Challenging Negative Gender Stereotypes: A Study on the Effectiveness of Automated Counter-Stereotypes

Author1, Author2, Author3
Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com
{author1, author5, author9}@abc.org

Abstract

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

Gender stereotypes are pervasive notions about individuals based on gender and play a significant role in shaping societal attitudes, behaviours, and even opportunities. Recognizing the negative implications of gender stereotypes, particularly in online communications, this study investigates various strategies to automatically counteract and challenge these views. We assess the feasibility and efficacy of eleven counter-strategies by presenting AI-generated gender-based counter-stereotypes to (self-identified) male and female study participants. The strategies of counter-facts and broadening universals emerged as the most robust approaches, while humour, perspective-taking, counter-examples, and empathy were perceived as less impactful. Also, the differences in ratings were more pronounced for stereotypes about men vs. stereotypes about women than between the genders of the study participants. Also, alarmingly, certain counter-strategies led to increased perceived offensiveness and implausibility, especially when rated by a participant of the opposite gender. Our analysis and the collected dataset offer foundational insight into counter-stereotype generation, guiding future efforts to develop strategies that effectively challenge gender stereotypes in online conversations.

Keywords: Gender Stereotypes, Counter-Stereotypes, Social Influence, Online Conversations

1. Introduction

Stereotypes involve attributing certain characteristics to a person purely on the basis of their perceived membership in a certain social category, often defined by demographic features such as race, ethnicity, age, or religious affiliation. In particular, perceived *gender*[1] continues to be one of the most salient features by which these conscious and subconscious social categorizations are made, despite growing recognition that gender is not necessarily apparent from a person's appearance, is not a binary categorization, and in most cases is not relevant to the situation (Ellemers, 2018). Children as young as five years-old show a strong tendency to sort people into male and female categories (Aina and Cameron, 2011), and as young as six years-old make assumptions about a person's intelligence (Bian et al., 2017) based on this categorization. Gender stereotypes can be harmful to people of all genders, as they define particular expectations for how people can and should behave, regardless of individual strengths and weaknesses. Thus, from a young age, girls and women are expected to be friendly, nurturing, deferential, and concerned with presenting a feminine appearance, while boys and men are expected to be strong, competitive, and unemotional.

Studies show that gender stereotypes lead to biased perceptions of women's intellectual and leadership performance, limiting their career opportunities; for example, when the same CV is submitted with a typically male versus female name, the male candidates are judged as more competent (Moss-Racusin et al., 2012). Similar examples of gender discrimination against women in the professional sphere are unfortunately quite common (Correll et al., 2007; Buffington et al., 2016; Bobbitt-Zeher, 2011). Meanwhile, the pressure experienced by men to conform to masculine stereotypes can lead to impaired mental health and substance abuse (Wong et al., 2017). Stereotypical beliefs about masculine gender roles also lead to a lack of help-seeking by male victims of intimate partner violence and sexual assault (Bates et al., 2019).

Stereotypes are reinforced by repeated exposure. On the other hand, stereotypical associations can be weakened by exposure to *counter-stereotypes* or information that disrupts or challenges the stereotype. Several counter-strategies can be employed, such as providing factual information to contradict the stereotype, asking questions to motivate critical thinking, or encourag-

[1] This study is limited to binary gender stereotypes, yet we recognize the prevalence of harmful non-binary stereotypes and believe their complexity warrants a dedicated study.

ing the speaker to "put themselves in the target group's shoes". While many of the social psychology studies on counter-stereotypes involve in-person interventions, we are interested in the question of how to effectively generate counter-stereotypes in online spaces, such as on social media platforms, where such content is prevalent (Felmlee et al., 2020; Kerkhof and Reich, 2023). To that end, we present the results of an online annotation experiment to assess whether generative AI technology (in our case, ChatGPT) can be used to generate appropriate and plausible counter-stereotypes and which counter-strategy is judged to be most effective at countering negative gender stereotypes. Our results indicate that 1) ChatGPT can be used to generate effective counter-stereotypes, 2) there are differences in terms of annotator ratings depending both on whether the stereotype targets men or women and whether the annotators themselves identify as male or female, and 3) the most promising strategies for future work are presenting counter-facts, or stating that all people can have the stereotypical trait regardless of their gender. Our annotated dataset will be publicly available at [Anonymized Github Repository].

