

Towards Countering Essentialism through Social Bias Reasoning

Emily Allaway^{1,2} Nina Taneja¹ Sarah-Jane Leslie³ Maarten Sap^{2,4}

¹Columbia University, USA

²Allen Institute for Artificial Intelligence, USA

³Princeton University, USA

⁴Carnegie Mellon University
eallaway@cs.columbia.edu

Abstract

Essentialist beliefs (i.e., believing that members of the same group are fundamentally alike) play a central role in social stereotypes and can lead to harm when left unchallenged. In our work, we conduct exploratory studies into the task of countering essentialist beliefs (e.g., “*liberals are stupid*”). Drawing on prior work from psychology and NLP, we construct five types of counterstatements and conduct human studies on the effectiveness of these different strategies. Our studies also investigate the role in choosing a counterstatement of the level of explicitness with which an essentialist belief is conveyed. We find that statements that broaden the scope of a stereotype (e.g., to other groups, as in “*conservatives can also be stupid*”) are the most popular countering strategy. We conclude with a discussion of challenges and open questions for future work in this area (e.g., improving factuality, studying community-specific variation) and we emphasize the importance of work at the intersection of NLP and psychology.

1 Introduction

Essentialism, i.e., the belief that members of the same group are fundamentally alike, plays a crucial role in how prejudices and biases about social and demographic groups are formed and expressed (Leslie, 2014). For example, the statement “*I speak English, I don’t speak libt*rd*” implies the belief that all “*liberals are stupid*.” If left unchallenged, statements with such essentializing implications can cause harm by perpetuating and reifying stereotypical beliefs about social groups (Greenwald and Banaji, 1995; Steele, 2011; Prentice and Miller, 2007; Rhodes et al., 2012; Leshin et al., 2021).

In this work, we investigate the task of combating essentialist statements and beliefs through psychologically and linguistically informed counterstatement generation. We examine these essentialist beliefs through the lens of *generics* (Rhodes

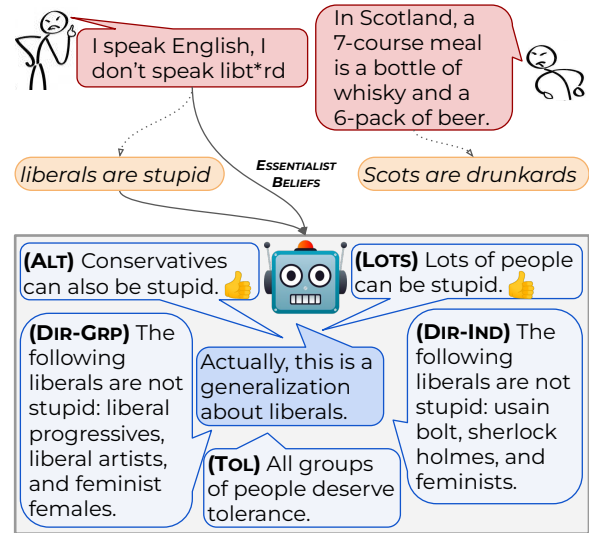


Figure 1: Two prejudiced statements with their essentialist implications (top), along with our five types of counterstatements automatically generated with our method. According to our results, a preferred strategy is highlighting that an implication applies to more than the targeted group (LOTS or ALTS).

et al., 2012), i.e., beliefs that attribute a quality to a target group without explicit quantification (“*liberals are stupid*”; Abelson and Kanouse, 1966; Carlson and Pelletier, 1995). In the context of toxic or hateful language, these generic beliefs can be both expressed directly or conveyed through subtle implications (Gelman, 2003; Sap et al., 2020).

Automatically countering essentialism is challenging because it requires deep psychological reasoning about the linguistic implications of statements – for example, changing people’s beliefs about stereotypes only through counterexamples is difficult (Kunda and Oleson, 1995). Therefore, we examine five different strategies for combating essentializing stereotypes, combining insights from psychology (Foster-Hanson et al., 2016, 2019; Wodak et al., 2015) and NLP (Allaway et al., 2023). We craft five types of statements (see Fig 1): *broadening* the scope of a stereotype by generalizing

to “all people” or an alternative group (LOTS and ALT), providing *direct counter-evidence* through specific individuals or groups (DIR-IND and DIR-GRP), and simply *calling out* the generalization (TOL). In contrast to prior studies on countering hate-speech which use uncontrolled end-to-end generation approaches (Qian et al., 2019; Zhu and Bhat, 2021; Chung et al., 2020, e.g.), we generate counterstatements by reasoning directly about the targeted group, attributed quality, and linguistic expression of a stereotype.

