ABSTRACT

Title of Dissertation: Towards Effective and Inclusive AI:
Aligning AI Systems with User Needs
and Stakeholder Values Across Diverse Contexts

Yang Trista Cao
Doctor of Philosophy, 2024

Dissertation Directed by: Professor Hal Daumé III
Department of Computer Science

Inspired by the Turing test, a long line of research in AI has focused on technical improvement on tasks thought to require human-like comprehension. However, this focus has often resulted in models with impressive technical capabilities but uncertain real-world applicability. Despite the advancements of large pre-trained models, we still see various failure cases towards discriminated groups and when applied to specific applications. A major problem here is the detached model development process — these models are designed, developed, and evaluated with limited consideration of their users and stakeholders. My dissertation is dedicated to addressing this detachment by examining how artificial intelligence (AI) systems can be more effectively aligned with the needs of users and the values of stakeholders across diverse contexts. This work aims to close the gap between the current state of AI technology and its meaningful application in the lives of real-life stakeholders.

My thesis explores three key aspects of aligning AI systems with human needs and values:

identifying sources of misalignment, addressing the needs of specific user groups, and ensuring value alignment across diverse stakeholders. First, I examine potential causes of misalignment in AI system development, focusing on gender biases in natural language processing (NLP) systems. I demonstrate that without careful consideration of real-life stakeholders, AI systems are prone to biases entering at each development stage. Second, I explore the alignment of AI systems for specific user groups by analyzing two real-life application contexts: a content moderation assistance system for volunteer moderators and a visual question answering (VQA) system for blind and visually impaired (BVI) individuals. In both contexts, I identify significant gaps in AI systems and provide directions for better alignment with users' needs. Finally, I assess the alignment of AI systems with human values, focusing on stereotype issues within general large language models (LLMs). I propose a theory-grounded method for systematically evaluating stereotypical associations and exploring their impact on diverse user identities, including intersectional identity stereotypes and the leakage of stereotypes across cultures.

Through these investigations, this dissertation contributes to the growing field of human-centered AI by providing insights, methodologies, and recommendations for aligning AI systems with the needs and values of diverse stakeholders. By addressing the challenges of misalignment, user-specific needs, and value alignment, this work aims to foster the development of AI technologies that effectively collaborate with and empower users while promoting fairness, inclusivity, and positive social impact.

Towards Effective and Inclusive AI: Aligning AI Systems with User Needs and Stakeholder Values Across Diverse Contexts

by

Yang Trista Cao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Hal Daumé III, Chair/Advisor
Professor Hernisa Kacorri
Professor Katie Shilton
Professor Rachel Rudinger
Professor Kai-Wei Chang

Dedication

Psalms 121:2

Acknowledgments

In reflecting on the journey that has been my doctoral studies, I am deeply aware that this achievement is not mine alone. It is a tapestry woven from the kindness, support, and wisdom of many remarkable individuals who have guided, encouraged, and believed in me along the way. This acknowledgment section is a small token of my gratitude to each of them. From my family and friends to my mentors and peers at the university, each has played a crucial role in shaping my academic and personal growth. Here, I extend my heartfelt thanks to those who have been part of this extraordinary chapter of my life.

First off, a huge thanks to my advisor, Hal Damué III. My first dive into research was with him during the summer research experience (REU) program back in 2017. Even though the AVW building wasn't the coziest place to work (anyone who knows AVW building knows what I am talking about), the experience I had working with Hal really opened my eyes to the joys of natural language processing and led me to apply for PhD programs. I still light up thinking about the time Hal asked if I wanted to continue working with him, even though I was just a newbie in the research world at that time. I feel incredibly lucky to have spent the past six years learning from Hal what great research looks like and experiencing the best of collaboration.

I also want to extend a heartfelt thank you to my committee members—Rachel Rudinger, Katie Shilton, Michelle Mazurek, Hernisa Kacorri, and Kai-Wei Chang. Each of you has contributed uniquely and significantly to my growth and the success of my projects. Rachel, from

my very first research project on stereotypes, you've been an incredible source of ideas and inspiration, always ready to guide and enrich our work with your innovative thinking. Katie and Michelle, joining the readtheroom meetings for our content moderation project with Sarah and Lovely was always a highlight for me. Your encouragement and advice during my tougher days were like beacons of light. Working with Hernisa Kacorri was an absolute pleasure too. Our project on Visual Question Answering (VQA) for accessibility not only excited me but also made me feel like my work was making a real, positive impact. Hernisa, you taught me so much about how to engage effectively with people who have visual impairments. The lessons I've learned from you are invaluable, and I hope to keep contributing to the field of accessibility. Our time together at ASSETS 2023, where you introduced me to many fantastic researchers, was filled with enriching conversations and is something I will always cherish. And Kai-Wei, what a thrill it was to work with you during my internship at Amazon in 2021. Even though we couldn't meet in person because of COVID-19, collaborating with you was a dream come true. Your ability to find solutions when our project hit snags and to always find a positive angle helped me keep pushing forward.

Moreover, a huge thanks to all my collaborators: Anna Sotnikova, Kyle Seelman, Kyungjun Lee, Sarah Ann Gilbert, Lovely-Frances Domingo, and Jieyu Zhao. A special shout-out to Anna, with whom we worked on two projects during the entire COVID period. Those late-night calls filled with both worries and laughter were something I will never forget. Kyungjun, it was so random how we met, but our chat was so engaging that we just had to find a way to work together. You've opened my eyes to how my research can help people with disabilities, and I'm so excited to keep pushing this forward.

Also, I am more than honored to have been part of the halfolks group alongside Sudha

Rao, Amr Sharaf, Kianté Brantly, Khank Nguyne, Chen Zhao, Navita Goyal, Connor Baumler, Lingjun Zhao, Huy Nghiem, Ruijie Zheng, Yu Hou, Tin Nguyen, and Sandra Sandoval. Thank you all for the unwavering support and the wonderfully friendly atmosphere that made navigating my PhD journey so much easier and more enjoyable. Each one of you contributed to making our group truly exceptional. A special mention to Sudha, who was like a big sister to me during my first two years at UMD. Your warm welcome and invaluable guidance on my initial project were more than I could have asked for. Chen, you were the first PhD student I met during the REU program back in 2017. Despite the challenges in that dark lab room at AVW, your perseverance and determination served as a great inspiration to me. Navita and Connor, your sweet nature, witty jokes, and non-stop support and encouragement always lifted my spirits. This group has been the best part of my academic life, and I sincerely wish I could stay with you all a bit longer.