In summary, our main contributions are as follows:

- A dataset presenting the ratings of offensiveness, plausibility, and effectiveness from 75 annotators for 220 counter-stereotypes. The counter-stereotypes are generated automatically according to 11 counter-strategies for 10 negative common stereotypes about men and 10 negative common stereotypes about women.
- An annotation study indicating the most (and least) effective counter-strategies to challenge gender-based stereotyping online.
- Fine-grained analysis showing that while there are differences in the ratings depending on whether the annotators identify as male or female, the more salient differences correspond to whether the stereotype targets men or women.

2. Related Work

This work draws on research from the social sciences as well as NLP and computer science. The processes by which stereotypes are formed, spread, and potentially disrupted have been studied extensively in social psychology, and we will provide only a brief overview of some key findings here. Researchers in NLP have recently begun translating these findings into new methods for countering hate speech, microaggressions, and stereotypes in online discourse, and this work will be discussed as well.

2.1. The Psychology of Stereotypes

Stereotypes arise from cognitive processes that developed to help humans immediately categorize unknown strangers and determine their potential threat level ("friend" or "foe") (Fiske et al., 2018). However, these cognitive shortcuts have limited utility in most modern social contexts and should be refined or discarded when additional information is available to make more accurate judgements about an individual. Nonetheless, gender stereotypes, whether consciously or subconsciously held, persist across different cultures and contexts.

Numerous studies evaluate different methods for reducing the effect of stereotypes. Studies of racial bias have reported the effectiveness of strategies such as exposure to anti-stereotype exemplars (examples of people who disconfirm the stereotype in question) (Dasgupta and Greenwald, 2001), exercises that involve thinking from the target group's perspective (Todd et al., 2011), and setting explicit goals for cooperation and equality (Blincoe and Harris, 2009; Wyer, 2010).

Palffy et al. (2023) conducted a field experiment to examine the effectiveness of counter-stereotypical framing and role models for adolescents choosing future occupations. They found that the intervention successfully increased the number of women who applied for typically male jobs in STEM fields. However, it did not increase the number of men who applied to typically female jobs in health and care-taking occupations.

Foster-Hanson et al. (2022) examined the question of how to reduce gender stereotypes in children, specifically focusing on the *essentialist* nature of such beliefs: the assumption that all members of a group are fundamentally the same due to some underlying essential nature. They observe that the statement "Girls can be good at math too" is a common response to the stereotypical statement of "Boys are good at math." However, it only challenges the content of the specific claim about math skills while not addressing (and possibly even reinforcing) the essentialist belief that gender is a meaningful way to categorize people. They suggest instead the strategies of narrowing the scope of the statement ("Well, *John* is good at math") or broadening the scope of the statement ("Well, *lots of kids* are good at math"), and show that these strategies are more effective at reducing prescriptivist beliefs about gender in 4-year-old children.

2.2. Countering Stereotypes with NLP

The NLP community began exploring the automatic generation of *counter-speech* (statements challenging hate speech) with the work of Qian et al. (2019), Mathew et al. (2019), and Chung et al. (2019) and followed by studies by Tekiroğlu et al. (2020) and Zhu and Bhat (2021), among others. While still an active field of research, a new branch also recognizes the need to counter less extreme forms of abuse, such as stereotyping and microaggressions. Critically, while toxic content classified as "hate speech" can usually be removed from an online platform according to the terms of service, stereotyping and microaggressions are more likely to remain visible on the platform and thus need to be handled differently in order to mitigate their potential harms. Additionally, in many cases, the writers of such content have no intent to offend anyone, and therefore, there is also a component of education and empathy that can be useful in such cases.

Ashida and Komachi (2022) automatically generated counter-speech as well as 'micro-interventions', a term which specifically refers to a statement countering a microaggression. They compared few-shot versus zero-shot approaches using GPT-2, GPT-3, and GPT-neo, and found that GPT-3 produced the least offensive and most informative responses, although they caution that fact-checking is necessary to avoid hallucinations or misinformation.

Allaway et al. (2022) examined five strategies for countering essentialist claims in generic statements about groups. In line with (Foster-Hanson et al., 2022) above, they found that broadening statements, which remind the reader that these characteristics are not unique to one particular group, were rated quite highly. In contrast, annotators did not generally prefer providing direct counter-evidence, partly due to a high number of incorrect or subjective examples in the automatically generated text.

Fraser et al. (2023) surveyed the literature and identified 11 strategies for countering stereotypes that could potentially be implemented using generative language models. They used ChatGPT to automatically generate counter-stereotypes for 18 stereotypes common in North America, spanning the dimensions of negative–positive, descriptive–prescriptive, and "more" accurate–"less" accurate. In a small ($n = 4$, all female) annotation study, they determined that the strategies of denouncing stereotypes, warning of consequences, and using an empathetic tone were the most promising strategies overall. We make use of Fraser et al.'s list of counter-strategies, but we focus specifically on negative, descriptive, gender-based stereotypes and conduct a much larger and more representative annotation study ($n = 75$, 50% male–female split).

