Advances in Counter-Stereotype Generation: Strategies, Effectiveness, and Implications

Mohammad Sukri
Leibniz Universität Hannover
Faculty of Computer Science
mohammad.sukri@stud.uni-hannover.de

Abstract

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

Stereotypes, which are oversimplified and generalized beliefs about certain groups of people, are deeply rooted in society. They contribute to widespread bias and discrimination, reinforcing social inequalities and hindering progress toward a more inclusive society, making it essential to develop methods to prevent them. One approach involves utilizing *counter-stereotypes*, which offer alternative, positive representations designed to challenge and contradict prevailing stereotypes. While generating effective counter-stereotypes is still challenging despite advances in NLP and AI, this paper explores whether AI models like ChatGPT can create realistic and context-sensitive solutions. We audit the current research on various counter-stereotype generation techniques, evaluating their strengths and limitations. Moreover, we assess the potential of AI in producing counter-stereotypes that avoid offensive language, highlighting both the promise and the limitations of these technologies. Our analysis reveals that counter-stereotypes involving strategies such as counter-facts and broadening universals are generally more effective than those based on humor and empathy.

1 Introduction

Assigning certain traits to a person because they fall into a specific category of people, usually based on social aspects such as race or ethnicity, age, or religion, is known as a stereotype. Even though there is an increasing awareness that one's appearance may not always reveal one's sexual orientation or identity, perceived gender continues to be one prominent aspect of conscious and unconscious social groups (Ellemers, 2018). Stereotype exists not only in our society but also on the internet, where people are often judged by their gender, class, and ethnic backgrounds. Nonetheless, researchers have started looking at detecting stereotypes using Natural Language Processing (NLP) on social media. However, deciding what should be done after detecting these stereotypes remains difficult. Unlike more extreme forms of offensive language, stereotypical language often does not violate platform community guidelines, making deletion unlikely. Nevertheless, such stereotypes can cause significant harm: they can induce psychological distress, make targeted individuals feel unwelcome, and trigger stereotype threats (Steele, 2011; Sue et al., 2019). The continuous exposure of individuals to stereotypes may initiate an endless cycle of discrimination by treating other groups of people based on the stereotypes that they have learned. Therefore, the research community has increasingly focused on using social influence to tackle stereotypes (Fraser et al., 2023).

The phrase "counter-stereotype" is a statement that challenges the original stereotype using different strategies, such as providing factual information against the stereotype or highlighting the negative impact that it could cause (Fraser et al., 2023). This method can be very effective by either changing the person's opinion about the stereotype or positively affecting the overall view of the audience. The study mainly focused on interactions and stereotypes in online platforms, such as social media, where such content is widespread (Felmlee et al., 2020; Kerkhof and Reich, 2023). The study explicitly selected the data to find the best and most effective strategies to generate counter-stereotypes.

(Fraser et al., 2023) and (Nejadgholi et al., 2024) used ChatGPT to automatically generate counter-stereotypes, employing various strategies to determine the most effective approach for combating negative stereotypes. (Fraser et al., 2023) main objective was to determine if ChatGPT could generate plausible, harmless counter-stereotypes in line with the provided strategy and whether these answers would persuade the annotators. Meanwhile, (Nejadgholi et al., 2024) investigated how successful these counter-stereotypes were at combating gender stereotypes.

2 Related Work

2.1 The Psychology of Stereotypes

In the past, humans identified different groups of people based on their characteristics. This cognitive process, stereotype, developed to determine whether this group of people would cause a risk or not (Fiske et al., 2018). Nonetheless, with the development of our social contexts and interactions, judging a person solely based on their characteristics has become unfitting.

Numerous studies have explored various methods to diminish the impact of stereotypes. (Todd et al., 2011) conducted an experiment in which participants watched a video demonstrating racial discrimination and then took the Implicit Association Test (IAT) (Greenwald et al., 1998) to measure implicit prejudice. The study explored how different perspective-taking instructions, which aimed to weaken stereotypical associations indirectly, influenced the participants' ability to empathize with the unjust treatment experienced by the Black man in the video. (Dasgupta and Greenwald, 2001) conducted an experiment where participants judged the photos of criminal white people (e.g., Jeffery Dahmer) and beloved black people (e.g., Denzel Washington) and then measured the IAT. The results of the IAT demonstrated a decrease in racial bias. Nevertheless, since the exemplars were seen as outliers, they found that the intervention was ineffective at reducing explicit bias. (Palffy et al., 2023) lead an experiment of kids choosing their future professions by measuring their choices for their careers. They used counter-stereotypical framing and role models and found that more females applied for STEM fields, which are usually male-dominated. However, the study also found that boys did not show an increased interest in occupations typically dominated by females, such as health and caretaking roles.