I owe much of my academic growth to the CLIP Lab at Maryland. I am profoundly thankful to have received insightful feedback on my work from the brilliant professors in CLIP, including Marine Carpuat, Philip Resnik, and Jordan Boyd-Graber. Huge thanks also go to my labmates in the CLIP lab: Sweta Agrawal, Pranav Goel, Eleftheria Briakou, Alexander Hoyle, Haozhe An, Yoo Yeon Sung, Neha Srikanth, Pedro Rodriguez, Shi Feng, Weijia Xu, Xing Niu, and Chenglei Si. Each of you has contributed significantly to my research and personal development. The previous times we've had together have not only enriched my PhD experience but have also been a source of constant motivation and joy. In particular, I want to highlight my mates who joined the same year in 2018: Sweta, Pranav, and Eleftheria. I vividly remember our shared anxiety during our first year, worrying about the uncertainty of our futures. However, we supported each other through those challenging times, and looking back now, I am amazed at how far we have come. I am so proud to be colleagues and friends with such incredibly talented individuals.

Outside of the lab, I am incredibly grateful to be surrounded by my precious friends, who have become like family to me here in Maryland: Shuhan Kou, Jiaxin Yuan, Jiajin Guan, Bo He, Haozhe An, Jerry Zheng, Lingjun Zhao, Jieyu Zhao, Xiaoyu Liu, Hou Yu, Yuancheng Xu (Prof. Zha) and Yuanyuan, Yue Feng, and Xijun Wang. Cheers to all of our crazy nights. A special mention to my roommate of five years, Shuhan Kou. Shuhan, you have been the most wonderful companion I could ever hope for. Together, we have navigated the challenges of COVID-19, shared countless memories, and raised our beloved Bobo. You have always been our cherished Aunty Kou. I deeply hope that our paths continue to intertwine long after we graduate and move on from UMD. The thought of not having you all nearby is something I will definitely miss.

I also owe immense gratitude to my undergraduate advisors: Dianna Xu and Lynn Butler. Initially a Math major, I would not have considered double-majoring had it not been for Dianna. She taught my first computer science class and her encouragement and confidence in us ignited my interest in continuing with computer science courses. Lynn was my mathematics advisor at Haverford College. Her probability class was the most challenging I have ever taken, yet it was her rigorous approach and mentorship that profoundly deepened my understanding and skills. Before meeting them, I never thought I was smart or dedicated enough for a PhD program, but both Dianna and Lynn inspired me to think, "Why not?" This pivotal encouragement eventually led me to pursue this path. I also want to extend my thanks to Professors Steven Lindell, David (Dave) Wonnacott, Robert Manning, and Joshua Sabloff. Your exceptional teaching has provided me with a lifelong educational foundation, and the “confidence-boosting shot” you gave, helping me believe in my capabilities and encouraging me not to limit myself to my own imagination.

Last but not least, I would like to express my deepest gratitude to my family: my wonderful parents and my forever loyal fans—my grandparents and my little boy, Bobo. Thank you, Baba

and Mama, for always believing in me and providing unwavering support. From my childhood, you both have given me the warmest companionship. Every night, I would share everything that happened during my day with you, and you would help me navigate through all my happiness, worries, sadness, and anxiety. This has ensured that I go to bed worry-free, a habit that continues to comfort me to this day. Even now, I feel I can still share everything with you, and you always strive to understand me. I am incredibly fortunate to be your child. Special thanks to Baba, who introduced me to God, who has become my reliance, my shepherd, and my resting home. I am here today because I have a loving and just God by my side. I know I will be okay tomorrow and beyond, because I have my Lord.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

NLP	Natural Language Processing
AI	Artificial Intelligence
HCAI	Human-Centered Artificial Intelligence
ML	Machine Learning
VQA	Visual Question Answering
BVI	Bline and Visually Impaired
LLM	Large Language Model
LPM	Large Pre-trained Model
MLLM	Multilingual Large Language Model
HF	Hugging Face
LM	Language Model
ILPS	Increased Log Probability Score
CEAT	Contextualized Embedding Association Test
SeT	Sensitivity Test
ABC	Agency-Belief-Communion
MAP	Maybe Ambiguous Pronoun
GICoref	Gender Inclusive Coreference

Chapter 1: Introduction

With the rapid development of artificial intelligence (AI) technologies, especially the advancement of large pre-trained models (LPMs), real-world usage and deployment of AI systems is expanding across many industries and applications. For example, AI algorithms are used to assist in the discovery and legal analysis of documents for both criminal and civil courts [Liu et al., 2019a], analyzing medical images to detect cancerous tumors and aid diagnostics [Kaur and Garg, 2023], and providing personalized teaching and tutoring to fill knowledge gaps for students [Kokku et al., 2018]. As the adoption of AI technologies continues to expand, they are set to have substantial and wide-reaching impacts on both individuals and society. This growth brings with it pressing concerns about the ethical implications and societal consequences of these technologies [Garibay et al., 2023]. For instance, large language models have the capability to generate highly persuasive and impactful propaganda messages at a lower cost [Goldstein et al., 2023]. Similarly, recommendation systems have the potential to perpetuate social biases by influencing user preferences and guiding choices [Milano et al., 2020]. Studies also indicate that human judgments and inferences can be influenced by reasoning about model behaviors [Shafto et al., 2012, Sun et al., 2020]. Given AI's potential for such significant impact, the AI research field bears the responsibility to develop systems that support positive social outcomes [of Economic and Affairs, 2018]. In particular, AI systems should be designed to augment and amplify

human abilities while ensuring their safe and ethical usage [Wang, 2019, Shneiderman, 2020].

Meanwhile, the AI field has been focusing on algorithmic improvement. From the Turing test to pursuing Artificial General Intelligence, model developers have been focusing on enhancing models' capability to mimic human-like behaviors, such as image tagging and captioning, and various forms of reasoning. There is no doubt about the necessity of improving models' technical capacities, and we have seen great strides in models' abilities in recent years. However, it is not yet clear how these increasingly capable models will lead to positive social impact beyond technical achievements.

Hence, there has been a notable increase in efforts to align AI with human stakeholders, aiming to enhance the usability of AI systems. According to ISO 9241-11 [ISO, 1998], usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." In this framework, a useful AI system should not only *align with users' needs* to foster effective and efficient human-AI collaborations but also *align with stakeholders' values* to ensure that these collaborations are both safe and enjoyable. Addressing both aspects is particularly challenging, given the specialized and diverse nature of the stakeholders involved.

