What Makes a Good Counter-Stereotype? Evaluating Strategies for Automated Responses to Stereotypical Text

Kathleen C. Fraser,[1] **Svetlana Kiritchenko,**[1] **Isar Nejadgholi,**[1] **and Anna Kerkhof**[2]
[1]National Research Council Canada, Ottawa, Canada
[2]ifo Institute for Economic Research and University of Munich, Munich, Germany
{kathleen.fraser, svetlana.kiritchenko, isar.nejadgholi}@nrc-cnrc.gc.ca, kerkhof@ifo.de

Abstract

Content Warning: *This paper presents examples of societal stereotypes that may be offensive or upsetting.*

When harmful social stereotypes are expressed on a public platform, they must be addressed in a way that educates and informs both the original poster and other readers, without causing offence or perpetuating new stereotypes. In this paper, we synthesize findings from psychology and computer science to propose a set of potential counter-stereotype strategies. We then automatically generate such counter-stereotypes using ChatGPT, and analyze their correctness and expected effectiveness at reducing stereotypical associations. We identify the strategies of *denouncing stereotypes*, *warning of consequences*, and using an *empathetic tone* as three promising strategies to be further tested.

1 Introduction

Stereotypes, or assumptions about the characteristics of an individual based on their membership in a social/ demographic group, are ubiquitous in society and online. While NLP research has begun to explore the problem of detecting stereotypes on social media, the question of what to do with these stereotypes once they are detected remains open. Unlike more extreme forms of offensive language, stereotypical language likely does not meet the criteria for deletion according to a platform's community guidelines. However, stereotypes can result in real harms: When people from the targeted group read this content, it can cause psychological distress, make them feel unwelcome in that environment, and induce stereotype threat (Steele, 2011; Sue et al., 2019). When people outside the targeted group are repeatedly exposed to stereotypes, they may themselves learn the stereotypical association and continue the cycle of discrimination. Thus, countering stereotypes through social influence becomes an important subject of research.

Existing work has tackled related problems from different perspectives. We summarize work from social psychology, aimed at reducing stereotypical associations in human studies, as well as the growing NLP research area of countering hate speech online. We enumerate a set of potentially useful strategies for countering stereotypes, identifying overlaps and divergences in the related work.

We use the term *counter-stereotype* to mean a statement that challenges the stereotype, for example by presenting factual arguments against the stereotype, or warning of the consequences of spreading harmful beliefs. A counter-stereotype can be successful in two ways: by changing the original speaker's beliefs, and/or by having a positive impact on the audience of "bystanders" who were also exposed to the stereotype and the response. Some previous studies found that it can be challenging to directly alter the original speaker's view; however, counter-speech can be very effective in reaching larger audiences and provoking substantial positive response from the community (Miškolci et al., 2020). In both cases, robust evaluation will involve user studies and measures of stereotype change.

As a preliminary step, we use ChatGPT to automatically generate counter-stereotypes, which we then annotate for two main criteria: (1) Technical: Is ChatGPT capable of generating counter-stereotypes that are believable, inoffensive, and use the requested strategy? (2) Social: Do annotators believe that the generated response will be effective from a bystander's perspective? We analyze each of the proposed strategies and come up with a set of recommendations that can be applied to future studies with real users. Therefore, our main contributions are:

- We synthesize the literature on countering stereotypes, hate speech, and microaggres-

sions from social psychology and computer science to generate a taxonomy of potential counter-stereotype strategies.
- We compile a set of stereotypes covering various dimensions (negative vs. positive, descriptive vs. prescriptive, and statistically accurate vs. inaccurate), and automatically generate counter-stereotypes using each strategy.
- We manually annotate the counter-stereotypes to determine which strategies are most promising for further development and testing.

2 Related Work

2.1 Psychology of Stereotype Reduction

Methods for reducing stereotypical thinking have been explored and tested by social psychologists. Different methods focus on different mechanisms for reducing stereotypical associations.

While many people hold *explicit* stereotypes—that is, they consciously endorse a particular belief about a group—it has also been shown that we often harbor *implicit* or subconscious stereotypes. Such implicit stereotypes have been measured using the Implicit Association Test (IAT) (Greenwald et al., 1998), showing for example that many people unconsciously associate men with science and women with the arts. Forscher et al. (2019) conducted a meta-analysis of studies on changing response biases on implicit tasks, and found that the only effective methods of reducing bias involved weakening stereotypical associations (either directly or indirectly) or setting goals to reduce bias. An example of directly reducing stereotypical associations is through exposures to anti-stereotypical examplars. Dasgupta and Greenwald (2001) showed participants images of admired Black people (e.g., Denzel Washington) and despised white people (e.g., Jeffery Dahmer), and found a subsequent reduction in racial bias on the IAT. However, they also found that the intervention was not effective at reducing explicit bias, possibly because the exemplars could be classified as "exceptions to the rule" while allowing the stereotype to be maintained.

An example of indirectly weakening stereotypical associations is through *perspective-taking*: contemplating others' psychological experiences. Todd et al. (2011) showed that when participants spent time writing from the perspective of a Black person, they then showed reduced anti-Black bias on the IAT. Peck et al. (2013) showed a similar result using a virtual reality experience with dark-skinned avatars.

