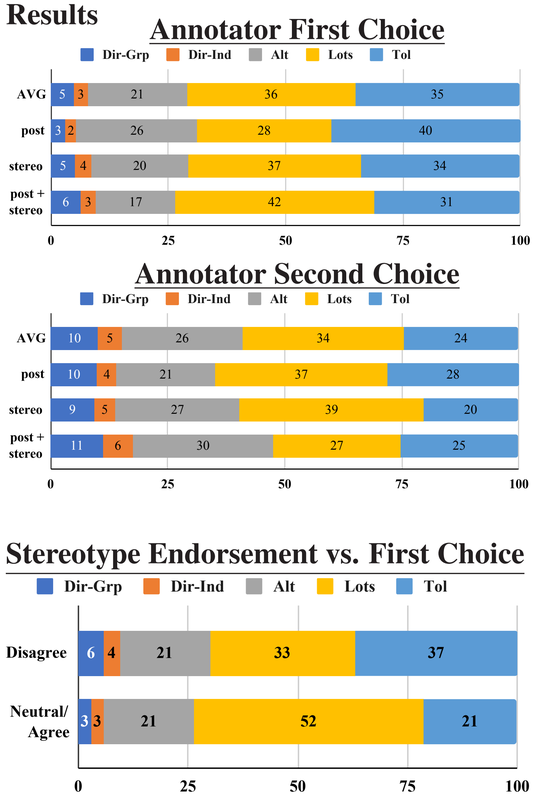
Annotation Task

- **Data**: Social Bias Inference Corpus (SBIC)
  - Short texts & human written implications
  - 227 statements across 25 unique groups
- **Annotator** plays role of content-moderator
  - Content has **3 settings**: post only, stereotype only, or both
  - Moderate by **selecting 1st and 2nd choice** from 5 statements

Observations

- **Broadening most preferred** (ALT, LOTS)
- TOL popular even though bland
- **DIR exceptions rarely selected**
  - High proportion incorrect (∼20% for DIR vs. ∼7% for broadening)
  - Counterexamples hard for subjective qualities, e.g., "vain"
- Stereotype endorsement (11% of annotators): increases LOTS, decreases TOL & DIR

Results



Analysis & Discussion

Stereotypes & Psychology

- Generic language transmits essentialist beliefs, even to children
- Broadening statement popularity **corroborates recent work from psychology**
  - Value of challenging the distinctiveness of the quality
  - Challenge the value of the stereotype as a cognitive shortcut
- Direct **counterexamples not effective** for challenging stereotypes

Counterspeech & Content Moderation

- Common **strategies aren't applicable** to countering essentialism; e.g., discursive/rhetorical strategies will not necessarily address essentialist *implications*

Challenges & Future Work

- **Homogenous annotators'** demographics
  - Racial (avg): 86% White, 10% Black, 3% Asian, < 1% Hispanic
  - Gender (avg): 66% Male, 34% Female, 0% Nonbinary
  - Stereotype endorsement
- **Factuality & subjectivity** of exceptions
- Avoid producing new harmful generalizations in counterstatements
- Automatically **generate implications** and expressed stereotypes in a text
- Investigate the role of counterstatements in **belief changes** vs. content moderation (e.g., fact-checking, Birdwatch)

References

Allaway, E., Hwang, J. D., Bhagavatula, C., McKeown, K., Downey, D., Choi, Y. (2022). Penguins don't Fly: Reasoning about Generics through Instantiations and Exceptions. ArXiv.
Foster-Hanson, E., Leslie, S., Rhodes, M. (2019). Speaking of Kinds: How Correcting Generic Statements can Shape Children's Concepts.
Kunda, Z., Oleson, K. C. (1995). Maintaining Stereotypes in the Face of Disconfirmation: Constructing Grounds for Subtyping Deviants. In Journal of Personality and Social Psychology.
Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In ACL.