On one hand, AI system users are specialized, each possessing unique skills, needs, and use cases that drive their interactions with these systems. For example, healthcare professionals require interpretability to understand system logic before trusting diagnosis recommendations [Rudin and Radin, 2019]; users with disabilities need accessibility-aware interfaces [Oswal, 2019]; businesses expect integration with existing workflows and protection of sensitive data [Cheng et al., 2017]. Hence, a universal, one-size-fits-all approach to AI frameworks is impractical. It is essential to investigate methodologies to accurately capture users' specific needs

and to innovate technical approaches that are in alignment with these needs.

Meanwhile, the stakeholders of AI systems — encompassing both individuals directly and indirectly affected by these systems — represent a broad spectrum of backgrounds, identities, and values. AI systems that neglect to integrate these diverse values run the risk of causing harm to various stakeholders [Friedman and Nissenbaum, 1996a], particularly to marginalized groups. This is evidenced in previous literature, which includes reports of systems generating Islamophobic content [Samuel, 2021], auto-suggesting racist and sexist web search queries [Hazen et al., 2020], and producing images with limited cultural diversity that reinforce stereotypes [Jha et al., 2024a]. Moreover, the AI Incident Database [McGregor, 2020] has over 3,000 similar incidents reported up to 2024 [Incident Database, 2024]. Additionally, the complexity intensifies when considering that the values of different stakeholders sometimes conflict [Fleisig et al., 2023]. This necessitates a careful balancing of AI systems that not only align with users' values but are also universally safe and responsible, incorporating the values of all stakeholders[1].

To address these challenges, my thesis delves into both aspects of *alignment with users' needs* and *alignment with stakeholders' values*, aiming to bridge the existing gap between current AI technologies and their deployment to real-life stakeholders. I explore users' and stakeholders' needs and values under four distinct settings, each approached from a model improvement perspective. First, I investigate possible causes of misalignment in AI system development, particularly focusing on how a natural language processing (NLP) system may misalign with real-life stakeholders' perceptions of gender. Then, to understand how AI systems can be tailored to meet specific user needs, I analyze two real-life AI applications with their specific user groups: a con-

[1]The term "values" can encompass different meanings, including ethical values and instrumental values. In my thesis, unless stated otherwise, "value" specifically refers to ethical values.

tent moderation assistance system used by volunteer content moderators and a visual question answering (VQA) system for blind and visually impaired (BVI) users. In the content moderation use case, my research centers on aligning the system's functionalities with the moderators' needs to facilitate effective collaboration. For the visual question answering system, I investigate the alignment challenges when adapting to users with disabilities and thus unique requirements. Finally, to explore how to develop safe AI systems that align with human values, I assess the stereotype problem within large language models (LLMs), which are widely used and have a broad impact across diverse groups of people. I propose a stereotype measurement methodology to investigate prevailing stereotype issues within LLMs.

1.1 Bias Entering Machine Learning Lifecycle

In the development of AI systems, developers tend to focus on algorithmic technical advancement while overlooking or making false assumptions about their system stakeholders. Hence, when such systems are deployed, they can easily result in misaligned systems that can cause harm to stakeholders, particularly to those minority groups. In my first study, I investigate how such false assumptions can cause a misaligned system through each stage of the machine learning (ML) lifecycle. In particular, I focus on gender assumptions in natural language processing (NLP) systems.

NLP systems tend to assume that gender is binary (male or female), immutable (gender is assigned at birth and never changes), and, thirdly, that social gender is in perfect correspondence to gendered linguistic forms, which is not how people actually experience gender in the real world. First, gender is *not* binary — there are, for example, agender people who do not experience

gender. It is not immutable either — for example, there are genderfluid people whose gender identity may shift by time or context. And the inner-relationships of social gender and gendered linguistic forms can be very complicated. I thus first disentangle the concepts of gender with respect to sociological and sociolinguistic gender and discuss how they are realized in languages in connection to gender bias in NLP.

Further, I delve into the coreference resolution systems as an illustrative example. Coreference resolution systems are NLP systems that determine which textual references resolve to the same real-world entity. When those entities are people, coreference resolution systems run the risk of making unlicensed inferences, possibly resulting in harms to system stakeholders. I thus assess six stages of the ML lifecycle, including task definition, data collection, data labeling, model definition, model evaluation, and feedback loop, to investigate how false assumptions of gender can cause gender biases to enter the system. Take the data labeling stage as an example, I conduct a human study with an ablation technique. I aim to identify which forms of linguistic knowledge are crucial for human annotators when making coreference judgments, and to understand how these can result in problematic training datasets for the system. Findings from this study demonstrate that without careful consideration of the system's real-life stakeholders, we (system developers) may introduce gender biases into every stage of the machine learning lifecycle when developing systems and thus lead to a misaligned and unsafe system.

1.2 Alignment with Users' Needs through Specialized User Contexts

In my second line of study, I seek to explore how AI systems can be tailored to the unique needs of specialized user groups. This exploration is conducted in two different application

contexts, each involving distinct user groups: volunteer content moderators using moderation detection assistance systems, and blind and visually impaired (BVI) users engaging with visual question answering (VQA) systems. The first context focuses on AI systems enhancing humans in accomplishing the moderation task, whereas the second context focuses on AI systems complementing human abilities by providing visual information to BVI users.

Content moderators guard online forums against hostility and extremism while maintaining community norms, ensuring the forums remain healthy and open to all participants. However, with the workload and continuous exposure to online harassment Gilbert [2020], Wohn [2019], Dosono and Semaan [2019] and toxic content, this moderation work can easily result in burnout. Thus, with the goal of using AI technologies to lighten the load for moderators, the NLP field has a long line of study on developing automated systems to identify toxic, offensive, and hateful content [e.g. Jigsaw, 2019a, Pavlopoulos et al., 2020]. Yet, content moderation encompasses more than toxicity detection, particularly in platforms that leverage volunteer moderation within smaller communities hosted by the site. For example, Reddit is a platform consisting of various communities, known as "subreddits," focused on a diverse set of topics, and each subreddit has its own moderation rules. Hence, in order to have a system that aligns with moderators' needs in detecting potential rule-violating content, such a system needs to support much more than just toxicity detection. In this study, I thus aim to assess to what extent current NLP models can serve the needs of a wide spectrum of moderation rules that exist. To do so, I conduct a model review to reveal the availability of existing NLP models to cover various moderation rules and guidelines from three exemplar Reddit forums. I further put state-of-the-art LLMs (GPT and Llama) to the test, evaluating how well these models perform in flagging violations of platform rules from one particular forum. Finally, I conduct a user survey study with volunteer moderators to gain insight

into their perspectives on useful moderation systems.