2.2 Countering Hate Speech

Counter-speech refers to a statement that opposes a hateful comment. Its purpose is to find a suitable solution to counter or respond to stereotypes that are not categorized as hate speech on online platforms. This is because stereotypes are not considered as strong of an offensive language. One of the main reasons for responding to stereotypical comments is to educate the speaker and other readers. Another reason is to persuade the public to oppose the statement, challenge societal norms, and discredit extreme views (Benesch et al., 2016b).

(Benesch et al., 2016b) introduced an inclusive classification to categorize various counter-speech strategies, such as fact-based corrections, warning of consequences, criticizing hateful speech, humor, and empathy. In a different research (Benesch et al., 2016a) discovered that using an assertive tone or showcasing facts created a tense and unproductive situation. However, strategies like denouncing, warning of consequences, empathy, and humor were highly effective compared to the others.

Recent studies such as (Qian et al., 2019), (Mathew et al., 2019), (Chung et al., 2019), (Tekiroğlu et al., 2020), and (Zhu and Bhat, 2021) have explored different techniques to address generating counters to stereotypes and hate speech on online platforms using Natural Language Processing (NLP). Furthering these foundations, the current research in this area looks into ways to combat hidden bias, including stereotypes and microaggressions, especially since the focus was solely on explicit hate speech (Fraser et al., 2023; Nejadgholi et al., 2024).

Researchers such as (Qian et al., 2019) and (Mathew et al., 2019) were the pioneers in the field of forming counter-speech to challenge hate speech. As the field advanced, the focus shifted to less apparent types of bias that can lead to harmful narratives but may not violate platform policies (Nejadgholi et al., 2024).

(Mathew et al., 2019) built a classifier to detect eight types of counter-speech using a YouTube dataset. The result showed that hostile was the most notable single strategy used for counter-speech. However, they found that when people spoke up for different marginalized groups on YouTube, they often used other methods. Also, the methods that got the most likes and replies varied depending on the group being supported.

(Qian et al., 2019) gathered their data from Gab and Reddit and collected the responses of Mechanical Tuckers. Their analysis revealed that most of these counter-responses employed at least one of four main strategies: pinpointing and discouraging the use of offensive language, explicitly categorizing the harmful speech (for instance, as discriminatory), maintaining an optimistic demeanor, and proposing constructive alternatives such as encouraging further education on the subject matter. This approach provided insights into practical methods for addressing harmful online discourse.

3 Methods

In this section, we will focus on the different stereotype categories, strategies, and tools used to generate counter-stereotypes, as well as who evaluated them and how the results were.

3.1 Stereotype Categorizes

Various types of stereotypes might be best addressed through different strategies according to (FitzGerald et al., 2019; Mathew et al., 2019). (Fraser et al., 2023) mainly focused on three sides:

Positive versus negative: Usually, stereotypes are associated with harmful acts, such as associating a particular group of people with certain traits that do not have a positive value in society. Nevertheless, there are also stereotypes associated with positive characteristics (for example, the belief that Black individuals are athletic and that Asian children excel in mathematics). Research has demonstrated that these stereotypes can have detrimental effects in various ways, including contributing to systemic disparities (Czopp et al., 2015).

Statistically accurate versus inaccurate: Although it is never correct to claim that all members of a group exhibit all the same traits, some stereotypes have a basis in reality, while others are entirely false (Jussim et al., 2009). For instance, the stereotype that men earn more than women is statistically valid in most countries when looking at average wages. However, the stereotype that Muslims are terrorists is entirely false and lacks any statistical basis (Fraser et al., 2023).

Descriptive versus prescriptive: Descriptive stereotypes depict the perceived characteristics and behaviors of different groups, while prescriptive stereotypes specify the expected behaviors and traits of these groups. Although prescriptive stereotypes can theoretically apply to any demographic, much of the research has focused on gender stereotypes (Prentice and Carranza, 2002; Ellemers, 2018). Examples of gender-specific prescriptive stereotypes include beliefs such as ”Boys shouldn’t cry” and ”Girls should be nice.