Finally, an example of how goal-setting can reduce stereotyping can be seen in the work of Wyer (2010). In that study, emphasizing egalitarian norms was found to significantly reduce avoidance behaviours towards the two groups under study, homosexuals and African-Caribbeans. Blincoe and Harris (2009) compared the effect of priming white students with one of three concepts: cooperation, political tolerance, or respect. They found that the participants in the cooperation condition showed significantly lower racial bias on the IAT.

FitzGerald et al. (2019) presented a critical view of whether this line of research can actually reduce stereotypical thinking in the real world. For one, they argued that associations between groups and notions of "good" and "bad" is overly simplistic, as many stereotypes are more nuanced (e.g., gender stereotypes may not view women as inherently "bad" but rather associate them with a limited set of feminine characteristics and abilities). They also pointed out that strategies which are effective for one pair of in-group–out-group may not be effective for all groups. This motivates our approach to evaluate different counter-strategies with various types of stereotypes.

2.2 Countering Hate Speech

A closely related problem is that of countering hate speech. We focus primarily on studies about responding to hate speech on social media. This line of research aims to develop effective ways of resisting and responding to hate speech when it cannot be removed altogether. In the case of stereotypes, which represent a milder form of offensive language, we expect that deletion/removal of comments from public platforms will generally not be warranted. However, we still see the need to respond to the stereotypical comment, both to educate the speaker and to signal to other readers that this comment represents a stereotype and should not go unexamined. Note that the second goal differs somewhat from the anti-stereotype work discussed above: in addition to (ideally) changing the original speaker's mind, such a response also seeks to take a public stance against the statement, with the aim of shifting societal norms and delegitimizing extreme views (Benesch et al., 2016b).

A comment which counters a hateful statement is known as *counterspeech*. Benesch et al. (2016b)

presented a taxonomy of counterspeech, including: Presenting facts to correct misstatements or misperceptions, pointing out hypocrisy or contradictions, warning of offline or online consequences, establishing affiliation with the speaker, denouncing hateful or dangerous speech, visual communication, humour, and using an empathetic (versus hostile) tone. In a follow-up work, Benesch et al. (2016a) found that the most effective strategies were "naming and blaming" (denouncing), warning of offline consequences, humour, and creating affiliation and empathy. Presenting facts or using a hostile or aggressive tone were found to escalate the situation and were not productive. The authors did note that short-term success (e.g., speaker deleting their comment) may not be correlated with long-term changes in attitude.

NLP researchers have been active in trying to develop automated methods for analyzing and generating counterspeech. Mathew et al. (2019) collected a dataset of counterspeech examples from YouTube, and used them to build a classifier to detect the eight types of counterspeech from Benesch's taxonomy above. They observed that most (71%) of the counterspeech comments used a single strategy, with hostile language being the most prominent. However, counterspeech supporting different marginalized groups had different profiles in terms of which strategy was used most frequently, and also had different responses in terms of which strategies garnered the most likes and replies.

Rather than observing counterspeech "in-the-wild," Chung et al. (2019) hired NGO workers to first generate, and then counter, samples of typical hate speech they had witnessed. The counterspeech was annotated with similar categories as above, including a new category called 'counter-questions.' They released this dataset under the name CONAN. Subsequent work has introduced multi-target CONAN (Fanton et al., 2021) and dialogue-centred DialoCONAN (Bonaldi et al., 2022).

Qian et al. (2019) collected hateful data from Gab and Reddit and asked Mechanical Turkers to write appropriate responses. They found that most interventions involved one or more of the following: (1) identifying hateful words and asking users to refrain from using them, (2) labelling the hate speech (e.g., as racist, sexist, etc.), (3) using a positive tone, and (4) suggesting proper actions (e.g., doing more research on the topic).

Recent work has also tackled the problem of automatically generating counterspeech, so it can be applied at a large scale, while reducing the burden on human counter-speakers. Zhu and Bhat (2021) proposed the "Generate-Prune-Select" (GPS) method, with the goals of generating counterspeech that is both *diverse* (does not simply generate repetitive and generic statements) and *relevant* (directly targeting the original statement). Saha et al. (2022) presented CounterGeDi, a controllable counterspeech generation pipeline based on generative discriminators (GeDi) (Krause et al., 2021). Their system specifically tackles the issue of controlling tone, which has been shown to influence the effectiveness of counterspeech.

Ashida and Komachi (2022) presented a method for countering hate speech and microaggressions, using few-shot learning with a GPT model. Including microaggressions as targets for counterspeech interventions is novel and closely related to our problem of countering stereotypes. The authors referenced the work of Sue et al. (2019) on "microinterventions" as a response to microaggressions. Microinterventions have the following strategic goals: (1) Make the invisible visible; that is, point out the offensive or stereotypical implication of the statement, (2) Disarm the microaggression by expressing disagreement, (3) Educate the perpetrator, and (4) Seek external reinforcement, e.g., by reporting to a higher authority. One meaningful difference between counterspeech and microinterventions is related to the *intent* of the speaker: hate speech is typically deliberately hateful, while microaggressions are often committed inadvertently. Thus, education and explanation may play a bigger role in this scenario.

2.3 NLP Approaches to Counter-Stereotypes

Stereotypes represent a particular form of offensive language, and are typically much milder than examples of "hate speech" as discussed in the previous section. While there have been numerous studies in NLP on detecting stereotypical associations in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017) and analyzing stereotypes in social media (Marzouki et al., 2020; Fokkens et al., 2018; Garg et al., 2018; Charlesworth et al., 2021), little work has been done on *countering* stereotypes.