On the other hand, visual question answering (VQA) is a task mostly developed to examine machine "understanding", i.e. machines' multi-modal comprehension abilities, but found to be potentially useful for blind and visually impaired (BVI) users in answering their questions about the visual world in real-time. VQA systems are designed to answer natural language questions based on image information. However, it is unclear whether VQA systems built for testing multi-modal comprehension abilities can align with the needs of BVI users. For example, previous studies have shown the information BVI users seek for help and the images they take vary greatly compared to the training datasets of machine "understanding" VQA systems [Brady et al., 2013, Zeng et al., 2020]. Thus, to explore the alignment gap, I conduct performance assessments on seven machine "understanding" VQA model architectures spanning from 2017 to 2021 using an accessibility VQA dataset, in which the image-question pairs are created by BVI users. I aim to investigate whether advancements in model architectures designed for machine understanding VQA also yield performance improvements on accessibility VQA datasets. I also conduct a mixed methods error analysis to identify particularly challenging classes as directions for future VQA systems to improve on.

In both contexts, I assess the corresponding AI systems in each use case with data collected from the users to understand potential challenges for adapting technology-focused systems to meet application-specific user needs. I also discuss directions to tailor the systems to better align with users' needs. As experts of their own experiences, specialized users (both through user feedback and data collected from them) offer invaluable perspectives on how AI can better empower them and align with their needs.

1.3 Alignment with Diverse Stakeholders' Values in General Large Language Models

Both general models and models for specific applications face stakeholders from diverse backgrounds, representing a wide spectrum of identities. Thus, beyond alignment with needs, to ensure a safe and enjoyable human-AI interaction, AI systems need to align with diverse human values of the stakeholders. However, LLMs trained with a large amount of internet texts easily pick up social norms and conventions that underline the data, which includes both factual information as well as social stereotypes about race, gender, religion, etc. For example, we have found bidirectional encoder representation transformers (BERT) associate groups like "Muslims" and "Black people" with more negative traits on average compared to other groups. Such biases may lead AI systems to make decisions that deny opportunities or resources to marginalized groups [e.g. Eubanks, 2017, Tamkin et al., 2023]. Thus, it is crucial we mitigate these potential harms. To do so, we first need a clear understanding of what we are measuring and seeking to mitigate.

Thus, in my research, I develop a systematic approach for assessing stereotypes toward diverse identities within Large Language Models (LLMs). I adopt the Agency-Belief-Communion (ABC) stereotype model Koch et al. [2016] from social psychology as a framework for the systematic study and discovery of stereotypic group-trait associations in LLMs. To measure these associations, I introduce a new measurement, the Sensitivity Test (SeT), which captures the robustness of word associations in large language models besides probabilities of co-occurrence. I also crowdsource to collect human stereotypes and develop a screening approach to filter out

unreliable annotations while keeping annotators' various perspectives on stereotypes. Since this measurement has the advantage of being easy to generalize to unstudied groups, I take a step towards exploring stereotypes of intersectional identities, finding some correspondence between model behavior and the literature on intersectional stereotypes. In my overarching aim to examine system alignment concerning stakeholders with diverse identities, I propose this approach to systematically evaluate stereotypical associations within LLMs.

Going beyond English stereotypes within the U.S. culture, LLMs are increasingly being trained on texts from diverse languages and cultures, where each culture possesses its own set of stereotypes. A potential implication of training models multilingually is the transmission of harmful stereotypes across cultures. This may reinforce existing stereotypes among stakeholders of the model, as well as create new stereotypes that are transmitted from a different language. I thus utilize the proposed measurement approach to assess how cultural stereotypes leak across various languages in multilingual language models (MLLMs). In this study, I define stereotype leakage as the effect of stereotypic word associations in MLLMs of one language impacted by stereotypes from other languages. I specifically examine stereotype leakages in English, Russian, Chinese, and Hindi within the models mBERT [Müller et al., 2020], mT5 [Xue et al., 2020], and GPT3.5 [Brown et al., 2020, OpenAI, 2023]. I conduct a human study in four languages to collect cultural human stereotypes and measure stereotypic associations in MLLMs. Interestingly, findings revealed that GPT3.5 is the model most affected by the influence of human stereotypes, encompassing both stereotype leakages and stereotypes originating from the target language itself. Additionally, it's noteworthy that Hindi, which is the relatively low-resourced language among the four languages, exhibits the most significant stereotype leakage. With the widespread usage of GPT models, these findings emphasize the necessity to align systems with stakeholders'

diverse identities and to be cautious regarding stereotype leakages, particularly from dominant cultures.

1.4 Summary of Contributions

This thesis makes several key contributions to the field of human-centered artificial intelligence, focusing on the alignment of AI systems with human needs and values:

- Revealing user needs in specific AI applications and identifying gaps in existing AI systems: Through in-depth studies of two specific user groups — blind and visually impaired users of visual question answering systems and volunteer content moderators using moderation assistance tools — I have uncovered unique user requirements and highlighted the shortcomings of current AI technologies in meeting these needs. The frameworks developed for studying these user groups can be extended to investigate user needs across a wide range of AI applications.
- Proposing methodologies for assessing bias and stereotypes in large language models: I have introduced novel datasets and methods to systematically evaluate bias and stereotypes in AI systems. These contributions include the machine learning lifecycle framework for identifying sources of bias, the Agency-Belief-Communion (ABC) model for measuring stereotypical associations, and techniques for assessing stereotype leakage across cultures in multilingual models. These methodologies provide a foundation for researchers and developers to identify and mitigate harmful biases and stereotypes in AI systems.
- Emphasizing the importance of a human-centered approach in AI development: By demonstrating the significance of aligning AI systems with human needs and values across diverse

contexts, this thesis underscores the critical role of human-centered design in the development of AI technologies. The insights and methodologies presented here contribute to the growing body of research aimed at creating AI systems that effectively collaborate with and empower human users while promoting fairness, inclusivity, and positive social impact.

These contributions advance our understanding of the challenges and opportunities in aligning AI systems with human needs and values, providing valuable tools and insights for researchers, developers, and stakeholders working towards the responsible development and deployment of artificial intelligence technologies.