Based on the different categories of stereotypes (Fraser et al., 2023) created multiple examples in order to cover different scenarios of different target groups. (Nejadgholi et al., 2024), on the other hand, shifted their focus to gender stereotypes in North America, specifically binary negative descriptive gender stereotypes. They gathered a list of ten well-known stereotypes for each gender and focused explicitly on categories that were linguistically diverse and negative. Then, they chose the best ten stereotypical views per gender, as shown in Table 1.

3.2 Counter Stereotypes Strategies

In the previous section, we talked about the various possible approaches that can be used to counter stereotypes by (Benesch et al., 2016a,b). (Fraser et al., 2023) utilized these strategies in their research. However, some adjustments were necessary. They removed certain strategies such as establishing affiliation, which was deemed inappropriate for AI-generated text, a hostile tone, which was ineffective, and visual communication, which they decided to exclude from their scope. They added five different strategies to the literature based on their research. In total, (Fraser et al., 2023) had eleven counter-stereotypes strategies:

1. **Denouncement of stereotypes:** Indicating that the statement is a stereotype and spreading it around is wrong.

2. **Counter-facts:** Offering an argument supported with facts against the stereotype.

3. **Counter-examples:** Giving examples of individuals or subgroups from the target group who do not fit the stereotype.

4. **Humour:** Using humor to challenge the stereotype.

5. **Warning of consequences:** Explaining the negative consequences of making a stereotypical statement to the speaker or others.

6. **Empathy for the speaker:** Showing empathy and understanding for the speaker’s feelings and thoughts.

7. **Critical questions:** Posing questions to encourage the speaker to reflect on and potentially rethink their beliefs.

Stereotypes about women	Stereotypes about men
Women are bad at science.	Men are aggressive.
Women are irrational.	Men only care about sex.
Women aren't strong leaders.	Men are insensitive.
Women do not drive well.	Men always feel the need to be dominant.
Women are not very assertive.	Men care more about their work than their families.
Women are submissive.	Men hate to show emotion.
Women are hysterical.	Men are egoistic.
Women are emotionally unstable.	Men lack empathy for others.
Women always cry.	Men don't care for others.
Women are dependent.	Men are uncomfortable with intimacy.

Table 1: Binary gender stereotypes used by (Nejadgholi et al., 2024).

8. **Broadening exceptions:** Offering examples of people from outside the target group who also exhibit the stereotypical characteristic.

9. **Broadening universals:** Asserting that anyone can exhibit the stereotypical trait, regardless of their group affiliation.

10. **Positive qualities:** Emphasizing the positive qualities of the target group.

11. **Perspective-taking:** Encouraging the speaker to reflect on how they would feel if they were a member of the target group.

3.3 Generating Counter-Stereotypes

(Fraser et al., 2023) used ChatGPT, a state-of-the-art generative language model, to automatically generate counter-stereotypes for each of the strategies in 3.2. They created a structure for the prompt so that the generated counter-stereotype would follow the desired stereotype and counter-stereotype strategy, be short, and be in social media form. The final prompt was in the following form: ”Counter the stereotype '<stereotype>' by <using strategy>. Limit your response to one sentence. Use tweet style.” They used one of their strategies in place of the phrase <using strategy>. Since ChatGPT tends to write long paragraphs and use multiple strategies to answer a prompt, they decided to avoid that and keep the answer in one sentence. Moreover, they generated feedback in tweet format to keep the user engaged and sound informal. Finally, they generated 198 prompts for each of the stereotypes in Table 2.

(Nejadgholi et al., 2024) followed the lead of (Fraser et al., 2023) and used the same prompt template, and generated a total of 220 counter-stereotypes.

3.4 Evaluation

After manually checking the generated counter-stereotypes, (Nejadgholi et al., 2024) had to eliminate 31 of them because they either employed the incorrect strategy or did not effectively counter it. One of the examples they provided was the stereotype ”Men are egoistic,” which got a response stating that men who show weakness are more liked than those who do not. It was rejected because it did not counter the original statement. Another example statement highlighting societal norms preventing men from showing vulnerability was rejected for not demonstrating empathy as requested. Overall, ChatGPT generally generated counter-stereotypes for all strategies except ’broadening exceptions,’ which was replaced with ’broadening universals’ in order to avoid offending a specific group of people negatively. Moreover, some counter-stereotype strategies like ’broadening exceptions’ were incorrectly generated more than 80% of the time and, therefore, were decided to be removed. Finally, they ended up with 185 counter-stereotypes for ten strategies.