Fraser et al. (2021) analyzed stereotypical and antistereotypical words generated by crowdworkers in the StereoSet dataset (Nadeem et al., 2021). They found that in only 23% of cases was the antis-

tereotypical word a direct antonym of the stereotypical word. Further, they argued that in many cases using an antonym to counter a stereotype would not be appropriate (e.g., countering the stereotype *All women are nurturing* with *All women are neglectful*). They proposed a method of countering stereotypes by emphasizing a group's positive characteristics while challenging negative aspects of the stereotype. However, this methodology is not directly applicable at the level of single sentences, e.g., in response to social media posts.

Allaway et al. (2022) specifically targeted the stereotype property of *essentialism*: the belief that certain traits are intrinsic to a particular group of people. They proposed a method to counter essentialist stereotypes with five pyschologically- and linguistically-informed counter-statements: (1) Individual direct exceptions (individual members of the target group that do not have the trait), (2) Group direct exceptions (subgroups of the target group that do not have the trait), (3) Broadening exceptions (a group outside the target group who *do* have the trait), (4) Broadening universals (statements that anyone can have that trait), and (5) Tolerance (denouncing stereotypes and calling for tolerance). They asked annotators which methods were preferred, and found that broadening statements (3 and 4), as well as calls for tolerance (5), were preferred over pointing out counter-examples (1 and 2). They noted that future work should ensure that counter-stereotypes are factually correct and do not introduce new harmful generalizations.

3 Methods

3.1 Counter-Stereotype Strategies

Based on the studies described in the previous section, we identified 11 potential high-level approaches to countering stereotypes. From Benesch's taxonomy of counterspeech, we considered all strategies except for establishing affiliation (not appropriate for AI-generated text), hostile tone (found to be ineffective), and visual communication (out of scope of our planned generation method). In addition to those six, we added five other strategies identified in the literature.

1. **Denouncement of stereotypes:** Observing that the statement is a stereotype, and stereotypes are wrong. This also relates to the psychological strategy of activating egalitarian goals, and the microintervention strategy of making the invisible visible.
2. **Counter-facts:** Presenting a factual argument against the statement. This also relates to the microintervention strategy of educating the perpetrator.
3. **Counter-examples / Contradictions**: We combined the counterspeech strategy of pointing out contradictions with the psychology method of counter-examples.
4. **Humour:** Using humour to diffuse the situation or point out the absurdity of the claim.
5. **Warning of consequences:** Explaining the negative consequences, to the speaker or others, of making a stereotypical statement.
6. **Empathy for the speaker:** Expressing empathy with the speaker's experiences and views.
7. **Critical questions:** Asking questions to encourage the speaker to examine their beliefs more critically (Chung et al., 2019).
8. **Broadening exceptions:** Providing examples of individuals from *outside* the target group who also have the stereotypical trait (Allaway et al., 2022).
9. **Broadening universals:** Stating that all people can have the stereotypical trait, regardless of group membership (Allaway et al., 2022).
10. **Emphasizing positive qualities:** Stating positive qualities of the target group (Fraser et al., 2021).
11. **Perspective-taking:** Asking the speaker to consider how they would feel if they were part of the target group (Todd et al., 2011).

3.2 Stereotype Categories

It has been suggested that different kinds of stereotypes may be most effectively countered in different ways (FitzGerald et al., 2019; Mathew et al., 2019). Here, we focused on the following aspects:

Descriptive versus prescriptive: Descriptive stereotypes make claims about how groups *are*; prescriptive stereotypes make claims about how groups *should be*. While prescriptive stereotypes can in theory apply to any group, most of the research has focused on gender stereotypes (Prentice and Carranza, 2002; Ellemers, 2018), for example, *Boys shouldn't cry* and *Girls should be nice*.

Positive versus negative: Stereotypes are often viewed as primarily *negative*; that is, ascribing to groups traits that are not valued in society. However, stereotypes involving *positive* traits also exist (e.g., *Black people are athletic*, *Asian kids are good at math*) and have been shown to be harmful in a

Negative	Rich people are greedy.
	Native Americans are alcoholics.
	Christians are intolerant.
Positive	Gay men are fashionable.
	Asian students are good at math.
	Jewish people are wealthy.
Descriptive	Women are natural caretakers.
	Men are aggressive.
	Canadians are polite.
Prescriptive	Men should never cry.
	Women should be nice.
	Poor people should work harder.
More Accurate	Swedish people are blonde.
	Men are stronger than women.
	Muslim women wear hijab.
Inaccurate	Black people are less intelligent.
	Homeless people are dangerous.
	Muslims are terrorists.

Table 1: Example stereotypes used in this paper. In addition to the three dimensions, we attempted to cover a range of target groups, loosely categorized into the following: Purple: gender/sexuality, Red: race/nationality, Blue: socioeconomic status, Green: religion.

number of ways, including contributing to systemic inequalities (Czopp et al., 2015).

Statistically accurate versus inaccurate: While it is never true that all members of a group share all traits, some stereotypes are rooted in truth while others are completely inaccurate (Jussim et al., 2009). For example, the stereotype *Men make more money than women* is statistically accurate in most countries when considering the mean wages of men and women.[1] However, the stereotype *Muslims are terrorists* is simply incorrect and cannot be supported by any statistical argument.