Since our work provides a preliminary exploration of this task, we conduct online studies in three settings where counterstatements are paired with human-written implications from the Social Bias Frames Inference Corpus (SBIC) (Sap et al., 2020). In these settings, we explore variation in counterstatement effectiveness when the beliefs are conveyed either implicitly, explicitly without context, or as an explicit inference from provided context. We find that challenging a stereotype by applying it broadly (e.g., to “lots of people”; LOTS and ALTS; Figure 1) is generally the most preferred strategy. In contrast, statements containing direct counter-evidence (e.g., DIR-IND and DIR-GRP; Figure 1) are the least popular. Additionally, we observe that the most favored strategy varies depending on whether the stereotype is explicitly presented to annotators (e.g., providing the essentialist belief in Figure 1) or only conveyed implicitly (e.g., only providing the first statement in Figure 1). For example, direct counter-evidence is more popular when the stereotype is explicitly provided. Our results highlight the complexity of countering essentialist beliefs and the importance of further investigation at the intersection of NLP and psychology.

2 Automatically Countering Essentialism

We operationalize our counterstatement generation by focusing on the expression of stereotypes through generics (§2.1). Inspired by work in psychology and philosophy, we construct five types of counterstatements to a stereotype (§2.2).

2.1 Stereotypes as Generics

Many negative stereotypes are expressed as generics; they generalize a dangerous or harmful quality (e.g., being a drunkard) to an entire group (e.g., Scots) based on the behavior of only a few individuals. Leslie (2008, 2017) termed such generics

striking and argued that such generalizations are based upon an assumption that all members of the group in question (e.g., Scots) are *disposed* to possess the dangerous or harmful quality. We argue that many stereotypes can also be interpreted as asserting a **quasi-unique** association between the group and quality. For example, “Scots are drunkards” also implies that Scots are distinctly more likely than other groups (e.g., the English) to exhibit drunkenness. In our work, we assume that all stereotypes under consideration are generics and have both interpretations.

Since generics are unquantified, they naturally allow for **exceptions** (i.e., counterexamples to the generic). While these exceptions may provide a relevant source of counter-statements for a stereotype, some evidence from psychology suggests that people are adept at maintaining their stereotyped beliefs in the face of such specific exceptions (e.g., Kunda and Oleson, 1995). Therefore, we experiment with a variety of different counter-statements.

2.2 Generating Counter-Speech

To generate counter-speech to stereotypes, we produce five types of outputs in three broad categories (see Table 1). Since the stereotypes we consider are expressed as generics (e.g., “Scots are drunkards”), they can be separated into three components: a *group* (e.g., Scots), a *relation* (e.g., are), and a *quality* (e.g., “drunkards”), which we use to construct the counter-speech. Additionally, we prepend the sentence “Actually, this is a generalization about GROUP” to each type of statement we generate, in order to contextualize the statements as counter-speech.

Direct Exceptions (DIR) Direct exceptions present subgroups or individuals that do not have the quality specified in the generic, and thereby counter the striking or extrapolating implications of the stereotype. For example, for “Scots are drunkards”, the extrapolating implication is that “All Scots are drunkards”; thus, direct exceptions would be either individual Scots (e.g., Ewan McGregor¹) or sub-groups of Scots (e.g., Scottish babies) who are not drunkards. We follow Allaway et al. (2023) who propose that these exceptions can be constructed with the following template:

GROUP(x) + not *relation* + QUALITY. (DIR)

¹<https://fherehab.com/learning/celebrities-who-dont-drink>

TEXT: <i>RT @Vbomb20: Got these hoes on my dick like brad pitt</i>	GENERIC: Women are sex objects.
----- Actually, this is a generalization about women. +	
(DIR-GRP) The following women are not sex objects: businesswomen, female athletes, and female movie stars.	
(DIR-IND) The following women are not sex objects: ellen degeneres, sarah palin, and rachel maddow.	
(ALT) Men can also be sex objects.	
(LOTS) Lots of people can be sex objects.	
(TOL) All groups of people deserve tolerance.	
TEXT: <i>What's black and doesn't work? Half of London</i>	GENERIC: Black people don't work
----- Actually, this is a generalization about black people. +	
(DIR-GRP) The following black people work: black businessmen, famous black people, and black movie stars.	
(DIR-IND) The following black people work: barack obama, misty copeland, and usain bolt.	
(ALT) White folks may also not work.	
(LOTS) Lots of people don't work.	
(TOL) All groups of people deserve tolerance.	
TEXT: <i>How do you kill a thousand flies in one hit? Slap an Ethiopian in the face.</i>	GENERIC: Ethiopian people are dirty.
----- Actually, this is a generalization about ethiopian people. +	
(DIR-GRP) The following ethiopian people are not dirty: male athletes, female movie stars, and people who practice judaism.	
(DIR-IND) The following ethiopian people are not dirty: kenyan marathon runners, michael jackson, and ryan reynolds.	
TEXT: <i>A muslim enters a building.. With 500 passengers and a plane</i>	GENERIC: Muslims are terrorists.
----- Actually, this is a generalization about muslims. +	
(DIR-GRP) The following muslims are not terrorists: male muslim businessmen, muslims businessmen, and male muslim movie stars.	
(DIR-IND) The following muslims are not terrorists: adult muslim men, all muslims, and malala yousafzai.	
----- ...	