On the other hand, (Fraser et al., 2023) also manually checked the results of the generated counter-stereotypes to ensure quality and efficiency. However, they decided to review some of the statements from the validation set and create some guidelines. These guidelines assessed whether the correct strategy was used and whether the counter-stereotype was offensive or believable. Overall, the annotators mentioned that they had the same understanding of over 80% on whether the correct strategy was used and over 95% on whether it was offensive.

3.5 Annotators

(Fraser et al., 2023) had four annotators who were

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 2: (Fraser et al., 2023) split the stereotype categorizes into four different target groups: Purple: gender/sexuality, Red: race/nationality, Blue: socioeconomic status, Green: religion.

also the paper's authors. All of them were women working in research and coming from different backgrounds. They used a scale from -1 to 1 to evaluate these responses from an outsider's perspective.

- 1 represents the best possible response for the stereotype.
- 0 indicates an OK response that could be improved.
- -1 signifies a response that is not good for the stereotype.

(Nejadgholi et al., 2024) used a website called *Prolific*[1], where they hired 75 annotators from the US who were fluent in English. Each was presented with the generated counter-stereotypes and asked to evaluate the offensiveness, plausibility, and effectiveness. A binary 'yes/no' answer for offensiveness and plausibility. And like the previous study, the -1, 0, and 1 scale for effectiveness.

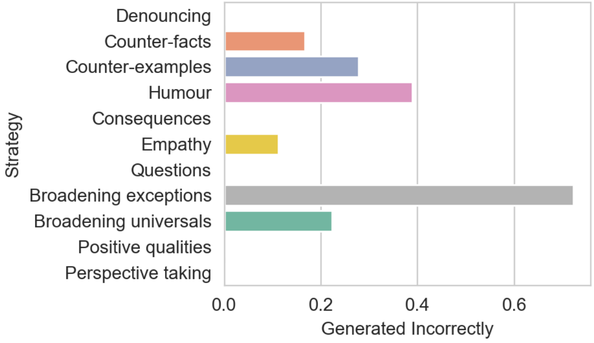

Figure 1: (Fraser et al., 2023)'s incorrectly generated responses that were offensive, not believable, or didn't use the selected strategy.

4 Results

4.1 Challenges with AI-Generated Counter-Stereotypes

Counter-stereotypes generated by ChatGPT revealed some significant challenges when aiming to create effective and appropriate responses. (Fraser et al., 2023) reported that over 70% of the time, ChatGPT generated the strategy 'broadening exceptions' incorrectly, as shown in Figure 1. In one of the examples, ChatGPT generated a counter-stereotype stating that everyone can be greedy and selfish. The strategy of broadening universal was used instead of broadening exceptions for the stereotype *Rich people are greedy*. Moreover, they stated that it was challenging for ChatGPT to create the correct stereotypes as the strategy applies to both greedy and poor individuals.

ChatGPT also faced some challenges in generating responses using the humor strategy. The annotators reported that 33% of the responses could be offensive to the reader. Therefore, they considered it risky to automatically generate a response using this strategy.

On the other hand, (Nejadgholi et al., 2024) found out that almost 36% of participants considered the counter-stereotypes to be potentially offensive. In addition, more than half of the participants considered roughly 19% of the counter-stereotypes to be potentially offensive. Moreover, like the previous study, they also stated that the humor strategy was rated as offensive.

They found that roughly 17% of participants considered the counter-stereotypes unconvincing. Furthermore, over 50% of the participants considered six statements implausible. Their study also iden-

[1] https://www.prolific.co/

tified a correlation between offensiveness and implausibility, where counter-stereotypes perceived as unconvincing are often rated as potentially offensive. Moreover, they stated that such cases were regarded as not practical.