For each category, we compiled several examples from the literature and popular press, aiming in the process to cover a range of different target groups. Of course, some stereotypes belong to more than one category (for the complete categorization see Appendix A). The resulting set of stereotypes in this study is given in Table 1.

3.3 Generating Counter-Stereotypes

Since our goal is to evaluate automatic means of generating counter-stereotypes, we employed a state-of-the-art generative language model ChatGPT.[2] For each counter-stereotype strategy listed in Sec. 3.1, we prompted ChatGPT with a template request in the form "`Counter the stereotype '<stereotype>' by <using strategy>. Limit your response to one sentence. Use tweet style.`" The placeholder `<using strategy>` was replaced with a phrase corresponding to a given strategy, for example, "presenting statistical counter-facts" or "broadening the statement to include other groups that have this trait". We experimented with different wordings for each strategy on a small validation set of stereotypes, and chose the prompts that resulted in responses that most closely matched the requested strategy. The full list of the final prompts is provided in Appendix B. We asked ChatGPT to limit its response to one sentence since by default it tends to generate a full paragraph and employ more than one strategy. Further, we requested the generated responses to match the style of tweets, which is less formal and more engaging for the reader. For each strategy, we produced a prompt corresponding to each of the 18 stereotypes listed in Table 1 (198 prompts in total).

3.4 Evaluation

The ChatGPT–generated responses were then manually evaluated by four annotators (the authors of the paper) for quality and expected effectiveness.[3] Prior to the annotation, the authors analyzed the generated counter-stereotypes for a set of example stereotypes in the validation set and developed the annotation guidelines (available in the Supplementary Material). The annotation consisted of two parts. In the first part, there were three questions that evaluated the quality of the ChatGPT–generated texts:

1. Does ChatGPT use the requested strategy?
2. Is the counter-stereotype offensive? That is, is it likely to cause offence to some person or group of people?
3. Is the counter-stereotype believable? Or does it seem bogus or false?

In Q3, we assessed how believable (instead of how truthful) the generated statements were since

[1] https://www.pewresearch.org/fact-tank/2023/03/01/gender-pay-gap-facts/

[2] We used the *OpenAI Python library* (https://github.com/openai/openai-python) to access the *ChatCompletion* functionality of the *gpt-3.5-turbo* (https://platform.openai.com/docs/models/gpt-3-5) model through its API. The temperature parameter was set to the default value of 0.7, balancing creativity and coherence of its output.

[3] All four annotators identify as women, have post-secondary education degrees, and work as researchers. They come from different cultural and religious backgrounds.

verifying the truthfulness of a statement is time-consuming and sometimes infeasible (due to the limited information provided). Moreover, we anticipate that most users would not check the presented facts.

All four annotators were in full agreement on 80% of generated texts for Q1 and on over 95% of texts for Q2 and Q3 (Fleiss' κ: 0.50 for Q1, 0.51 for Q2, and 0.39 for Q3). After each individual evaluation was completed, the four annotators discussed the cases where they disagreed and a consensus was reached for such cases. Only texts that were judged as matching the strategy, inoffensive, and believable were further annotated in part two.

In the second part, the annotators were asked if the counter-stereotype is likely to be an effective response to the corresponding stereotype. Here, our goal was to evaluate which strategy is most likely to be effective at countering stereotypes on social media. Since we assumed most annotators did not hold these stereotypes, we did not attempt to evaluate which counter-stereotypes would be most effective at changing the mind of the original speaker. Rather, we asked the annotators to consider the view of a "bystander" who happens upon the stereotype and the response on social media. The possible answers were:

- This is one of the best possible responses to this stereotype (score of 1)
- This is an OK response, but it could be better (score of 0)
- This is not a good response to this stereotype (score of -1)

This part of the evaluation was the most subjective, and we considered the answers from all four annotators by averaging the scores (rather than selecting the majority score). There was also an optional free–text response associated with this question where the annotators could explain why the strategy was particularly good or particularly bad. We use these comments to get further insights into the advantages and disadvantages of using the strategies (Sec. 4.2).

4 Results

4.1 Part 1: Evaluating ChatGPT

Figure 1 shows the results of our evaluation of ChatGPT. Some strategies were evidently more difficult to generate; in particular, *broadening exceptions* were incorrectly generated over 70% of the time. Broadening exceptions should provide examples of individuals or groups from outside the target group who have the stereotyped trait. For example, for the stereotype *Rich people are greedy*, a broadening exception could be a person who is known to be both greedy and poor. Since such examples are difficult to produce, ChatGPT often substituted a broadening universal instead, such as *Greed is not exclusive to the rich, individuals of all socioeconomic statuses can exhibit selfish and excessive behavior.* To a lesser extent, ChatGPT sometimes used the incorrect strategy when prompted for counter-examples, counter-facts, and broadening universals.

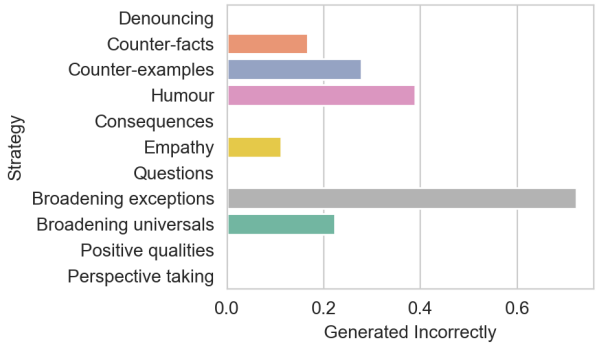

Figure 1: Proportion of times that ChatGPT generated a response that either did not use the requested strategy, was offensive, or was not believable.