1.5 Roadmap

This thesis is organized into five chapters. Chapter 2 provides background on AI alignment, discussing key concepts and related work in alignment. Chapter 3 investigates gender bias in NLP systems, examining how biases can enter at each stage of the machine learning lifecycle. Chapter 4 explores the alignment of AI systems with specific user needs in two application contexts: content moderation assistance for volunteer moderators and visual question answering for blind and visually impaired users. Chapter 5 addresses the alignment of AI systems with human values, focusing on the assessment of stereotypes in large language models across diverse user identities and cultural contexts. Finally, Chapter 6 concludes the thesis, summarizing the main findings, contributions, and future directions.

Chapter 2: Background

The widespread adoption of personal computers since the 1980s initially captivated people with their technical advancements but soon highlighted significant usability issues. Users frequently expressed frustration with unintuitive computer interfaces and complex features [Baecker et al., 1995]. To address these challenges, the field of Human-Computer Interaction (HCI) was established, initially aiming to overcome these considerable obstacles in computer software design [Xu et al., 2021]. This led to developments such as graphical user interfaces (GUIs), which simplified user interaction, and the creation of user-friendly desktop environments and interactive devices like the mouse, vastly improving the user experience.

As pointed out by Xu et al. [2021], the pattern is repeating with AI development. Since the Turing test [Turing, 1950], artificial intelligence (AI) technologies have been focusing on algorithmic advancement aiming to simulate human cognitive abilities and behaviors. Humans are thus viewed as “cognitive machines” that could be mimicked through AI systems [Winograd, 2006]. Early "Good Old Fashioned AI" (GOFAI) approaches focused on symbolic reasoning, logic systems, and knowledge representation to model human expertise [Haugeland, 1989]. As computational power increased, statistical and machine learning models demonstrated more robust capabilities on specialized tasks like computer vision and natural language processing. In recent years, the development of deep neural networks, as well as large pre-trained language mod-

els, have demonstrated remarkable advances on benchmark tests of perception, reasoning, and language understanding [Bommasani et al., 2021]. Advances in AI have thus clearly expanded computational capabilities, yet translation to real-world benefits outside of narrow applications remains unclear. According to a survey by the Pew Research Center [Rainie et al., 2022], "less than half of the public believes technologies would improve things over the current situation." Key concerns among the public include the safe and ethical usage of the technologies, human over-reliance on them, and uncertainty about the real social impact they may bring.

To address these issues, researchers suggested once again going back to humans — taking a human-centered approach to improve human-AI collaborations, which led to the emergence of human-centered AI (HCAI) as an important area of research [Shneiderman, 2022a]. It emphasizes the complexity of human users, both internally, as they have different demographic backgrounds, prior experiences, needs, and interests, and externally, as they interact with various social environments and contexts [Auernhammer, 2020]. Recognizing this complexity, researchers in HCI sought a human-centered, design-oriented approach to improve human-computer interactions [Winograd, 2006]. Specifically, the field examines how users engage with computational systems and aims to make these interactions as effective, efficient, safe, and satisfying as possible.

Within the broad spectrum of Human-Centered Artificial Intelligence (HCAI), a critical topic is AI system alignment. The concept of alignment first emerged in discussions about superintelligence, emphasizing the moral alignment of AI systems with human values, particularly focusing on concepts like friendliness [Bostrom, 2003, Yudkowsky, 2004]. Subsequent research has explored various facets of alignment. In the realm of reinforcement learning, the focus has been on agent intention alignment, wherein reward functions are developed to ensure AI agents align with human values and intentions in task execution [Dewey, 2011, Leike et al., 2018]. The

advent of large pre-trained models has further broadened the scope of alignment, introducing dimensions such as safety and harm. Gabriel [2020] suggested that alignment should encompass human instructions, intentions, revealed preferences, ideal preferences, interests, and values. Terry et al. [2023] examined alignment in interactive AI systems, considering it through user interaction cycles as specification alignment, process alignment, and evaluation support. In my dissertation, I concentrate on AI systems' alignment with users' needs, akin to intention alignment, as well as alignment with human ethical values.

2.1 Alignment with User Needs

Emerging AI technologies are unlocking promising new possibilities for human-machine collaboration. AI systems can bring multifaceted benefits when thoughtfully integrated to complement human strengths. It can be used to reduce human labor for repetitive and tiring work. For example, content moderation is a timely- and emotionally-consuming work for moderators. AI technology can help alleviate moderation burdens and reduce moderator burnout [Kiene and Hill, 2020, Seering et al., 2018]. For example, prior study has shown that automation is helpful when communities experience unprecedented or unexpected growth, such as in cases where communities receive an influx of new users [Kiene et al., 2016] and in cases where communities are subject to sustained brigades of bad actors spamming hateful content [Han et al., 2023]. Moreover, AI can complement human skills in areas such as data analysis. Systems can rapidly process volumes of data unmanageable for people to surface insights around trends, correlations, and patterns [Cui et al., 2021]. Humans then leverage strengths like contextual reasoning and creative abstraction to interpret implications to inform decisions or recommendations. This demonstrates

how AI and human collaborators can enhance one another — AI expands analytic breadth and capacity while humans supply guidance, oversight, and nuanced perspective. Furthermore, AI can enhance human capabilities, such as creativity. In creative writing, for instance, AI writing assistants can provide idea-generation support by offering inspirational prompts and suggestions when authors experience lulls in their workflow [Clark et al., 2018].

However, adapting technologies to real-life applications presents significant challenges. One key issue is the frequent misalignment between system functions and users' needs. For instance, users often need better system explanations or greater transparency. In AI programming assistants, a user study revealed that users face difficulties in debugging generated code and integrating it into their projects due to insufficient understanding of the system's capabilities or framework [Lin, 2017, Ferdowsifard et al., 2020]. Another challenge is the diminished human agency in certain contexts [Shneiderman, 2022b]. A study [Liang et al., 2024] found that software developers struggled to steer the system towards generating desired outputs. Furthermore, the need for human agency versus automation can vary widely depending on the application, the user, and their specific use cases [Kang and Lou, 2022]. These challenges highlight the necessity for AI technologies to be designed with a user-centric approach in order to align with users' needs and be truly useful.

2.2 User Involved Design

To tackle these challenges, researchers have developed various methods for incorporating user perspectives into system design. The concept of user-involved design was first introduced in the book "Participatory Design," which highlights the importance of involving end-users in

the design process, particularly those who will utilize the technology in their everyday lives and work [Schuler and Namioka, 1993]. Kujala [2003] further developed a taxonomy of different user-involved design approaches, analyzing their respective challenges and advantages. Amershi et al. [2019] propose applicable guidelines for designing AI systems with users in the loop.