3. Human Study Design

We describe next the study design: stereotype selection, counter-stereotype generation using automatic means (ChatGPT), and a questionnaire to gather participants' perceptions of counter-stereotype effectiveness.[2]

3.1. Gender Stereotypes

In this study, we focus on negative, descriptive, binary gender stereotypes (i.e., stereotypes portraying either men or women in a negative way), common in North America. We compiled a list of ten well-known stereotypes for each gender using the existing literature and online sources. Table 1 shows the stereotypes selected for the study.

3.2. Generating Counter-Statements

Following Fraser et al. (2023), we evaluate eleven counter-stereotype strategies:

1. **Broadening exceptions:** Stating that the stereotypical trait is not unique to the target group by providing examples of other socio-demographic groups that share the trait.
2. **Broadening universals:** Stating that the stereotypical trait is not unique to the target group and that all people, regardless of group membership, can have the trait.
3. **Warning of consequences:** Pointing out possible negative outcomes of perpetuating the stereotype for the speaker, the target group, or the society.
4. **Counter-examples**: Providing examples of individuals or subgroups from the target group who do not have the stereotypical trait.
5. **Counter-facts:** Providing facts that contradict the stereotype.
6. **Critical questions:** Asking questions to motivate the speaker to review and possibly reconsider their beliefs.
7. **Denouncement:** Pointing out that the statement is a stereotype and perpetuating stereotypes is wrong.
8. **Empathy for the speaker:** Expressing empathy with the speaker's feelings and thoughts.
9. **Humour:** Challenge the stereotype using humour.

[2]This study was approved by the Ethics Boards at the authors' institutions.

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 in this study.

10. **Perspective-taking:** Asking the speaker to consider the stereotype from the target group's perspective.

11. **Emphasizing positive qualities:** Highlighting the positive characteristics of the target group.

For each strategy and each stereotype, we prompted ChatGPT[3] to generate one sentence in a social-media style. We used the prompts provided by Fraser et al. (2023). In total, 220 counter-statements were generated.

Next, we manually checked each counter-statement and excluded 31 statements that were not countering a given stereotype or that were generated using a strategy other than requested. For example, given the stereotype "*Men are egoistic"*, the statement "*Research shows that women and men have equal levels of self-esteem, and that men who expressed vulnerability were actually more well-liked than those who did not. #byeegoisticstereotype*" was rejected since it did not counter the corresponding stereotype. Likewise, "*Men do experience emotions, but societal expectations often discourage them from showing vulnerability and expressing themselves. #empathyforall*" was rejected as not showing empathy for the speaker when it was the requested strategy.

Table 2 shows the breakdown of the rejected counter-statements by strategy. Overall, ChatGPT was generally able to successfully generate counter-statements for all strategies, except 'broadening exceptions'. For most of the stereotypes, in place of 'broadening exceptions', ChatGPT used a related strategy of 'broadening universals'. We speculate that ChatGPT chose this strategy to avoid making negative statements about particular social groups. Since 80% of the counter-statements for 'broadening exceptions' were generated using incorrect strategy, we decided to exclude this strategy from further study, leading to the exclusion of 35 counter-stereotypes in total. The remaining 185 counter-statements for 10 strategies were presented to the participants in the survey.

Strategy	# rejected
Broadening exceptions	16
Broadening universals	2
Consequences	0
Counter-examples	2
Counter-facts	3
Critical questions	0
Denouncement	1
Empathy with speaker	4
Humour	1
Perspective-taking	2
Positive qualities	0
Total	**31**

Table 2: The number of counter-statements rejected after manual assessment.

3.3. Questionnaire

After providing informed consent, the participants were first presented with general instructions about the task. They were told that they would be shown gender stereotypes accompanied by the corresponding counter-stereotypes. Counter-stereotype was defined as follows: *"A counter-stereotype challenges a gender stereotype. E.g., a counter-stereotype could present factual arguments against the gender stereotype, provide counter-examples or ask the speaker how they would feel if they were part of the target group. A counter-stereotype is* ***not*** *just the opposite of a gender stereotype."* Also, one example pair of stereotype–counter-stereotype was shown.