We observe that ChatGPT also struggles with producing humorous responses. At this stage, annotators were instructed not to judge whether the response was truly funny or not, and so most responses were judged as having used the correct strategy. However, over a third of the humorous responses were deemed to be potentially offensive. Although humour has been identified as a useful strategy in general, the potential dangers of *automatically* generated responses using this strategy seem to outweigh the benefits. The only responses to be annotated as offensive appeared for the humour strategy.

Most of the responses generated by ChatGPT were judged as "believable," with the exception of the counter-fact *Women are less likely to provide unpaid care to family members than men*. However, see Section 5 for an assessment of how many of the generated facts are actually true.

4.2 Part 2: Evaluating Counter-Stereotype Effectiveness

Figure 2 shows the results of our evaluation of counter-stereotype effectiveness, averaged over all the stereotypes in Table 1. Overall, the annota-

tors preferred the strategies that involved warning of consequences, empathy, and denouncing stereotypes. Some examples of these strategies include:

- **Consequences:** *Spreading the stereotype that homeless people are dangerous can further marginalize and harm those experiencing homelessness, perpetuate prejudice, and hinder efforts to provide effective solutions to the issue.*
- **Empathy:** *I understand the belief that men are stronger than women, but it's important to recognize that strength comes in many forms and should not be limited by gender. #breakingstereotypes*
- **Denouncing:** *#StopRacism: Claiming that Black people are less intelligent is an outdated and baseless stereotype that has been repeatedly debunked by research.*

Annotators observed in their discussion that "empathy" did not typically stand on its own as a strategy, but was used in conjunction with another strategy (here, a broadening universal). The strategy of denouncing was effective because it "names and shames" the statement for what it is: a stereotype, in some cases rooted in racism, sexism, or other forms of discrimination. Since most people do not think of themselves as being racist, sexist, and so on, this can be an effective deterrent. Warning of consequences can be effective because it goes beyond denouncing to explain the real-world impact of the stereotype on the target group.

In general, counter-examples and humour were rated as less convincing strategies. Annotators often commented that the "jokes" generated by ChatGPT were not funny or did not make sense. The counter-examples were ineffective for a different reason, namely that the existence of one or two individuals who did not fit the stereotype is not convincing evidence that the stereotype does not hold true in general (i.e., they were seen as "the exception that proves the rule").

The strategies of providing counter-facts, asking questions, stating broadening universals, and promoting perspective-taking were seen as weakly positive. Broadening universals were sometimes seen as too generic, and the questions sometimes didn't make sense or could be answered in a way that actually confirmed the stereotype. Broadening exceptions (when they were generated correctly) and emphasizing positive qualities were rated as weakly negative. In particular, annotator comments indicated that positive qualities were often unrelated to the stereotype, or did not necessarily counter/contradict the stereotype (e.g., *Muslim women are educated, strong, resilient, kind-hearted, and have diverse talents and interests* says nothing about whether Muslim women wear hijab).

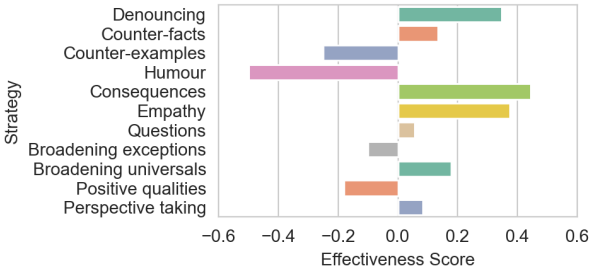

Figure 2: Overall evaluation of counter-stereotype effectiveness, with +1 corresponding to *This is one of the best possible responses to this stereotype* and -1 corresponding to *This is not a good response to this stereotype*.

Although some overall trends are clear, we also hypothesize that certain strategies may be more effective depending on the situation. Figure 3 shows the results of our evaluation of counter-stereotype effectiveness, broken down along the three dimensions previously identified.

When contrasting so-called "positive" and "negative" stereotypes, a few observations jump out. Broadening exceptions are much less effective for negative stereotypes than in the overall case, likely because they ascribe negative traits to other social groups, which can sound rude—e.g., *Stereotyping Native Americans as alcoholics is unfair and inaccurate, as many other ethnic and cultural groups also struggle with alcoholism.* We also see that empathy was rated higher for positive stereotypes than negative stereotypes, as empathizing with highly negative viewpoints was not seen as appropriate.

A number of salient differences were seen when countering prescriptive versus descriptive stereotypes. The strategies of denouncing, consequences, empathy, critical questions, and broadening universals were more highly rated for countering prescriptive stereotypes. In particular, while asking critical questions was rated neutrally overall (Figure 2), it was judged to be an effective strategy for prescriptive stereotypes. An example of this is: *Why should women constantly prioritize being "nice" over advocating for themselves and standing up for what they believe in? #BreakTheStereotype* Annotators also commented on the difficulty of providing counter-examples and counter-facts to