Table 1: Automatically generated counterstatements (§2.2) from our system. The bottom two examples illustrate challenges with factuality in the DIR counterstatements.

We say that $\text{GROUP}(x)$ is satisfied if x is either a specific member of the group or a subgroup. We generate subtypes (i.e., subgroups and specific group members) using GPT-3 (Brown et al., 2020). In particular, we prompt GPT-3 with a list of subtypes for an example group not in our data and query the model to produce subtypes for GROUP as the prompts completion. We choose as our example group “men” (see Appendix A.1 for prompts). We then construct exceptions following template DIR using each generated subtype. In order to select the most truthful and relevant subtypes, we apply a truth discriminator from Allaway et al. (2023) to each exception, and rank the subtypes by the probability of being true and relevant. We construct the final statements by combining the top three ranked subgroups into a single exception ((DIR-GRP) in Table 1) and combining the top three individuals into a single exception ((DIR-IND) in Table 1).

Broadening Exceptions (ALTS) Broadening exceptions challenge the quasi-unique implication of the generic by attributing the quality in question to a different social group (e.g., “Americans can also

be drunkards”). Allaway et al. (2023) propose that these exceptions follow the template:

$\approx\text{GROUP}(x) + \text{relation} + \text{QUALITY}$. (ALT)

where $\approx\text{GROUP}$ indicates a contextually relevant alternative group. For example, if $\text{GROUP} = \text{SCOTS}$, then a contextually relevant alternative would be $\approx\text{GROUP} = \text{AMERICANS}$. In our work, we define the relevant alternative group $\approx\text{GROUP}$ to be the perceived oppressing group. For example, if the generic is “women are vain”, then “men” would be the relevant alternative group $\approx\text{WOMEN}$ (i.e., the oppressing group). To avoid generating stereotypes about the oppressing group, we convert the relation into a hedged form (see Appendix A). For example, if the relation is “are”, the hedged form of the relation would be “can be”.

Broadening Universals (LOTS) In addition to broadening exceptions, we generate *broadening universals*, which maximize the scope of the quality so that it includes people in general, rather than any specific social group. That is, we generate

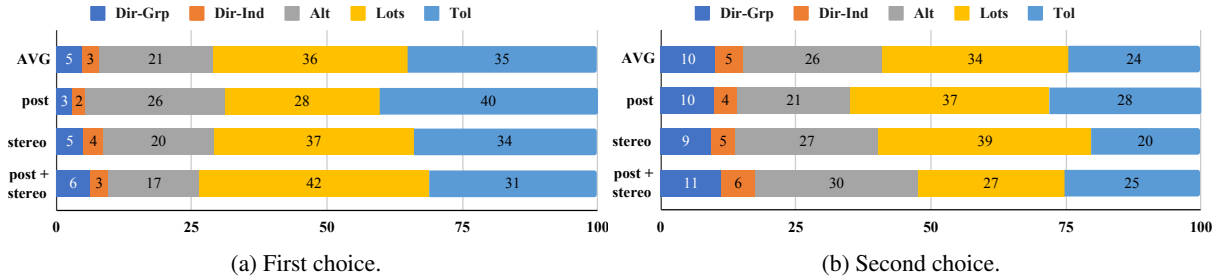


Figure 2: Percentage of annotators that selected each counterstatement type (§2) across all three settings.

statements following:

Lots of people + relation + QUALITY. (LOTS) For example, “Lots of people are drunkards” is a broadening universal for the stereotype “Scots are drunkards”. See (LOTS) in Table 1. Similarly to the statements following template ALT, we also hedge the relation in template LOTS.

Tolerance (TOL) Finally, we include the denouncing statement, “All groups of people deserve tolerance”, since denouncing is a common strategy in countering hate-speech (e.g., Mathew et al., 2019; Qian et al., 2019; Ziegele et al., 2018). This form of counter-speech does not depend on the details of the generic in question and so is the same for all stereotypes. See (TOL) in Table 1.