Then, the annotation task was presented and explained as shown in Figure 1. The task included

[3] https://platform.openai.com/docs/models/gpt-3-5

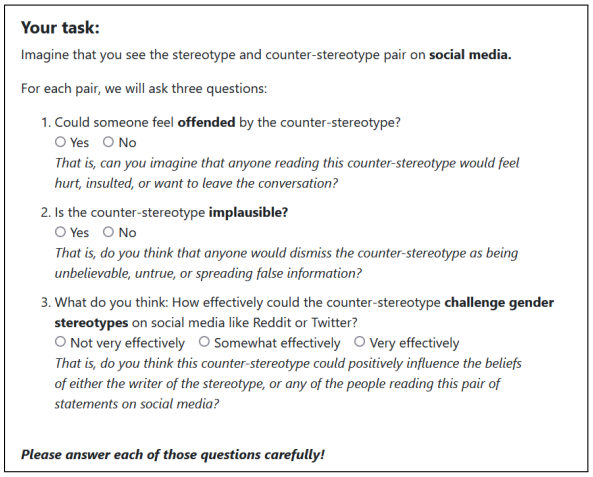

Figure 1: The description of the task presented to the participants.

questions to evaluate the **offensiveness**, **implausibility**, and potential **effectiveness** of each pair of stereotype–counter-stereotype. The questions related to offensiveness and implausibility required a binary answer, 'yes/no'. The third question had three options: (1) very effectively (assigned a score of 1), (2) somewhat effectively (score of 0), (3) not very effectively (score of -1).

To ensure the quality of responses, we employed three strategies. First, attention check questions were included in the survey. Second, we announced that a small number of randomly drawn participants had the chance of earning a bonus payment of 5 USD (on top of their participation reimbursement) if their responses to the questions were within 20% bandwidth of the average response. Finally, we monitored the amount of time spent by a participant on the task.

3.4. Participants

We recruited a total of 75 participants using *Prolific*, an online platform that connects researchers and survey takers. The platform offers a large pool of pre-screened participants from OECD countries who wish to participate in various studies[4].

Since our study focuses on stereotypes prevalent in North America, we recruited participants solely from the U.S. and requested fluency in English. Out of 75 participants, 37 were male and 38 were female. The mean (median) age was 40.41 (38) years. Each participant was asked to assess 30 stereotype–counter-statement pairs, which, on average, took around 15 minutes to complete. The participants were paid 3.00 USD (around $12 per hour), which corresponds to the average reimbursement participants receive on Prolific. All participants passed the attention check, and each stereotype–counter-statement pair was rated on average by 12 participants (6 male and 6 female).

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

4. Results

In the following sections, we present and discuss the results of human annotations for ten counter-strategies. (Recall that the strategy 'broadening exceptions' was excluded due to ChatGPT's inability to consistently generate statements using this strategy (Sec. 3.2)).

4.1. Offensiveness and Implausibility of Counter-Stereotypes

Many counter-statements were rated as potentially offensive by at least some participants. On average, a counter-statement was perceived as potentially offensive by 35.8% of participants who rated it, and 35 (out of 185) counter-statements were perceived as potentially offensive by more than half of the raters. Counter-statements generated for the humour strategy were rated as offensive most frequently.

These numbers are surprisingly high, especially taking into consideration the amount of effort put into guard-railing ChatGPT from generating offensive outputs. To clarify, none of the generated statements were explicitly offensive,[5] yet participants indicated that some statements could potentially offend or upset certain social groups or users (e.g., *'Just watched my husband try to fold a fitted sheet - dominance is not in his DNA. #NoDominateGene #HumorWins*, which was generated for the strategy 'humour').

In fact, we note that there is a delicate boundary between what is considered offensive and non-offensive in this context. Even minor nuances can significantly influence how participants perceive a counter-statement. For example, the counter-statement *"Women are strong, resilient, intelligent, nurturing, ambitious, and capable leaders who can handle their emotions without shedding a tear. #WomenEmpowerment"* was perceived as offensive by the majority of the participants, but the sentence *"Women are intelligent, intuitive, and capable decision-makers who excel in both emotional intelligence and logical reasoning. #WomenAreNotIrrational"* was rated as non-offensive by all participants. Though seemingly positive, the former example promotes a narrow understanding of strength and leadership, implying that showing emotion or shedding a tear is a sign of weakness or incapability. In contrast, the second sentence is more inclusive and acknowledges a broad range of capabilities in women without falling into

[5] We consider a text explicitly offensive if it includes direct and unambiguous words or expressions, such as overtly derogatory language, intended to insult, degrade, or belittle someone.