One fundamental approach in this domain is "usability engineering" [Wixon and Wilson, 1997], which involves defining, measuring, and enhancing a system's usability by incorporating user input throughout the process. This approach shares similarities with the modern process of machine learning (ML) model development. In a similar vein, Gould and Lewis [1985] promote a user-centered design approach, emphasizing the importance of focusing on potential users early in the development process to prevent unnecessary efforts.

Going beyond mere user input, participatory and co-operative design methodologies involve a deeper collaboration between designers and users [Floyd et al., 1989]. In these approaches, both parties work together in planning and crafting the design, ensuring a more holistic and user-centric development process. Such approaches have been successfully implemented in various fields. For instance, in healthcare, participatory design involves patients, caregivers, and healthcare professionals in shaping healthcare services and facilities [Neuhauser and Kreps, 2011]. Similarly, in the realm of educational technology, participatory design is utilized to incorporate the perspectives of teachers, students, and education experts into the development of tools and platforms [Christian Dindler and Iversen, 2020].

Extending beyond considering the needs and feedback of direct end-users, value sensitive design methods involve identifying and integrating the values of all stakeholders, both direct and indirect, in the design process [Friedman et al., 2017]. They acknowledge that technology design and deployment can have far-reaching impacts and thus aim to be proactive in understanding and

addressing the values, needs, and ethical considerations of a broader group of stakeholders.

2.3 Accessibility

Building upon the ISO [1998] definition of usability, its revision emphasizes the significance of user experience [Bevan et al., 2015]. Notably, this revision broadens the scope of usability by incorporating *Accessibility*, mandating systems to be "effective, efficient and satisfying for users with the widest range of capabilities." However, given the diverse and progressive nature of disability, coupled with the challenges in collecting and understanding data on disability, there exists a persistent risk that the development of AI technologies may leave people with disabilities underserved.

Disability is not homogeneous — individuals may experience a range of disabilities, each varying in intensity and impact, which can also fluctuate over time [Trewin, 2018]. This makes it challenging to collect representative data to cover all disabilities. This diversity presents a significant challenge in collecting data that accurately represents all types of disabilities. Additionally, as reported in the case study by Blaser and Ladner [2020], there is a lack of standardized practices for collecting disability data. Many studies prioritize demographics like gender and race, often overlooking disability data. Furthermore, disability information can be sensitive; in some instances, it's considered confidential, and Institutional Review Boards (IRBs) may restrict its collection.

Consequently, disability data is frequently absent in many studies. This gap poses a risk in the development of AI systems, where data concerning disabilities might be misinterpreted as outliers and disregarded [Trewin, 2018]. Such oversight can lead to considerable frustrations for

disabled individuals when using these AI systems. For instance, Glazko et al. [2023] observed that generative AI systems often struggle to produce authentic-looking happy pictures of people with disabilities. Additionally, they noted that these systems frequently require human verification, a task that can be particularly challenging for people with disabilities. This requirement significantly reduces the utility of these systems for people with disabilities. Furthermore, privacy and security is another significant concern when it comes to sharing disability data. Such data often includes sensitive personal information, yet there is a noticeable absence of accessible methods for users to review and give their consent for data sharing [Kamikubo et al., 2023].

Many studies have contributed to closing the gap. To expand the dataset for accessibility, Park et al. [2021] proposed design guidelines for collecting data from people with disabilities through online infrastructure, and Kacorri et al. [2020] constructed a data surfacing repository to aggregate accessibility datasets. Many datasets on disability were also published, such as the visual question answering dataset collected from blind people [Bigham et al., 2010] and American sign language datasets collected from deaf people [Desai et al., 2023, Martinez et al., 2002].

2.4 Alignment with Human Values

Aligning AI systems with user needs enables the systems to augment human capabilities in various task performances. In addition to that, ensuring these systems align with human values is also crucial for fostering safe and enjoyable interactions for all the stakeholders. This prevents the systems from causing direct harm to users, such as by producing harmful, insulting, or stereotypically biased language. Moreover, it diminishes the risk of AI systems perpetuating or

exacerbating the cycle of reinforcing stereotypes. Hence, a great number of studies have been dedicated to understanding and quantifying the fairness levels of systems, as well as to developing strategies to improve the alignment.

2.4.1 Fairness and Inclusiveness

Gender bias stands as the most explored domain within the field of fairness. It has been the subject of extensive research in various NLP systems, including but not limited to coreference resolution [Rudinger et al., 2018, Levesque et al., 2012, Zhao et al., 2018a, Webster et al., 2018], natural language inference [Rudinger et al., 2017], occupation classification [De-Arteaga et al., 2019a], sentiment analysis [Kiritchenko and Mohammad, 2018], and machine translation [Font and Costa-jussà, 2019, Prates et al., 2019, Dryer, 2013, Frank et al., 2004, Wandruszka, 1969, Nissen, 2002, Doleschal and Schmid, 2001]. Overall, these studies reveal that NLP systems tend to encode substantial gender biases, manifesting in ways such as associating gender pronouns with stereotypical roles and professions, and defaulting to male references in gender-neutral scenarios in machine translation. Additionally, some researches also investigate gender biases in general language models, both intrinsically through word embedding associations [Bolukbasi et al., 2016a, De-Arteaga et al., 2019a, Gonen and Goldberg, 2019] and extrinsically in generative outputs [Tamkin et al., 2023]. However, much of this research has traditionally viewed gender bias through a binary lens, often implicitly. This limitation has been highlighted by Larson [2017a] and May [2019].

Going beyond gender bias, studies have also delved into other domains of identity, notably racial bias. The majority of these studies concentrate on the dichotomy of "Black" versus "White"

or "African American" versus "Caucasian American" [Field et al., 2021]. Findings from this body of research indicate that systems often exhibit representational biases against African Americans [Caliskan et al., 2017a, Manzini et al., 2019]. Several studies have also demonstrated that systems tend to underperform when processing African American English (AAE) [Blodgett et al., 2018, Groenwold et al., 2020, Sandoval et al., 2023], frequently misclassifying AAE as toxic or abusive [Davidson et al., 2019, Dixon et al., 2018].

