Overview. The prevalence of gender bias has made it an important social issue. Recent examples include the numerous complaints against Uber and the infamous memo questioning Google's diversity policies. Furthermore, advances in AI exacerbate this problem, as AI systems can absorb biases in training data and quickly become sexist. I propose a three-part approach to developing a system for combating gender bias in language: 1. Use theoretical and data driven methods to define types of gender bias; 2. Build classifiers to identify bias in text; 3. Build a language generation system that can analyze any bias text contains, and if necessary, suggest alternative phrasing. This approach has far-reaching applications in that it aims to foster civility in both human and machine-generated text. It also involves advancing state of the art NLP techniques by developing new methods for text processing with a focus on subtext.

Background. Most NLP research on gender focuses on the author, attempting to infer or obfuscate gender of the author [1, 2]. Studies that specifically address gender bias have examined word embeddings [3, 4]. These studies demonstrate the prevalence of bias in language by showing how a word like "receptionist" is linked more closely to "she" than "he", but they are limited to a single type of language representation. One study more broadly highlights differences in Wikipedia coverage of men and women [5], but none of these works provide a method for identifying bias in a given text.

One of the main challenges in identifying bias is its lack of a clear definition. However, understanding the relationship between social variables and language is a major field in sociolinguistics, resulting in linguistic models that language processing methods can leverage [6]. Prior work has characterized other subjective qualities, such as respect and politeness, by using hand annotators to identify traits and derive models based on linguistic theory [7, 8]. Similar techniques have the potential to identify bias in Phases 1 and 2 below.

Regarding Phase 3, removing bias from text is important not only for fostering civility in human-human interactions, but also for preventing AI systems from developing human-like biases. As demonstrated by studies of word embeddings, learning algorithms can absorb and even amplify biases present in training data [3, 4]. Recent research in dialogue systems has shown how we can use neural models to control the output of a language generation system. These methods have successfully made dialogue generation systems more consistent and more human-like [9, 10], and can help pinpoint the most relevant parts of sentences [11]. I can draw from this research to develop a method for removing bias in Phase 3.

Data. I have access to a dataset consisting of over 27M comments drawn from posts on Ted Talks, Twitter, Facebook, and Fitocracy. For example, comments on TED Talks include: "hot presenter" and "she has no inclination or understanding". Each comment is a response to a particular person and is tagged for the gender of the addressee. By looking at trends across comments addressed to men and women, I can isolate gender differences.

Phase 1. I propose to define bias by drawing on stereotypes from social science research and devising a series of questions to determine the presence of each stereotype, such as "Does the comment focus more on the speaker or on the subject matter?" I will ask annotators to answer these questions about comments from the data set through a crowd-sourcing platform like CrowdFlower. I can determine the consistency of these stereotypes by collecting multiple annotations for each comment and measuring inter-annotator agreement. I can then determine which characteristics are associated more with men or women in my collected annotations. For example, the annotations might show that comments on the speaker's appearance are more likely to be directed to women. This approach will lead to a typology of bias in a framework typical of critical discourse analysis.

Phase 2. The second part of this proposal is to develop a classifier to recognize bias, which I can model off of classifiers for similar characteristics [7, 8]. Given an input text, the ideal classifier will score the level of bias in the text and explain which features contribute to any bias detected. Understanding what motivates the classification is essential for Phase 3 and for usability in real-world applications. I can test the classifier on the annotated comments collected in Phase 1.

Phase 3. Finally, I will build a paraphrase system that can remove bias from text. I can draw from research in dialogue generation, which has devised methods for combining a generative model with a discriminator, where the generative model aims to fool the discriminator [9]. These techniques are applicable here, where the discriminator is the classifier built in Phase 2, and the generative model has the added challenge of preserving the meaning of the input text. Outputs will be evaluated for how well they reduce bias by using the classifier from Phase 2, and for how well they preserve meaning by using metrics standard in machine translation and summarization, e.g. METEOR. I will also recruit human annotators to evaluate outputs.

Intellectual Merit. This project drives at fundamental questions in NLP, including how to consider text within a social context. It will result in the creation of a bias-annotated dataset and a typology of bias, which currently do not exist. Additionally, building a bias classifier and a natural language generation system involves distinguishing core meaning from implicit sentiments. These methods are applicable to many NLP fields, like summarization and paraphrasing. Findings from this project will be presented in relevant journals and conferences and datasets will be made publically available. My background in both computer science and language as well as my past research experiences have prepared me to investigate these questions. Further, as a student at CMU LTI, I will benefit from interacting frequently with experts in the field, including my doctoral adviser, Yulia Tsvetkov.

Broader Impact. A system for reducing bias has expansive real-world applications. By helping everyone realize the potential impact of their comments, it can lead to fairness and civility in many forms, from online posts to employee performance reviews. Furthermore, socially aware language generation is critical for developing AI systems. Without it, AI systems can learn gender and racial biases, like Microsoft's AI bot Tay, which had to be shut down for perpetuating racism. While this proposal focuses on gender bias because of the availability of data, the project can also easily be expanded beyond gender to other types of bias, including race, nationality, religion, and age. Overall, this project has the potential to improve state of the art NLP techniques while building tools with real-world applications.

References. [1] Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender and variation in social media. *J Sociolinguistics*, *18*(2), 135–160. **[2]** Reddy, S., & Knight, K. (2016). Obfuscating Gender in Social Media Writing. *NLP+ CSS*, 17. **[3]** Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *American Association for the Advancement of Science*, *356*(6334), 183–186. **[4]** Bolukbasi, T. et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS*. **[5]** Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. *ICWSM*, 454–463. **[6]** Nguyen, D., Rosé, C. P., & De Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*. **[7]** Voigt, R., et al. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(25), 6521–6526. **[8]** Danescu-Niculescu-Mizil, C., et al. (2013). A computational approach to politeness with application to social factors. CoRR, abs/1306.6. **[9]** Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017). Adversarial Learning for Neural Dialogue Generation. *Proc. EMNLP*. **[10]** Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A Persona-Based Neural Conversation Model. *Proc. ACL*. **[11]** Li, J., Monroe, W., & Jurafsky, D. (2016). Understanding Neural Networks through Representation Erasure. *arXiv Preprint arXiv:1612.08220*