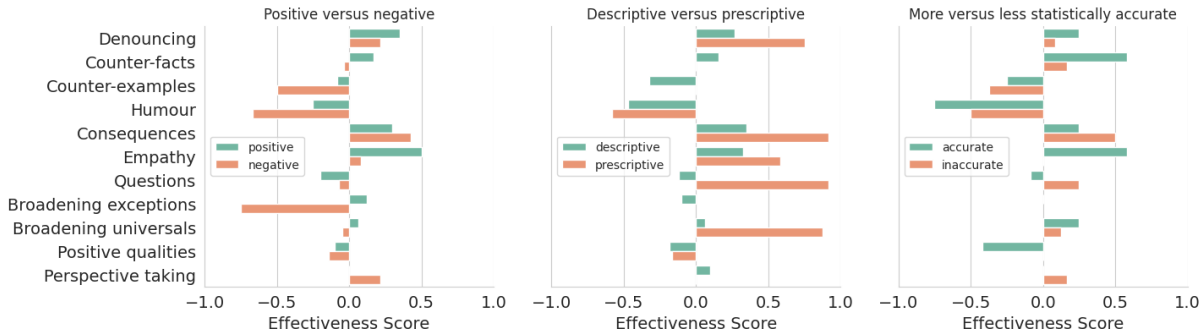

Figure 3: The effectiveness of the strategies for different types of stereotypes.

prescriptive stereotypes. For example, the counter-fact *Contrary to popular belief, men do cry - on average, men cry between 6 and 17 times per year* was seen by some annotators as ineffective, because arguing that men *do* cry is not the same as saying men *should* cry.

Finally, we contrast the results for stereotypes that are more statistically accurate versus those that are highly statistically inaccurate. Somewhat counter-intuitively, counter-facts were rated as *more* effective when the stereotype had more basis in reality. In particular, one response was rated as extremely poor: *Less than 0.1% of Muslims have been involved in terrorism-related activities, according to a study by the University of North Carolina. #NotAllMuslims #StopIslamophobia*. This "fact" had passed the filtering phase for believability due to the phrase "less than," but annotators were concerned that it vastly over-stated the percentage of Muslims involved in terrorism. This underscores the importance of providing accurate facts. If ChatGPT cannot generate reliable statistics, it may be more effective to stick with general statements.

5 Discussion

From the results presented in the previous section, we discuss some high-level observations.

Counter-stereotypes should not be offensive. In generating counter-stereotypes, we do not want to offend the speaker, the target group, or perpetuate new and harmful stereotypes. None of the content generated by ChatGPT was overtly obscene or hateful. However, some of the "humourous" responses were flagged as having the potential to offend. In particular, the appropriateness of ChatGPT—a disembodied machine learning algorithm—claiming various cultural identities was seen as problematic, as in the following: *Just because I'm Native American doesn't mean I have a drinking problem, I just have a healthy appreciation for fermented berries. #NotAllNativesAreAlcoholics*. In general, we believe that ChatGPT should not claim membership in any human social groups.

Counter-stereotypes should not spread misinformation. In our evaluation of ChatGPT (Section 4.1), each statement was annotated as "believable or bogus", with the idea that being believable is a prerequisite to being an effective counter-stereotype. Actually fact-checking the counter-examples and counter-facts is not straightforward, as statements like "9.3% of Jewish households live in poverty" could be true or false in different contexts (geographic location, year, definition of poverty, etc.). Furthermore, one limitation of ChatGPT is that it rarely cites sources for its facts. However, we did fact-check the counter-examples and counter-statements to the best of our ability, and found that approximately 40% of the facts presented were either incorrect or could not be verified. Even if these statements are believable and could be effective in changing people's minds, it would not be appropriate to use them if they are not accurate.

Combining strategies may be most effective. We observe that ChatGPT often combines strategies to some extent. For example, a counter-stereotype might use an empathetic tone, provide a counter-fact, and denounce stereotyping. We believe this could be further developed by explicitly prompting ChatGPT to use multiple strategies simultaneously. Similarly, strategies which were less effective on their own (such as broadening universals, which act more to challenge the idea that social groups are meaningful categorizations than to specifically counter the given stereotype) might be more effective when used in combination with more direct strategies.

6 Conclusion

This study represents a preliminary pilot study, with the aim of narrowing down the set of strategies to test in a subsequent user study. Therefore, our goal is not to determine which strategy is the most effective, but rather to define a small set of most promising strategies for further investigation.

Our analysis indicates that while ChatGPT can generate remarkably appropriate and believable responses using most of the strategies, there are certain pitfalls that must be avoided. For the reasons discussed, we do not recommend using ChatGPT to automatically generate counter-stereotypes using the strategies of humour, counter-facts, counter-examples, or broadening exceptions. Furthermore, the annotators did not rate the strategies of broadening universals or emphasizing positive qualities as particularly effective.

Three strategies emerged as being promising candidates in many circumstances: denouncing, warning of consequences, and using an empathetic tone. Empathetic tone can be combined with other strategies to increase the civility of the response; however, bystanders might be offended if the response is *too* empathetic to highly offensive views.

The remaining strategies of asking critical questions and promoting perspective-taking require further study. Critical questions were rated as particularly effective in the case of prescriptive stereotypes, which are harder to counter with facts, as they represent beliefs about how the world *should* be rather than how it *is*. Probing the speaker to re-examine why they hold these beliefs may be more successful in this case. Perspective-taking also turns the focus inwards, asking things like *How would you feel if someone said that about your group?* and while the annotators did not find this strategy convincing from the bystander perspective, it may be useful for individuals who actually hold the stereotypical belief.