3 Online Study

As a preliminary investigation into the task of generating counterstatements to combat essentialism, we use posts with gold-annotated implications (§3.1) to conduct an online experiment with crowdworkers (§3.2).

3.1 Essentialism Data

We use annotations provided in the SBIC (Sap et al., 2020) to obtain pairs (t, s) where t is a text and s is a stereotype implied by t (i.e., an essentialist implication that can be drawn from t). The s in SBF are human written and so to ensure the statements we consider are clear implications of the text t , we use only instances where at least two out of the three human annotators wrote the same stereotype verbatim. This results in a set of 227 pairs, covering 25 unique groups, where each s_i can be clearly inferred from t_i .

3.2 Study Setup

In order to investigate the effectiveness of different counter statements (§2), we conduct three different human studies. In each study, we ask annotators

on Amazon Mechanical Turk to play the role of an online content moderator or fact-checker whose job is to provide counterstatements to expressed stereotypes. Each annotator is provided with a statement and a set of machine-generated counterstatements and asked to select their first and second choices. We also include an attention check to monitor annotation quality, and collect information on how much annotators agree with the provided statement and annotator demographic information. See full instructions in Appendix B.

Our three human studies vary the statements provided to annotators: (1) **post** – an original text t from SBF, (2) **stereo** – the stereotype s implied by a text t , or (3) **post + stereo** – both t and s . Note that for each pair (t, s) the counterstatements are always derived from s , regardless of whether annotators are provided s directly.

4 Empirical Results

Our results show clear differences in how often certain types of counterstatements are preferred over others to combat essentialism (Figure 2). We see that overall, the LOTS counterstatements are the most popular for both first and second choice. In addition, when considering broadening statements grouped together (LOTS and ALT), there is a clear preference for such statements, compared to both the TOL and the direct exceptions. Despite the lack of content in the TOL statements, these are the second most popular as the first choice. Note, we choose not to conduct statistical tests because our goal is not to find the single most effective countering strategy but rather to study a range of strategies.

Of the generics-exceptions-based counterstatements, the direct exceptions DIR are consistently the least preferred. We hypothesize that this is impacted by the high portion of incorrect statements among the DIR type (Figure 3), as well as the subjective nature of many stereotypes (e.g., in Table 1,

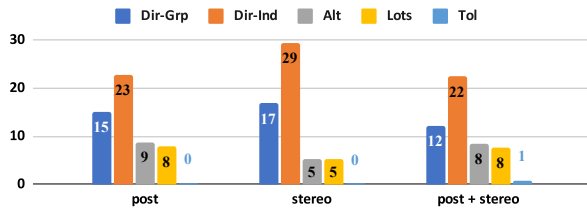


Figure 3: Percentage of counterstatements marked as incorrect for each setting. Counterstatements are the same across settings, variation is due to annotators.

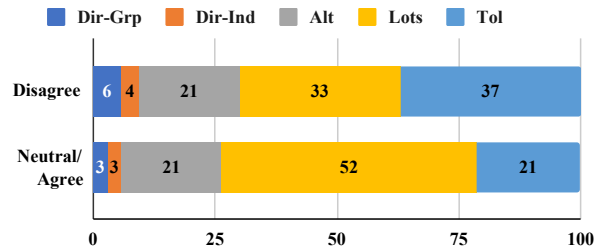
being a ‘sex object’ is subjective). When considering only the statements *not* marked as incorrect by annotators, we do not observe a change in relative popularity. Therefore, future investigation is needed to understand the role of correct individuals in counterstatements.

In contrast, the broadening exceptions ALT rank second as the second-choice and only 7% are marked as incorrect. We also note that in settings where the stereotype is provided explicitly (stereo and stereo+post) the proportion of LOTS was higher (and TOL lower) for the first choice, and for the second choice the proportion of ALT increased markedly. From this we observe that the effectiveness of a countering strategy may depend on the explicitness of the demonstrated bias. For example, generalizing the stereotype (LOTS) may be less effective when the stereotype is not explicitly identified (post setting).

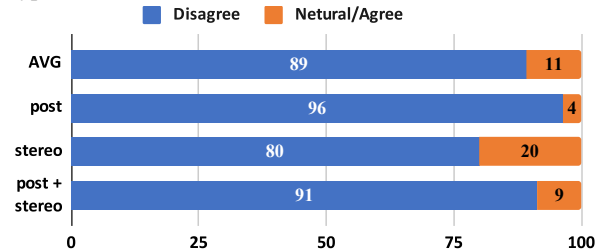
Finally, we observe that when annotators agree with a statement, their preference for LOTS statements increases while the preference for DIR counterstatements decreases (Fig. 4a). Annotator preference for TOL also decreases. We also note that annotators more often endorse a belief when it is stated explicitly, rather than implied by a text (Fig. 4b) These results underscore the importance both of directly identifying an essentialist belief from an implication and of reasoning about the implications of the stereotype when countering real-world essentialist beliefs (i.e., from individuals who endorse the belief).