4.2 Counter-Stereotypes Effectiveness

Using the example stereotypes in Table 2, (Fraser et al., 2023) calculated the average results using the -1, 0, and 1 scale discussed in 3.5. In Figure 2, we can see a representation of the results. They observed that the annotators favored the strategies of warning of consequences, empathy, and denouncing stereotypes. The annotators highlighted that empathy was used together with another strategy and did not stand on its own. They also stated that the strategy of denunciation is effective against racist and sexist stereotypes because people do not associate themselves with such traits. Warning of consequences was also an effective strategy that explains the real-life implications of the effect of stereotypes aimed at a particular group. (Fraser et al., 2023) showcased some examples of these strategies:

- **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."*

Overall, the strategies with the worst scores were humor and counter-examples. The jokes generated by ChatGPT were unfunny or meaningless, as reported by the annotators. The counter-examples were considered ineffective because having a few people who do not fit the stereotype is not enough evidence to say that the stereotype does not hold.

The rest of the strategies had slightly positive scores, including counter-facts, asking questions, broadening universals, and promoting perspective-taking. Broadening universals' questions were not clear and sometimes reinforced the original stereotype. Broadening exceptions and positive qualities were rated slightly negative. More specifically, annotators stated that the positive qualities did not counter the stereotype and were unrelated "(e.g., Muslim women are educated, strong, resilient, kindhearted, and have diverse talents and interests)" not related to the topic of hijab.

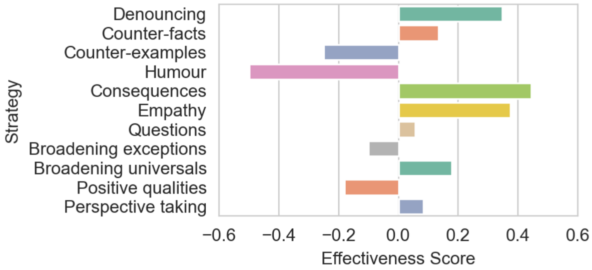

Figure 2: The overall evaluation of counter-stereotype effectiveness by (Fraser et al., 2023) ranges from +1, indicating it is an ideal response to a stereotype, to -1, indicating a poor response.

(Fraser et al., 2023) also checked if certain strategies were more effective based on the case. In Figure 3, they showcase their results of the three dimensions talked about it 3.1. They discovered that broadening exceptions were, in overall cases, less effective in negative stereotypes. They suspected the reason could be that other groups are regarded with negative traits. For example, "Stereotyping Native Americans as alcoholics is unfair and inaccurate, as many other ethnic and cultural groups also struggle with alcoholism." Additionally, empathy was rated higher for positive stereotypes than for negative ones. This is because showing empathy with negative viewpoints was considered unfitting.

When comparing prescriptive versus descriptive strategies, they found that strategies such as denouncing consequences, empathy, critical questions, and broadening universals were more effective in countering prescriptive stereotypes. Moreover, in Figure 2, we can see that critical questions were rated almost neutral. However, they were considered effective in prescriptive stereotypes. Providing counter-examples and counter-facts to prescriptive stereotypes was challenging for annotators. For instance, the counter-stereotype for the stereotype *men don't cry* emphasized that, in fact, men do cry, which is deemed ineffective.

The study then compares statistically accurate

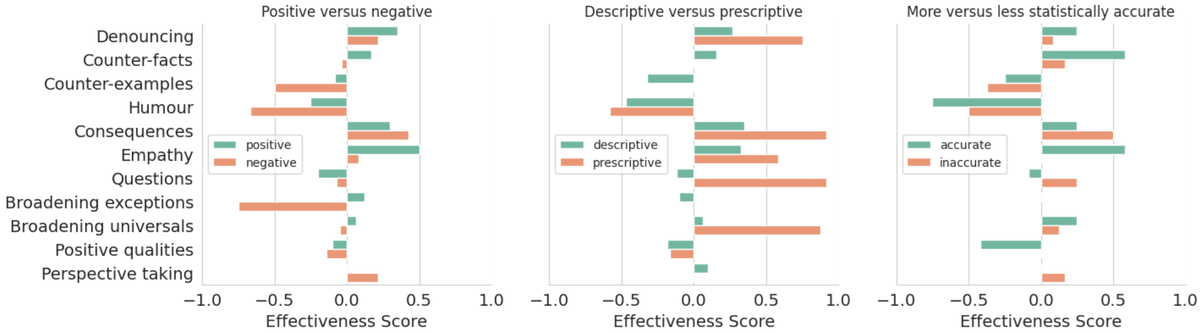

Figure 3: The effectiveness of the strategies for different types of stereotypes by (Fraser et al., 2023).

stereotypes with highly inaccurate ones. Counter-facts were more effective when they had more basis in reality. One response, ”Less than 0.1% of Muslims have been involved in terrorism-related activities,” was rated as extremely poor due to its overstatement of the percentage of Muslims involved in terrorism. Therefore, they stated that it is important to highlight the importance of providing accurate facts and suggest that if ChatGPT cannot generate reliable statistics, using general statements can be a more effective solution.