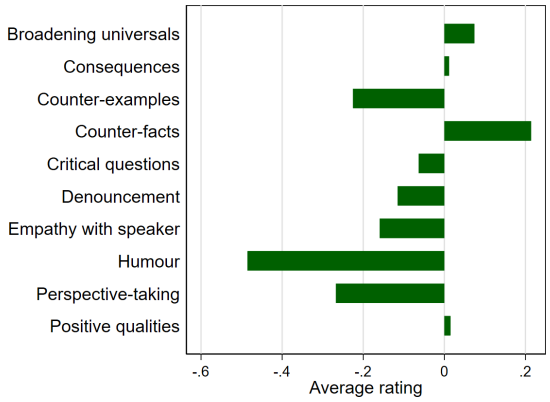

Figure 2: Average ratings of potential effectiveness for the ten counter-strategies.

the prescribing of certain emotional responses. This observation highlights the complexity and the importance of validating the non-offensiveness of counter-statements to ensure that counter-speech results in positive exchanges and does not escalate conflicts further.

A counter-stereotype was perceived as implausible on average by 16.7% of participants who rated it. Also, 6 statements were rated as implausible by more than 50% of participants. We found a correlation between the offensiveness and implausibility ratings: counter-stereotypes that participants perceived as implausible were also often rated as potentially offensive (Pearson ρ of 0.32). Counter-statements perceived as offensive or implausible were also often rated as ineffective (Pearson ρ of -0.21 and -0.29, respectively).

4.2. Effectiveness of Counter-Stereotypes

We measure the effectiveness of counter-statements by averaging the scores (1, 0, or -1) obtained by converting participants' answers to numerical values as described in Section 3.3.

Most and least effective strategies: Figure 2 shows the average ratings provided by the participants for each counter-stereotype strategy. Overall, two strategies, 'counter-facts' and 'broadening universals', received the highest positive ratings. The strategies of 'emphasizing positive qualities' and 'warning of consequences' received slightly positive average ratings, while all the other strategies were ranked negatively, on average. 'Humour' stands out as the most ineffective counter-stereotype strategy when automatically implemented by prompting ChatGPT.

Fine-grained results across subgroups: To get further insights, we split the ratings by the stereotyped group (stereotypes about women vs. stereotypes about men) and by participant-reported gender (male vs. female). Figure 3 and Table 3 present these detailed results.

Overall, we observe greater differences in the ratings for stereotypes about men versus stereotypes about women (mean absolute difference of 0.15) than between the ratings of male and female annotators (mean absolute difference of 0.10). Table 3 presents the fine-grained differences and their statistical significance. For example, although 'counter-facts' received mostly positive ratings for both subgroups of stereotypes, statements countering stereotypes about women were rated substantially higher than statements countering stereotypes about men, regardless of the gender of the annotator. Interestingly, female participants perceived counter-facts opposing stereotypes about men (e.g., *"According to a study by the American Psychological Association, men reported higher levels of intimacy overall, including emotional and physical intimacy, than women did. #menareintimate #breakthestereotype"*) as offensive and/or implausible at substantially higher rates than counter-facts opposing stereotypes about women. This was also at higher rates than male participants perceived counter-facts opposing stereotypes about both groups. Also, 'emphasizing positive qualities' was mostly perceived as effective for countering stereotypes about men, while 'warning of consequences' received higher ratings for stereotypes about women.

The strategies showing the greatest difference in ratings between male and female participants are 'critical questions' and 'denouncement', both of which were preferred more by female participants, and 'humour', which was disliked by all participants, but more so by females. 'Perspective-taking' and 'counter-examples' were rated higher by female participants for stereotypes about women and by male participants for stereotypes about men.

5. Discussion

Our research indicates a clear distinction in the potential effectiveness of various strategies when generated automatically. Certain strategies have a more universal appeal and efficacy, while others have limited or even negative impacts.

'Counter-facts' and 'broadening universals' were the two strategies that exhibited the most positive outcomes across different scenarios and audiences. 'Broadening universals' involves presenting a wider, more inclusive understanding of a particular trait or behaviour, challenging the limited scope of stereotypes. Similarly, counter-facts provide a logical and evidence-based method of disputing any unfounded claims. These approaches