In recent years, there has been an increasing focus on understanding biases in AI systems towards individuals with disabilities. Research in this area has highlighted various instances where AI systems demonstrate shortcomings in handling data related to people with disabilities across different tasks [Guo et al., 2020]. For instance, it has been observed that language models frequently misinterpret content about disabilities as toxic [Hutchinson et al., 2020] and tend to associate these identities with negative sentiments [Venkit et al., 2022]. Additionally, Massiceti et al. [2023] have revealed that Large Pre-trained Models (LPMs) are less adept at recognizing objects associated with disability compared to those not linked to disabilities, indicating a gap in the system's holistic representation of this demographic.

Recent studies have broadened the scope to encompass a wide range of identities, assessing stereotypical biases manifested in AI systems across these diverse domains. Research by Nadeem et al. [2020] and Nangia et al. [2020] has delved into identities related to nationality, race, religion, profession, orientation, disability, age, appearance, socioeconomic status, and gender. They proposed datasets specifically designed to evaluate whether large language models exhibit stereotypical biases toward these diverse identity groups. Additionally, Li et al. [2020] have probed transformer-based question answering models for stereotypes related to gender, nationality, religion, and ethnicity. Furthermore, Sotnikova et al. [2021a] explored stereotypes through inference

tasks, examining how potential targets of stereotypes are portrayed in realistic, neutral contexts.

The predominant focus on English in prior studies of gender bias in NLP overlooks the nuances and potential insights offered by other languages. Moving beyond the context of US culture, various works have examined stereotypes from different cultural perspectives. Levy et al. [2023] and Câmara et al. [2022] delve into stereotypic biases in sentiment analysis systems across multiple languages. Zhao et al. [2020] introduced a dataset to investigate bias in multilingual word embeddings across four languages. Additionally, González et al. [2020] presented a template-based anti-reflexive gender bias dataset for examining pronoun translations in four different languages. In a broader scope, Dev et al. [2023] and Jha et al. [2023] created datasets encompassing stereotypes from 178 countries. Extending their research further, Jha et al. [2024b] explore visual stereotypes on a global scale.

2.4.2 Fairness Assessment Methods

To quantify these biases, various fairness metrics and datasets have been proposed. They can be roughly categorized into two categories: *extrinsic* and *intrinsic* metrics [Goldfarb-Tarrant et al., 2021]. Intrinsic fairness metrics probe into the fairness of the language models [Guo and Caliskan, 2021, Kurita et al., 2019, Nadeem et al., 2020, Nangia et al., 2020], whereas extrinsic fairness metrics evaluate the fairness of the whole system through downstream predictions and generations [Dhamala et al., 2021, Jigsaw, 2019b, De-Arteaga et al., 2019b]. Extrinsic metrics measure the fairness of system outputs, which are directly related to the downstream bias that affects end users. However, they only inform the fairness of the combined system components, whereas intrinsic metrics directly analyze the bias encoded in the contextualized language mod-

els. In this section, I will discuss both intrinsic and extrinsic fairness evaluation methods.

2.4.2.1 Intrinsic Fairness Assessment

For measuring intrinsic model fairness, past work has generally taken one of two approaches. The first approach tests systems with hand-constructed templates like "The `[group]` is □", where `[group]` ranges over social groups (e.g., "woman" or "Hispanic"), and □ represents a "masked word" and ranges over occupations ("a professor" or "a nurse") [e.g., Bolukbasi et al., 2016b, May et al., 2019] or associations drawn from implicit association tests (IAT) (e.g., pleasant/unpleasant words or career/family-related words) [e.g., Caliskan et al., 2017b, Guo and Caliskan, 2021]. The second approach collects more natural sentences containing stereotypes, either by web crawling with crowdworkers annotations for social bias [Sap et al., 2019] or by having crowdworkers directly write stereotyping sentences [Nangia et al., 2020, Nadeem et al., 2020]. Both approaches measure stereotypes by calculating the word associations between the group identities and the stereotypic content. Following are two commonly-used word association measurements.

Increased Log Probability Score (ILPS) quantifies word associations in language models through masked word probabilities. It calculates the association score with a pre-defined template, "`[Group]` are □." [Kurita et al., 2019], where □ is a masked token. For example, given a group "Asian" and a trait `smart`, $P(\text{“}Asian\text{”}, \texttt{smart})$ measures the probability of `smart` given "Asians" by filling in the template. Since this probability is affected by the prior probability of `smart`, ILPS normalizes this probability by the "prior" probability of the trait given a masked group, as below:

$$\text{ILPS}(\mathtt{g}, \mathtt{t}) = \log \frac{P(\square = \mathtt{t} \mid \mathtt{g} \text{ are } \square.)}{P(\square_2 = \mathtt{t} \mid \square_1 \text{ are } \square_2.)} \tag{2.1}$$

Intuitively, ILPS measures how much each group raises the likelihood of a trait filling in the template. One can easily show that this equivalent to the *weight of evidence* of the trait in favor of the hypothesis that the group is the target: $s(\mathtt{g}, \mathtt{t}) = \text{woe}(\mathtt{g} : \mathtt{t} \mid \text{template})$ [Wod, 1985].

Contextualized Embedding Association Test (CEAT) estimates word associations with word embedding distances [Guo and Caliskan, 2021] . Intuitively, CEAT measures whether some groups are closer to certain traits in a latent vector space. Given two sets of target words defining groups X, Y (e.g. $X_{\text{male}} = \{$"man", "father", ...$\}$, $Y_{\text{female}} = \{$"woman", "mother", ...$\}$) and two sets of polar traits A, B (e.g. $A_{\text{pleasant}} = \{$ love, peace, ...$\}$, $B_{\text{pleasant}} = \{$ evil, nasty, ... $\}$), CEAT computes the effect sizes of the difference between X and Y being closer to A than B and corresponding p-values. Since contextualized word representations are affected by the contexts around the word, for each word in the four word sets, CEAT randomly samples 1000 sentences from Reddit, in which the word appears, and uses these to approximate the true effect size as below:

$$\text{CEAT}(A, B, X, Y) = \frac{\hat{\mathbb{E}}_{\mathtt{g} \sim X} s(\mathtt{g}, A, B) - \hat{\mathbb{E}}_{\mathtt{g} \sim Y} s(\mathtt{g}, A, B)}{\hat{\mathbb{S}}_{\mathtt{g} \sim X \cup Y} s(\mathtt{g}, A, B)} \tag{2.2}$$

$$s(\mathtt{g}, A, B) = \hat{\mathbb{E}}_{\mathtt{t} \sim A} \cos(\mathtt{g}, \mathtt{t}) - \hat{\mathbb{E}}_{\mathtt{t} \sim B} \cos(\mathtt{g}, \mathtt{t})$$

$\hat{\mathbb{E}}$ (resp. $\hat{\mathbb{S}}$) is the empirical expectation (resp. standard deviation), and $\mathtt{x}$ denotes the embedding

of $\mathtt{x}$.