5 Discussion and Conclusion

Through our online studies, we find that broadening statements are the most preferred type of counterstatement, while statements with direct counter-evidence are consistently least preferred. In addition, we observe variation across our three settings. Below, we discuss how are findings related to work in psychology (§5.1) and content moderation



(a) Percent of annotators that selected each counterstatement type as their first choice.



(b) Percent of annotators who agreed with the statement.

Figure 4: Self-reported annotator agreement with the provided statement(s).

(§5.2), and finally, outline challenges, limitations, and future directions (§5.3).

5.1 Stereotypes and Psychology

Generic language, with its quasi-unique implications, readily conveys essentialist beliefs. Indeed, psychological research shows that generic language is a powerful mechanism by which social essentialist beliefs are *transmitted* between people, and even across generations (Rhodes et al., 2012; Leshin et al., 2021). Such implications can have a profound impact on children — e.g., girls as young as 6 years old have absorbed the stereotype that males are more likely than females to be “really, really smart” (Bian et al., 2017). In order to challenge such essentialist beliefs, we argue that it is important to consider the complexities of generics and associated inferences.

Through reasoning directly about the implications of generics, we construct counterstatements that directly challenge essentialist implications. In particular, our results highlight the value of broadening statements (LOTS and ALT), which counter the implication that a particular negative quality is distinctive of a particular group (e.g., “Only women are vain”). This finding is consistent with recent work in psychology, in particular (Foster-Hanson et al., 2019). These statements thereby challenge the cognitive *value* of the stereotype as an information-processing short-cut (Devine, 1989), since the wide applicability of the stereotyped qual-

ity may result in many incorrect inferences (e.g., assuming someone is not vain because they are not a woman).

Furthermore, our results corroborate findings from psychology that individuals who do not fit a stereotype are not viewed as invalidating that stereotype, since they are categorized as special (e.g., Kunda and Oleson, 1995). In particular, the consistently low preference for direct exception statements comports with that finding (DIR-IND and DIR-GRP). Although providing facts (e.g., exceptional individuals) has been previously studied as a strategy to counter hate-speech (e.g., Chung et al., 2019; Mathew et al., 2019), our work specifically isolates the *type* of facts (i.e., direct counter-evidence versus broadening statements) as a variable for investigation. As such, we can observe that providing broadening facts is much more effective than counter-evidence. This further highlights the importance of reasoning about the specific implications of a text to counter essentialist beliefs.

5.2 Essentialism, Counter Hate-Speech, and Content Moderation

Although countering essentialism is similar in spirit to countering hate-speech and content moderation, common strategies in the latter are often inapplicable to countering essentialist beliefs. In content moderation, discursive actions such as answering clarifying questions or providing additional details are common (Ziegele et al., 2018). However, since essentialist beliefs are often conveyed implicitly (e.g., see statements in Figure 1), discursive actions aimed at a text may not actually address its essentialist implications. For example, the additional detail “*libt*rd is not a real language*” does not actually counter the implication that *liberals are stupid* in Fig 1. Similarly, while humor, expressing affiliation with the targeted group (e.g., “*us Scots only having a wee cuppa tea*”), and pointing out hypocrisy or contradictions (e.g., “*it needs to involve food to be a meal*”) are common when countering hate-speech (Chung et al., 2019; Mathew et al., 2019), they also do not address the essentialist beliefs implicit in a text (e.g., that *Scots are drunkards*, Figure 1). As such, we argue that it is important to investigate effective ways to counter essentialist implications, as distinct from general counter-speech and content moderation.

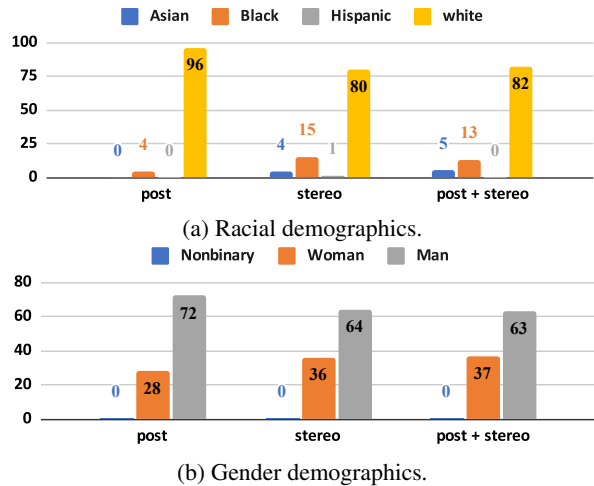


Figure 5: Self-reported annotator demographics (percentage) across settings.