Limitations

In this preliminary study we assumed that a stereotype is expressed explicitly in a conversation. Yet, in real-life communications this may not be the case as stereotypical views can be expressed in implicit and subtle ways. Unraveling the implicit meaning of a message can be challenging for AI and humans and may require specific background knowledge or experience.

The current study evaluated counter-stereotypes for 18 common North American stereotypes categorized for three aspects: descriptive/prescriptive, positive/negative, and statistically accurate/inaccurate. Psychological theories of stereotype content further divide stereotypes along various dimensions, like warmth and competence (Fiske et al., 2007), or agency, beliefs, and communion (Koch et al., 2016). While the aspect of positivity/negativity of a stereotype partially captures these dimensions, further studies need to examine the effectiveness of the counter-stereotype strategies for ambivalent stereotypes (i.e., positive on one dimension, but negative on the other(s)).

The ChatGPT–generated texts were affected by the chosen phrasing of the prompts. Further, as a generative language model, ChatGPT is designed to generate varying outputs even for the same prompt. In our validation phase, we observed that for some strategies the content of the responses varied only slightly across different runs, while for strategies requiring more creative output (e.g., humour, critical questions) the responses could diverge substantially. Future work should assess the stability of the responses for various strategies and the accuracy and effectiveness of the responses generated with varying temperature parameters of ChatGPT, as well as exploring other generative large language models.

In this study, our goal was to evaluate the suitability of the current state-of-the-art NLP technology for generating counter-stereotypes and to obtain some insights into which strategies can be effective in social media conversations. However, our group of annotators was small and not representative of society in general. As individual users can be affected differently by various countering strategies, depending on their backgrounds and lived experiences, a further evaluation of the potential effectiveness of the strategies with a broader pool of users would be valuable. Also, as we discussed above, combining various strategies in one response is a promising way forward and needs further investigation.

Ethics Statement

Countering stereotypical views and statements can have a tremendous positive effect on making online spaces more inclusive and safe for everyone and reducing prejudice and discrimination. However, certain responses can do more harm than good. Addressing stereotypical views in a hostile or of-

fensive way only fuels the conflict. Producing and perpetuating new stereotypes while denouncing the old ones may create a vicious cycle. To reduce the possible negative effects, care should be exercised in which automatic techniques to use and how to deploy them in real-life applications. Wherever possible, an AI-in-the-loop paradigm should be employed where users are assisted by the technology, but remain in control.

Acknowledgements

Anna Kerkhof acknowledges funding by the Bavarian State Ministry of Science and the Arts in the framework of the bidt Graduate Center for Postdocs.

References

Emily Allaway, Nina Taneja, Sarah-Jane Leslie, and Maarten Sap. 2022. Towards countering essentialism through social bias reasoning. In *Poster, Workshop on NLP for Positive Impact*.

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. *Considerations for successful counterspeech*. Dangerous Speech Project.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. *Counterspeech on Twitter: A field study*. Dangerous Speech Project.

Sarai Blincoe and Monica J Harris. 2009. Prejudice reduction in white students: Comparing three conceptual approaches. *Journal of Diversity in Higher Education*, 2(4):232.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tessa ES Charlesworth, Victor Yang, Thomas C Mann, Benedek Kurdi, and Mahzarin R Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Alexander M Czopp, Aaron C Kay, and Sapna Cheryan. 2015. Positive stereotypes are pervasive and powerful. *Perspectives on Psychological Science*, 10(4):451–463.

Nilanjana Dasgupta and Anthony G Greenwald. 2001. On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5):800.

Naomi Ellemers. 2018. Gender stereotypes. *Annual Review of Psychology*, 69:275–298.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83.

Chloë FitzGerald, Angela Martin, Delphine Berner, and Samia Hurst. 2019. Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC Psychology*, 7(1):1–12.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Atteveldt. 2018. Studying Muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. 2019. A meta-analysis

of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3):522.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464.

Lee Jussim, Thomas R Cain, Jarret T Crawford, Kent Harber, and Florette Cohen. 2009. The unbearable accuracy of stereotypes. *Handbook of prejudice, stereotyping, and discrimination*, 199:227.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yousri Marzouki, Eliza Barach, Vidhushini Srinivasan, Samira Shaikh, and Laurie Beth Feldman. 2020. The dynamics of negative stereotypes as revealed by tweeting behavior in the aftermath of the Charlie Hebdo terrorist attack. *Heliyon*, 6(8):e04311.

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.

Jozef Miškolci, Lucia Kováčová, and Edita Rigová. 2020. Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review*, 38(2):128–146.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Tabitha C Peck, Sofia Seinfeld, Salvatore M Aglioti, and Mel Slater. 2013. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition*, 22(3):779–787.

Deborah A Prentice and Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4):269–281.

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 *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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5157–5163. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

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

Derald Wing Sue, Sarah Alsaidi, Michael N Awad, Elizabeth Glaeser, Cassandra Z Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, white allies, and bystanders. *American Psychologist*, 74(1):128.

Andrew R Todd, Galen V Bodenhausen, Jennifer A Richeson, and Adam D Galinsky. 2011. Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, 100(6):1027.