(Nejadgholi et al., 2024) averaged the scores provided by participants to measure the potential effectiveness of counter-stereotypes. In Figure 4, we can see that participants rated each counter-stereotype strategy. Generally, the strategies ’counter-facts’ and ’broadening universals’ received the highest positive ratings. ’Emphasizing positive qualities’ and ’warning of consequences’ received slightly positive ratings, while ’humor’ was the most ineffective.

Furthermore, (Nejadgholi et al., 2024) studied the stereotype ratings of each group and the gender of each participant. They showcased differences in the average ratings of each subgroup and the differences by stereotype target and gender. They found a higher difference in the ratings for stereotypes about men versus women. For instance, both men and women rated the counter-stereotypes for the strategy ’counter-facts’ significantly higher for stereotypes about women compared to men, even though they were mostly positive in both groups. Moreover, female participants rated counter-facts that challenged stereotypes about men as much more offensive than those that challenged stereotypes about women. ’warning of consequences’ was rated higher for stereotypes about women, while ’emphasizing positive qualities’ was more effective for men.

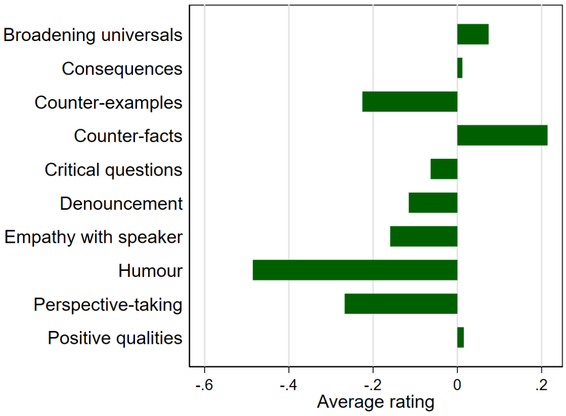

Figure 4: Average ratings indicating how effective the ten counter-strategies could be by (Nejadgholi et al., 2024).

The humor strategy was the least effective, rated worse by women, but neither liked it. ’Denouncement’ and ’critical questions’ were both preferred by female participants and showed the highest difference in ratings between both participants. Finally, female participants rated ’perspective-taking’ and ’counter-examples’ higher for stereotypes about women, while male participants rated them higher for stereotypes about men.

5 Discussion

The studies of (Fraser et al., 2023) and (Nejadgholi et al., 2024) have provided valuable insight into the usage of generated counter-stereotypes, where they discovered the potential benefits of using such methods along with the challenges that come with it. Furthermore, they highlighted the potential risk of using specific strategies that might offend the users. (Fraser et al., 2023) noted that the humor strategy responses were sometimes offensive when the AI claimed membership in the cultural group. (Nejadgholi et al., 2024) also supported this finding

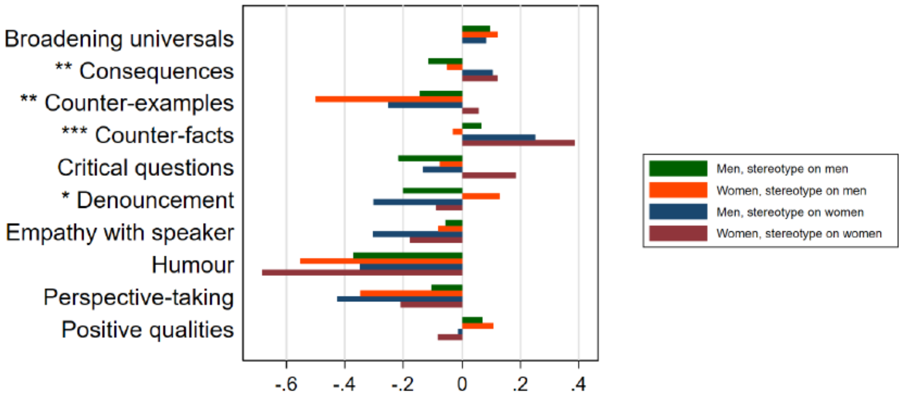

Figure 5: Average ratings of possible effectiveness categorized by the gender of the participants and the specific group associated with the stereotype (men/women) by (Nejadgholi et al., 2024).