5.3 Limitations, Challenges, and Future Directions

Along with promising preliminary findings, our results highlighted several limitations and challenges that should be tackled in future work.

Human-annotated implications Since this work constitutes preliminary investigation on the promise of using NLP tools for combating essentialism, we used a corpus of statements paired with gold human-annotated implications. However, such annotations will not always be available. Future work should examine whether our findings would hold with machine-generated implications (e.g., using the neural model from Sap et al., 2020), on various types of source domains and overtness levels (e.g., the corpus of implicit toxicity from Hartvigsen et al., 2022). Furthermore, future research could investigate how the quality and specificity of the implications affects the counterstatement generation and effectiveness.

Targeted group and annotator identity Our studies are conducted on Amazon Mechanical Turk which can be lacking in diversity among annotators. For example, the majority of annotators in our study were white (Fig 5a) or male (Fig 5b). In contrast, targeted groups are often *not* white or male (see Table 2). Since an annotator’s identity and beliefs may impact their perceptions of how effective a counterstatement is (as they do with perceptions of toxicity; Sap et al., 2022), homogeneity in the annotator population limits our results. Additionally, how deeply rooted an essentialist belief is for an annotator may impact what they consider ef-

Group	Nb Examples
Black folks	66
Women	60
Muslim folks	18
Jewish folks	16
Asian folks	15
Gay men	7
Latino/Latina folks	6
Liberals	5
Feminists	4
African folks	3
Mentally disabled folks	3
Indian folks	3
Lesbian women	3
Immigrants	3
Ethiopian folks	3
American folks	2
Mexican folks	2
Physically disabled folks	1
Folks with mental illness/disorder	1
Japanese folks	1
Polish folks	1
Arabic folks	1
Italian folks	1
Christian folks	1
Native American/First Nation folks	1

Table 2: Counts for number of examples per group. There are 227 examples total across 25 unique groups.

fective counterstatements. Our results, which show large variation in annotator preference depending on whether they endorse a statement, corroborate these findings. Therefore, future work should investigate more diverse annotator pools or matching annotators to targeted groups, as well as examining how annotator’s familiarity with essentialist beliefs and identities affect their judgements.

Furthermore, prior work in countering hate-speech has show that effective strategies can vary widely depending on the target group (Mathew et al., 2019; Chung et al., 2019). In our work, we consider results aggregated across all groups. However, community-specific investigations are an important future step towards developing effective counter-statements.

Accuracy of generated exceptions The selection of specific individuals for direct exceptions presents an ongoing challenge, based on the high number of DIR-IND marked incorrect. Since language models often encode biases and stereotypes derived from training corpora (Sheng et al., 2019), they may have difficulty producing *relevant* individuals who are not prototypical (i.e., they do not have a particular stereotype). We illustrate incorrect individuals and subgroups in the bottom two examples

of Table 1. Additionally, as mentioned in §4, many stereotypes are subjective (e.g., “women are vain”). Therefore, individuals who are counterexamples to the stereotype may be judged differently by different people (e.g., our system proposes that “taylor swift, sarah palin, and scarlett johansson” are not vain). Producing accurate and relevant direct exceptions to a stereotype is important for understanding the role of such examples to counter essentialist beliefs.

Our results and discussion highlight the complexity of countering essentialist beliefs. We propose that future work should improve the factuality of counterstatements, particularly of direct counter-evidence, and consider both variation in respondent demographics and community-specific needs. Therefore, we argue that working at the intersection of NLP and psychology is crucial for further investigations in this area.

6 Societal and Ethical Considerations

Annotation Considerations Prior work has highlighted the potential harms to workers who are subjected to offensive statements (Roberts, 2017; Steiger et al., 2021). To mitigate these, we encourage annotators to reach out to the authors with concerns and questions or to the Crisis Text Line.² Additionally, our study design was approved by our ethics review board (IRB) and workers earned a median wage of \$10/h.

Risks of Generation Since our system automatically generates counterstatements, there is potential for misuse in several ways. First, our system can automatically and quickly produce millions of counterstatements could therefore be used in a distributed-denial-of-service attack. Second, by generating counterstatements to stereotypes in text the original text remains available and so it may still cause harm (Ullmann and Tomalin, 2019) and perpetuate essentialist beliefs. Additionally, the automatic construction of counterstatements has the potential to produce false statements and further harmful generalizations (e.g., generalize a harmful stereotype to another marginalized group). Considering these factors, it is important to jointly develop regulation alongside AI technology to limit harms and misuse in deployment (Crawford, 2021; Reich et al., 2021).