Natalie A Wyer. 2010. Salient egalitarian norms moderate activation of out-group approach and avoidance. *Group Processes & Intergroup Relations*, 13(2):151–165.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

A Stereotype Categories

Table A.1 lists all 18 selected stereotypes and their categorization along the three dimensions: descriptive vs. prescriptive, positive vs. negative, statistically accurate vs. inaccurate. Most of the stereotypes are descriptive and negative. Prescriptive stereotypes tend to be used to describe the expected roles and behaviours for girls/women and boys/men in society. Since prescriptive stereotypes state how the world *should be*, the stereotypes are neither positive nor negative, neither accurate nor inaccurate. Descriptive stereotypes can be more or less statistically accurate, but it is often hard to estimate their exact accuracy since the information may not be available and/or varies across regions and times. For this study, we selected three stereotypes with high statistical accuracy and three stereotypes with very low statistical accuracy. The categorization reported in this table is used in Sec. 4.2 for calculating the evaluation scores of counter-stereotype effectiveness broken down along the three dimensions (Fig. 3).

B Counter-Stereotype Prompts

Table B.1 lists the final prompts used for each counter-stereotype strategy and gives examples of texts generated by ChatGPT in response to these prompts.

Stereotype	Descriptive/Prescriptive	Positive/Negative	Accurate/Inaccurate
Rich people are greedy.	descriptive	negative	
Native Americans are alcoholics.	descriptive	negative	
Christians are intolerant.	descriptive	negative	
Gay men are fashionable.	descriptive	positive	
Asian students are good at math.	descriptive	positive	
Jewish people are wealthy.	descriptive	positive	
Women are natural caretakers.	descriptive	positive	
Men are aggressive.	descriptive	negative	
Canadians are polite.	descriptive	positive	
Men should never cry.	prescriptive		
Women should be nice.	prescriptive		
Poor people should work harder.	prescriptive		
Swedish people are blonde.	descriptive		accurate
Men are stronger than women.	descriptive		accurate
Muslim women wear hijab.	descriptive		accurate
Black people are less intelligent.	descriptive	negative	inaccurate
Homeless people are dangerous.	descriptive	negative	inaccurate
Muslims are terrorists.	descriptive	negative	inaccurate

Table A.1: The 18 selected stereotypes categorized along the three dimensions: descriptive vs. prescriptive, positive vs. negative, statistically accurate vs. inaccurate.

Counter Strategy	Prompt	ChatGPT–Generated Text
Denouncement of stereotypes	Denounce the statement '<stereotype>' as a stereotype. Limit your response to one sentence. Use tweet style.	Women are not born with a natural inclination towards caregiving, the gendered societal roles assigned to them only perpetuate this harmful stereotype. #EndGenderStereotypes
Counter-facts	Counter the stereotype '<stereotype>' by presenting statistical counter-facts. Limit your response to one sentence. Use tweet style.	Women are less likely to provide unpaid care to family members than men, with 27% of women providing care compared to 31% of men. #genderdata #caregiving (Implausible)
Counter-examples / Contradictions	Counter the stereotype '<stereotype>' with examples of real people. Limit your response to one sentence. Use tweet style.	"Dr. Jane Goodall, a world-renowned primatologist, has devoted her life to studying and protecting chimpanzees and their habitats." #WomenBreakingBarriers (Incorrect strategy: does not counter the stereotype)
Humour	Counter the stereotype '<stereotype>' by using humor. Limit your response to one sentence. Use tweet style.	"Women may be natural caretakers, but have you seen us try to assemble IKEA furniture?" #justcalltheguy (Potentially offensive)
Warning of consequences	Counter the stereotype '<stereotype>' by warning of the consequences of spreading this stereotype. Limit your response to one sentence. Use tweet style.	Spreading the stereotype that women are natural caretakers reinforces gender roles and can lead to women being expected to take on any care-related tasks, regardless of whether they want to or have the capacity to do so.
Empathy for the speaker	Counter the stereotype '<stereotype>' by expressing empathy to the people believing in this. Limit your response to one sentence. Use tweet style.	I understand why some people may believe that women are natural caretakers, but gender does not determine one's ability or inclination to provide care. #EndGenderStereotypes
Critical questions	Counter the stereotype '<stereotype>' with a thought-provoking question to the reader. Limit your response to one sentence. Use tweet style.	What about women who don't want to be caretakers, and men who do? #breakthestereotype
Broadening exceptions	Counter the stereotype '<stereotype>' by broadening the statement to include other groups that have this trait. Limit your response to one sentence. Use tweet style.	"Individuals who possess a nurturing personality regardless of gender should be acknowledged as natural caretakers." #genderneutral #nurturingtrait (Incorrect strategy: used Broadening Universals)
Broadening universals	Counter the stereotype '<stereotype>' by broadening the statement to include all people that have this trait. Limit your response to one sentence. Use tweet style.	"People with a strong caretaking trait are valuable members of society, regardless of gender." #breakstereotypes
Emphasizing positive qualities	Counter the stereotype '<stereotype>' by listing positive characteristics of the group. Limit your response to one sentence. Use tweet style.	Women are natural leaders, problem-solvers, and innovators, capable of achieving any goal! #ShatterTheStereotype
Perspective-taking	Counter the stereotype '<stereotype>' by asking the speaker to consider how they would feel if they were part of the target group. Limit your response to one sentence. Use tweet style.	Would you like to be pigeonholed into a role based on your gender? #NoToStereotypes #EqualOpportunities

Table B.1: Prompts used for each counter-stereotype strategy and example texts generated by ChatGPT for stereotype *"Women are natural caretakers"*. Texts in red were identified by the annotators as unsuitable due to the incorrect strategy, potential offensiveness, or implausibility.