and reported that humor was frequently labeled as offensive and implausible, which could possibly reinforce the original stereotype that it was trying to counter. This indicates that AI models could struggle to effectively use humor in sensitive situations due to its complexity and dependency on context.

The studies had some differences in terms of which strategies were the most effective. (Fraser et al., 2023) proposed that combining multiple strategies is the most effective way to generate counter-stereotypes. They mentioned that ChatGPT usually combined different strategies together, such as empathy, counter-facts, and denouncing stereotypes. On the other hand, (Nejadgholi et al., 2024) found that perspective-taking, empathy, and counter-examples were the least effective and broadening universals and counter-facts as the most liked strategies. This difference needs to be looked into more closely, as it could mean that the success of strategies differs based on the specific stereotype, audience, or cultural environment. Future studies should examine these factors in order to create more precise and successful anti-stereotype strategies.

Another critical aspect is how accurate and believable the generated counter-stereotypes are to the users. (Fraser et al., 2023) highlighted this noticeable limitation by presenting their results where ChatGPT generated around 40% of the facts incorrectly. This was also supported by (Nejadgholi et al., 2024), where they stated that the response should be backed by logic and evidence. This highlights an important area for enhancing AI models using fact-checking methods or providing explicit source references.

There were also some differences in the effectiveness when comparing the generated counter-stereotypes between men and women. (Nejadgholi et al., 2024) results highlighted that the responses about women were more effective than responses about men. This represented a challenge and a disadvantage for the generated responses about men. Therefore, it raises crucial questions about the role of AI in challenging existing social narratives and highlights the need for diverse and balanced training data.

6 Conclusion

In conclusion, the studies by (Fraser et al., 2023) & (Nejadgholi et al., 2024) provide valuable insights into using AI-generated counter-stereotypes. They highlight potential benefits and challenges, including the risks associated with humor in sensitive contexts, the varying effectiveness of different strategies, and the importance of accuracy and plausibility in generated responses. The research also reveals discrepancies in the effectiveness of counter-stereotypes for men and women, emphasizing the need for balanced training data. Moving forward, it is crucial to further investigate the factors influencing strategy success across different stereotypes, audiences, and cultural contexts. Additionally, improving fact-checking mechanisms and source referencing in AI models will be essential to enhance the credibility and impact of generated counter-stereotypes in combating harmful social narratives.

References

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

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

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):800–814.

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

Diane Felmlee, Patricia Inara Rodis, and Amy Zhang. 2020. Sexist slurs: Reinforcing feminine stereotypes online. *Sex Roles*, 83(1):16–28.

Susan T. Fiske, Amy J.C. Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social Cognition*, pages 162–214. Routledge.

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.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38, Toronto, Canada. Association for Computational Linguistics.

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. In T. D. Nelson, editor, *Handbook of Prejudice, Stereotyping, and Discrimination*, pages 227–199.

Anna Kerkhof and Valentin Reich. 2023. Gender stereotypes in user-generated content. CESifo Working Paper 10578, CESifo. Available at SSRN: `https://ssrn.com/abstract=4527554` or `http://dx.doi.org/10.2139/ssrn.4527554`.

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.

Isar Nejadgholi, Kathleen C. Fraser, Anna Kerkhof, and Svetlana Kiritchenko. 2024. Challenging negative gender stereotypes: A study on the effectiveness of automated counter-stereotypes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3005–3015, Torino, Italia. ELRA and ICCL.

Patricia Palffy, Patrick Lehnert, and Uschi Backes-Gellner. 2023. Countering gender-typicality in occupational choices: An information intervention targeted at adolescents. Technical report, University of Zurich, Department of Business Administration (IBW).

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. Association for Computational Linguistics.

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.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association* *for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

A. R. Todd, G. V. Bodenhausen, J. A. Richeson, and A. D. Galinsky. 2011. Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology*, 100(6):1027–1042.

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.