Table 3: The differences and p-values of ratings for stereotype target (ratings of counter stereotypes about men - ratings of counter stereotypes about women) and for participants' gender (ratings by men - ratings by women), per strategy. Asterisks indicate whether the difference between ratings of stereotypes about men and women is statistically significant (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Strategy	Stereotype Target difference (p-value)	Participant Gender difference (p-value)	Observations
Denouncing	- 0.14 (0.09*)	0.26 (0.001***)	330
Counter-facts	0.30 (0.01***)	0.05 (0.67)	187
Counter-examples	0.26 (0.05**)	-0.04 (0.76)	142
Humour	-0.03 (0.72)	- 0.24 (0.01**)	206
Consequences	0.19 (0.03**)	0.05 (0.56)	259
Empathy	- 0.17 (0.09*)	0.05 (0.60)	207
Questions	0.16 (0.11)	0.22 (0.03**)	205
Broadening universals	- 0.07 (0.46)	- 0.03 (0.74)	216
Positive qualities	- 0.14 (0.13)	- 0.03 (0.75)	265
Perspective taking	-0.07 (0.49)	-0.03 (0.80)	187

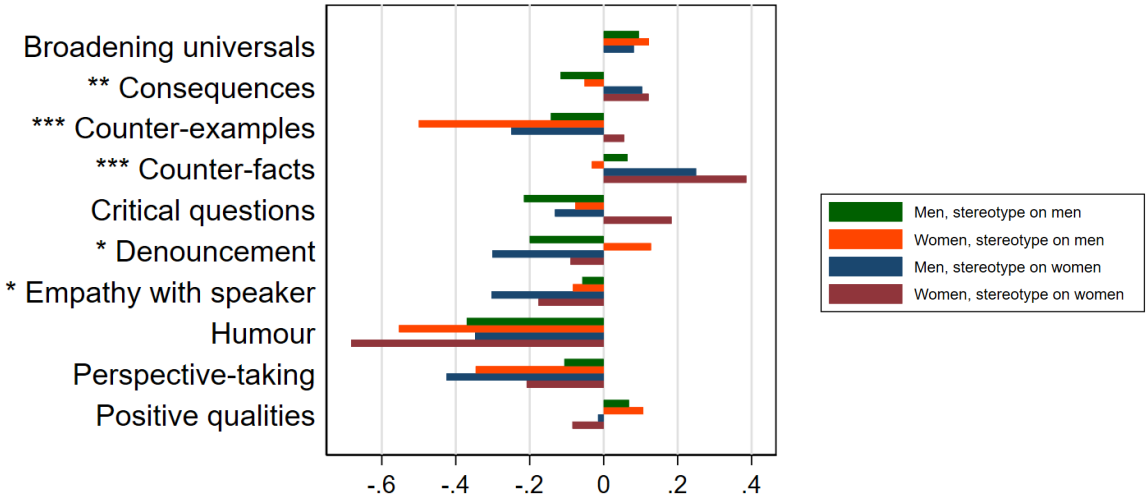

Figure 3: Average ratings of potential effectiveness broken down by the participants' gender and the group the stereotype is about (men/women). For example, "Men, stereotypes on men" refers to male participants' ratings for counter-stereotypes about men. Asterisks indicate whether the difference between ratings of stereotypes about men and women is statistically significant (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

immediately offer an alternative viewpoint, urging individuals to reconsider their biases. However, we observed that female participants sometimes found the counter-facts ineffective when used to counter stereotypes about men. This suggests that a one-size-fits-all approach may be insufficient and audience-specific nuances, where applicable, may enhance the effectiveness of a strategy.

On the other hand, strategies such as 'humour', 'perspective-taking', 'counter-examples', and 'empathy with the speaker' have been rated as ineffective. Among these, the use of 'humour' in counter-statements was particularly problematic. While humour can be a powerful tool for challenging societal norms, it also risks being offensive or misinterpreted, especially in sensitive areas like gender stereotypes. Also, producing high-quality humour is a difficult technical task for NLP models; even state-of-the-art generative models have yet to master this task. In our findings, automatically generated humorous counter-statements were interpreted as offensive and implausible by a significant portion of the audience. This highlights the precarious nature of using humour in automatic interventions; it can inadvertently reinforce the very stereotypes it aims to challenge.

Overall, we observe that, on average 35.8% of participants rated AI-generated counter-statements as potentially offensive, emphasizing the complexities of automating responses in delicate subject areas like gender stereotypes.

Besides humorous statements, counter-strategies that challenge a stereotype about one binary gender by directly attributing the negative trait to the opposite gender (e.g., by providing statistics on the other gender) may be seen as unnecessarily confrontational or belittling.

Furthermore, strategies like 'perspective-taking', which involves encouraging individuals to view a situation from another person's standpoint, and 'empathy', which focuses on fostering a deeper emotional understanding, fail to resonate with the audience in case of online gender stereotyping. While these strategies are valuable in interpersonal interactions, they might be interpreted as irrelevant, insincere or patronizing when automated and used in online communications.