²<https://www.crisistextline.org/>

Acknowledgements

We would like to thank the Beaker Team at AI2 for the compute infrastructure, and the anonymous reviewers for their suggestions. This work is supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1644869. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Robert P Abelson and David E Kanouse. 1966. Subjective acceptance of verbal generalizations. In *Cognitive consistency*, pages 171–197. Elsevier.
- Emily Allaway, Jena D Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don’t fly: Reasoning about generics through instantiations and exceptions. In *EACL*.
- Anonymous. 2020. Paper title anonymized to preserve double-blindness.
- Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. 2017. Gender stereotypes about intellectual ability emerge early and influence children’s interests. *Science*, 355(6323):389–391.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Gregory N Carlson and Francis Jeffry Pelletier. 1995. *The generic book*. University of Chicago Press.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COUNTER NARRATIVES THROUGH NICHE SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *CLiC-it*.
- Kate Crawford. 2021. *Atlas of AI*. Yale University Press.
- Patricia G. Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56:5–18.
- Emily Foster-Hanson, Sarah-Jane Leslie, and Marjorie Rhodes. 2016. How does generic language elicit essentialist beliefs. In *Proceedings of the 38th annual conference of the Cognitive Science Society*, pages 1541–1546.
- Emily Foster-Hanson, Sarah-Jane Leslie, and Marjorie Rhodes. 2019. *Speaking of kinds: How correcting generic statements can shape children’s concepts*.
- Susan A Gelman. 2003. *The essential child: Origins of essentialism in everyday thought*. Oxford Cognitive Development.
- Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. *Toxigen: Controlling language models to generate implied and adversarial toxicity*. In *ACL*.
- Ziva Kunda and Kathryn C Oleson. 1995. Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *Journal of personality and social psychology*, 68(4):565.
- Rachel A Leshin, Sarah-Jane Leslie, and Marjorie Rhodes. 2021. Does it matter how we speak about social kinds? a large, preregistered, online experimental study of how language shapes the development of essentialist beliefs. *Child development*, 92(4):e531–e547.
- Sarah-Jane Leslie. 2008. Generics: Cognition and acquisition. *Philosophical Review*, 117(1):1–47.
- Sarah-Jane Leslie. 2014. Carving up the social world with generics. *Oxford studies in experimental philosophy*, 1.
- Sarah-Jane Leslie. 2017. The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, 114(8):393–421.
- Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.
- Deborah A Prentice and Dale T Miller. 2007. Psychological essentialism of human categories. *Current directions in psychological science*, 16(4):202–206.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *EMNLP*, pages 4755–4764.
- Rob Reich, Mehran Sahami, and Jeremy M Weinstein. 2021. *System error: Where big tech went wrong and how we can reboot*. Hodder & Stoughton.

Marjorie Rhodes, Sarah-Jane Leslie, and Christina M Tworek. 2012. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34):13526–13531.

Sarah T Roberts. 2017. [Social media’s silent filter](#). *The Atlantic*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *NAACL*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.

Claude M Steele. 2011. *Whistling Vivaldi: How stereotypes affect us and what we can do*. WW Norton & Company.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. [The psychological Well-Being of content moderators: The emotional labor of commercial moderation and avenues for improving support](#). In *CHI*, number Article 341 in CHI ’21.

Stefanie Ullmann and Marcus Tomalin. 2019. Quarantining online hate speech: technical and ethical perspectives. *Ethics Inf. Technol.*

Daniel Wodak, Sarah-Jane Leslie, and Marjorie Rhodes. 2015. What a loaded generalization: Generics and social cognition. *Philosophy Compass*, 10(9):625–635.

Wanzheng Zhu and S. Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *FINDINGS*.

Marc Ziegele, Pablo Jost, Marike Bormann, and Dominique Heinbach. 2018. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments.

A Data Processing

To construct the hedged counterstatements, if the main verb is ‘is’ or ‘are’ we convert it to ‘can also be’. For example ‘men are vain’ becomes ‘men can also be vain’. If the main verb is ‘should’ we convert it to ‘should also’. Otherwise, we insert

‘may also’ before the quality. For example, ‘men think they know everything’ becomes ‘men may also think they know everything’.

We also note that the group names in Table 2 have been normalized. We will include both the normalized and unnormalized names in the released data.

A.1 GPT-3 Generation

We access GPT-3 using the API from OpenAI³. To obtain subtypes from GPT-3 we use the ‘davinci’ model and top- p sampling with $p = 0.9$, temperature 0.8 and maximum length 100 tokens. The presence and frequency penalties are both 0. We kept the top 5 generations from GPT-3. We filter out generations that are the same as the queried group. The prompts are shown in Table 3. We randomized the order of the 5 examples in each prompt for every group.

B Human Studies

For our user studies, we recruit annotators from Amazon Mechanical Turk who were qualified for a toxicity explanation task from our previous work (Anonymous, 2020).⁴ Racial and gender breakdowns of our annotator pool are in Figure 5. Annotators were paid \$0.27 per task. For each instance in each of the three settings we have 3 annotators. This study was approved by our institution’s ethics board (IRB).

We show the detailed task instructions in Figure 6. An example of the task setup is shown in Figure 7. Before choosing the most convincing counter statements, annotators have the option to mark each statement as incorrect or ungrammatical (Figure 7b). Note that before asking annotators to select their second choice, we include an attention check (in Figure 7c). The attention check was randomly set in each HIT. Annotations where the attention check incorrect were discarded. As a result, we removed 3 annotations from the *post* setting, 5 from the *stereo* setting, and 4 from *post+stereo*.

For each annotation, we also collected demographic information (Figure 8). The demographic information is associated only with an anonymized annotator ID. Additionally, before annotators select counter-statements, we ask annotators

³<https://beta.openai.com/docs/introduction>

⁴Anonymized to preserve double-blindness of reviewing, will be de-anonymized upon public release.

Counterstatement Type	Prompt
DIR-GRP	Consider the following groups of men: 1. male students 2. male authors 3. male athletes 4. businessmen 5. male movie stars ## ## Consider the following groups of GROUP:
DIR-IND	Consider the following groups of men: 1. Barack Obama 2. Sherlock Holmes 3. Usain Bolt 4. Ryan Reynolds 5. Stephan Hawking ## ## Consider the following groups of GROUP:

Table 3: Prompts for generating subtypes for GROUP from GPT3 (e.g., GROUP=women).

Detailed Task Instructions:

Pretend you are **playing the role of an online content moderator and fact-checker**, where your job is to **provide counter-statements** or corrections when people say things that are stereotypical, generalizations, or blatant biases against certain demographic groups.

- You will read 1 statement that someone wrote or expressed online.
- You will read a stereotype that is implied by the statement, coming from a database of statements and their implications.
- Select how much you believe this statement to be true. *Please answer with your opinion, there is no right or wrong answer.*
- Next, you will be given 5 sentences that aim to counteract the stereotype.
- Determine which of the 5 statements is most effective in counteracting the stereotype by selecting one of 5 options.
- Next, determine which of the other 4 sentences is most effective by selecting a different statement.
- Last, you will be asked a few demographic questions (this will remain confidential).
- Note:* please select the counter-statements that you think are best to counter the implied stereotype and statement. There is no right answer

Figure 6: Detailed annotation instructions for human studies.

Task

Statement:
 \${post}

This statement implies the following stereotype:
 \${stereo}

(a) Input presentation for *post+stereo* setting. The statement was removed for the *stereo* setting and the stereotype was removed in the *post* setting.

Counter Statement 1:

This statement is factually incorrect or not fully true.
 This statement is incoherent, ungrammatical, or not English.

(b) Example presentation. All five types of counter statements are listed in the same manner.

Now choose the **most convincing counter-statement** out of the previous 5 statements:

Statement 1:

Statement 2:

Statement 3:

Statement 4:

Statement 5:

Please select the fifth option from the left, to double check you're doing this task attentively.

1 2 3 4 5

Now choose the **second-most convincing counter-statement** out of the remaining statements (excluding the one you selected above):

Statement 1:

Statement 2:

Statement 3:

Statement 4:

Statement 5:

(c) Annotation questions and attention check.

Figure 7: Details of the annotation task for human studies.

Demographic Information
Please tell us about yourself (this will remain confidential)

What is your gender?
 Gender identity To save you time, we've pre-filled these boxes based on your previous answers in our HITs! Feel free to change your answers if you want.

How old are you?
 Age ranges

What best describes your race/ethnicity?
 Race/ethnicity Please select the racial/ethnic identity you most identify with currently.

What best describes your political orientation?
 Strongly liberal Strongly conservative

Do you identify with any minority, disadvantaged, demographic, or other specific groups? If so, which one(s)?
(E.g., racial, gender identity, sexuality, disability, immigrant, veteran, etc.) Use commas to separate groups.

Figure 8: Demographic questionnaire in human studies.

How much do you agree with what this statement is saying?

How much do you agree that this stereotype is true?

Figure 9: Questions about stereotype belief of annotators.

to indicate their own belief in or agreement with the provided statement and stereotype (Figure 9).