Another finding in our data is the noticeable difference in ratings when comparing stereotypes about women versus stereotypes about men. We observe that the stereotype's inherent nature and societal connotations play a larger role than the gender of the individual evaluating it. Due to a long history of the oppression of women and contemporary movements for reclaiming women's rights, societal discourse, both offline and online, includes extensive discussions on the issue. This has allowed models like ChatGPT to be adept at generating 'counter-facts' and 'warnings of consequences' that directly address these stereotypes. In other words, the established nature of stereotypes about women, paired with the rich database of discussions and rebuttals around them, appears to enable AI models to craft plausible and effective responses to counter such stereotypes.

In contrast, stereotypes that target men present a challenge for automatic countering. The counter-facts generated in response to stereotypes about men were often deemed less effective by participants. For example, in response to the stereotype *"Men don't care for others"*, ChatGPT generated the counter-fact, *"According to a study by the DoSomething.org, men are actually more likely than women to donate to charity and volunteer their time for causes. #MenCareToo"*, which might seem implausible or offensive to women. However, in response to the stereotype *"Women are irrational"*, it generates the counter-fact *"Studies show that men and women make decisions with similar levels of emotional involvement and rational thinking, debunking the stereotype that women are irrational. #GenderBias #FactsOverStereotypes"*, which has a neutral tone about both genders. Also, our qualitative assessment shows that counter-stereotypes generated for men are less specific than those generated for women. For example, for the strategy of mentioning the consequences of stereotypes, to counter the stereotype *"Men are aggressive."* ChatGPT generates *"Spreading the stereotype that men are aggressive can lead to harmful generalizations and discrimination, let's break the cycle. #StopStereotyping."*, but for stereotype *"Women are irrational."* it generates *"Spreading the stereotype that women are irrational can lead to women being underestimated and undervalued, ultimately hindering progress and equality."*. While the first statement broadly highlights the consequences of *all* stereotypes, it does not specify who exactly would be harmed, or how. On the other hand, the second statement explicitly mentions that women will be the ones "underestimated" and "undervalued" due to the stereotype and specifies its larger societal impact, suggesting that such stereotypes can "hinder overall progress and equality". This might stem from the fact that discussions on stereotypes about men, while present, are not as widely prevalent or as deeply ingrained as their counterparts, stereotypes about women. Thus, traditional strategies, like counter-facts, were not as impactful. However, our research found that other approaches, like listing the positive qualities inherent in men and emphasizing the fact that attributes like aggression and insensitivity can be found across all genders, were more effective.

6. Data Usability

We release our data publicly to be used for further research. This dataset is in English and was developed to understand the perceptions of individuals regarding counter-statements to negative, descriptive gender stereotypes about men and women. The primary goal was to evaluate the offensiveness, plausibility, and potential effectiveness of counter-statements when presented against stereotypes in online platforms.

This dataset includes 1) ten well-known North American stereotypes for each binary gender, sourced from existing literature and online platforms, 2) several one-sentence-long counter-statements for each stereotype generated in a social-media style by ChatGPT according to established counter-strategies to challenge stereotypes, and manually validated by the authors, 3) human ratings on the offensiveness, implausibility, and potential effectiveness of the counter-statements.

This data includes binary answers (yes/no) for offensiveness and plausibility and graded responses (score of -1, 0, 1) to measure the perceived effectiveness of counter-statements. It should be noted that these ratings capture perceptions towards stereotypes and their counter-statements and do not assert the truth or validity of these statements. The ratings provided by the participants are subjective and reflect the participants' opinions, which might be affected by their personal

experiences and beliefs. Further research is required to assess the actual effectiveness of these strategies in changing online users' stereotypical beliefs.

We anticipate several potential uses for this dataset. First, researchers can use this data to understand common gender stereotypes and the societal perception of counter-statements. Second, this data might be used for NLP tasks such as training and/or evaluating models to identify, respond to, or counteract stereotypes in digital content. Third, the insights learned from this data might be used by NGOs or community groups to craft more effective stereotype-countering campaigns.

Our data comes with limitations that need to be understood and mitigated before considering the above use cases. This dataset focuses exclusively on negative, descriptive, binary gender stereotypes. Therefore, it does not encompass all facets of stereotypes and is not inclusive of all gender identities. Also, the dataset is built on subjective opinions and might not be universally applicable or representative. Also, by nature, stereotypes can be sensitive and potentially offensive; care should be taken in their use and interpretation to avoid perpetuating harmful beliefs or norms.

7. Conclusion

We conducted a large-scale human study on the potential effectiveness of automatically generated statements to counter common gender stereotypes. We found that while some strategies offer a promising avenue to counter stereotypes, others require careful consideration to ensure they don't have an adverse effect. Future automatic interventions should prioritize strategies that have demonstrated consistent effectiveness while being wary of those that can be offensive or prone to include misinformation.

Confronting gender stereotypes requires a nuanced and tailored approach, considering stereotypes' historical, cultural, and societal context. The effectiveness of a counter-strategy may vary according to the social group being stereotyped, the expected audience, and other contextual features.

8. Ethics Statement

While our study represents a promising step in counteracting gender stereotypes online, it comes with limitations and ethical considerations that require ongoing attention. Emphasizing transparency, continuous evaluation, and social context will be essential as we navigate this intersection of technology and societal constructs.

We release a dataset that can be used for further research or social applications on countering stereotypes. Before using this dataset, users should familiarize themselves with the context and limitations of the dataset. The dataset should be handled and interpreted responsibly, considering the potential ethical implications (Kirk et al., 2022). We also strongly recommend cross-referencing with other relevant literature or datasets for a more holistic view.

One limitation of this research is its focus on binary gender stereotypes, including only male and female identities. We are fully cognizant of the widespread prevalence and significance of non-binary stereotypes in contemporary online discourse and lived experiences. Our decision to not delve into non-binary stereotypes within this study stems not from oversight but from an understanding of their intricate complexity. We believe these complexities, nuances, and variances inherent to non-binary stereotypes warrant an exhaustive and dedicated study separate from the constraints of our present research.

If successful, automatic methods to counteract gender stereotypes can reshape online conversations, fostering a more inclusive and equal digital environment. Such interventions can be instrumental in actively challenging entrenched biases, potentially influencing societal perceptions and behaviour. However, while AI models can mimic human language patterns, they may not always capture the subtleties and sensitivities needed when addressing deeply embedded societal constructs, so human oversight might be necessary, especially in critical applications. Furthermore, our sample's demographics might not be globally representative, limiting the generalizability of the findings.

As indicated, certain counter-strategies, especially when applied across genders, heightened perceptions of offensiveness and implausibility. These interventions could inadvertently perpetuate biases or spark unintentional controversies without careful calibration. It is essential to take caution, ensuring strategies do not polarize views further or cause distress to the audience. Also, even while attempting to challenge stereotypes, current AI models can inadvertently reproduce or reinforce societal biases present in the training data. Acknowledging this limitation and working continually to minimize such repercussions is crucial.

AI-generated strategies need consistent evaluation and refinement. What works today might not be as effective tomorrow due to evolving societal contexts. Regular reassessment ensures interventions remain relevant and impactful. It is imperative that users are informed when AI-generated strategies are being employed to counteract stereotypes, ensuring a transparent online interaction.

9. Bibliographical References

Olaiya E Aina and Petronella A Cameron. 2011. Why does gender matter? counteracting stereotypes with young children. *Dimensions of Early Childhood*, 39(3):11–19.

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.

Elizabeth A Bates, Kathryn R Klement, Linda K Kaye, and Charlotte R Pennington. 2019. The impact of gendered stereotypes on perceptions of violence: A commentary. *Sex Roles*, 81:34–43.

Lin Bian, Sarah-Jane Leslie, and Andrei Cimpian. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323):389–391.

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.

Donna Bobbitt-Zeher. 2011. Gender discrimination at work: Connecting gender stereotypes, institutional policies, and gender composition of workplace. *Gender & Society*, 25(6):764–786.

Catherine Buffington, Benjamin Cerf, Christina Jones, and Bruce A Weinberg. 2016. STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census. *American Economic Review*, 106(5):333–338.

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.

Shelley J Correll, Stephen Benard, and In Paik. 2007. Getting a job: Is there a motherhood penalty? *American Journal of Sociology*, 112(5):1297–1338.

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.

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

Susan T Fiske, Amy JC 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.

Emily Foster-Hanson, Sarah-Jane Leslie, and Marjorie Rhodes. 2022. Speaking of kinds: How correcting generic statements can shape children's concepts. *Cognitive Science*, 46(12):e13223.

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.

Anna Kerkhof and Valentin Reich. 2023. Gender stereotypes in user-generated content.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510.

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.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

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).

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.

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.

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.

Y Joel Wong, Moon-Ho Ringo Ho, Shu-Yi Wang, and IS Miller. 2017. Meta-analyses of the relationship between conformity to masculine norms and mental health-related outcomes. *Journal of counseling psychology*, 64(1):80.

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.