In our setting, since we care about social bias among multiple groups rather than the difference between two groups, we modify the CEAT to calculate the effect size of the distance difference between $\mathtt{g}$ with A and B for each group as below:

$$\text{CEAT}(\mathtt{g}, A, B) = \frac{\hat{\mathbb{E}}_{\mathtt{t}\sim A} \cos(\mathtt{g}, \mathtt{t}) - \hat{\mathbb{E}}_{\mathtt{t}\sim B} \cos(\mathtt{g}, \mathtt{t})}{\hat{\mathbb{S}}_{\mathtt{t}\sim A\cup B} \cos(\mathtt{g}, \mathtt{t})} \tag{2.3}$$

2.4.2.2 Extrinsic Fairness Assessment

Extrinsic fairness assessment methods evaluate a model's fairness based on its outputs, which can be either classification or generative outputs. In classification tasks, the primary focus is on addressing quality of service or allocation harms that may arise from model biases. As such, these metrics are designed to determine whether the model makes fair decisions across different groups. Common metrics include the demographic parity score, which measures the probability disparity of being classified with a positive label among various groups [Feldman et al., 2015], equal opportunity, assessing the variance in True Positive Rates (TPR) between groups [Hardt et al., 2016], and treatment equality, which examines the differences in the ratio of False Negative Predictions (FNR) to False Positive Predictions (FPP) among different groups [Berk et al., 2017].

In generative tasks, fairness metrics aim to address biases impacting quality of service or leading to representational harms. These metrics critically analyze whether models produce unwarranted stereotypical assumptions in their outputs. For instance, in question answering tasks, research often evaluates whether the model defaults to stereotypes instead of factual information,

particularly in ambiguous scenarios [Parrish et al., 2022]. In constrained text generation, studies investigate whether the model's content is unduly swayed by a character's identity [Dhamala et al., 2021, Yang et al., 2023]. Similarly, in natural language inference tasks, the focus is on examining whether the model makes inappropriate inferences influenced by people's identities [Sotnikova et al., 2021b, Sap et al., 2019].

Chapter 3: Bias Entering Machine Learning Lifecycle

Joint work with Hal Daumé III. Appeared at Computational Linguistics 2021

In pursuing the goal of aligning artificial intelligence systems with human needs and values, it is crucial to identify and address potential sources of misalignment. In this chapter, we examine how biases, particularly gender biases, can enter natural language processing systems at various stages of the machine learning lifecycle. We also discuss how these misalignments may harm various stakeholders of the systems. We use coreference resolution systems as an example to illustrate these issues.

3.1 Overview

Coreference resolution—the task of determining which textual references resolve to the same real-world entity—requires making inferences about those entities. Especially when those entities are people, coreference resolution systems run the risk of making unlicensed inferences, possibly resulting in harms either to individuals or groups of people. Embedded in coreference inferences are varied aspects of gender, both because gender can show up explicitly (e.g., pronouns in English, morphology in Arabic) and because societal expectations and stereotypes around gender roles may be explicitly or implicitly assumed by speakers or listeners. This can

lead to significant biases in coreference resolution systems: cases where systems "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [Friedman and Nissenbaum, 1996b, p. 332].

Gender bias in coreference resolution can manifest in many ways; work by Rudinger et al. [2018], Zhao et al. [2018a], and Webster et al. [2018] focused largely on the case of *binary* gender discrimination in trained coreference systems, showing that current systems over-rely on social stereotypes when resolving HE and SHE pronouns[1] (see section 3.2). Contemporaneously, critical work in Human-Computer Interaction has complicated discussions around gender in other fields, such as computer vision [Keyes, 2018, Hamidi et al., 2018].

Building on both lines of work, and inspired by Keyes's [2018] study of vision-based automatic gender recognition systems, we consider gender bias from a broader conceptual frame than the binary "folk" model. We investigate ways in which folk notions of gender—namely that there are two genders, assigned at birth, immutable, and in perfect correspondence to gendered linguistic forms—lead to the development of technology that is exclusionary and harmful of binary and non-binary trans and non-trans people.[2] We take the normative position that while the folk model of gender is even today wide-spread, building systems that adhere to it implicitly or explicitly can lead to significant harms to binary and non-binary trans individuals, and that we should aim to understand and minimize those harms. We take this as particularly important not just from the perspective of potentially improving the quality of our systems when run on doc-

[1]Throughout, we avoid mapping pronouns to a "gender" label, preferring to use the nominative case of the pronoun directly, including (in English) SHE, HE, the non-binary use of singular THEY, and neopronouns (e.g., ZE/HIR, XEY/XEM), which have been in usage since at least the 1970s [Bustillos, 2011, Merriam-Webster, 2016, Bradley et al., 2019, Hord, 2016, Spivak, 1997].

[2]Following GLAAD [2007], transgender individuals are those whose gender differs from the sex they were assigned at birth. This is in opposition to cisgender (or non-trans) individuals, whose assigned sex at birth happens to correspond to their gender. Transgender individuals can either be binary (those whose gender falls in the "male/female" dichotomy) or non-binary (those for which the relationship is more complex).

uments by or about trans people (as well as documents by or about non-trans people), but more pointedly to minimize the harms caused by our systems by reinforcing existing unjust social hierarchies [Lambert and Packer, 2019].

Because coreference resolution is a component technology embedded in larger systems, directly implicating coreference errors in user harms is less straightforward than for user-facing technology. Nonetheless, there are several stakeholder groups who may easily face harms when coreference is used in the context of machine translation or search engine systems (discussed in detail in subsection 3.4.5). Following Bender's [2019] taxonomy of stakeholders and Barocas et al.'s [2017] taxonomy of harms, there are several obvious ways in which trans exclusionary coreference resolution systems can hypothetically cause harm:

- *Indirect: subject of query.* If a person is the subject of a web query, relevant web pages about xem may be downranked if "multiple references to the same person" is an important feature in ranking and the coreference system cannot recognize and resolve xyr pronouns. This can lead to quality of service and erasure harms.
- *Direct: by choice.* If a grammar checker uses coreference features, it may insist that an author writing hir third-person autobiography is repeatedly making errors in refering to hirself. This can lead to quality of service and stereotyping (by reinforcing the stereotype that trans identities are not "real").
- *Direct: not by choice.* If an information extraction on job applications uses a coreference system as a preprocessor, but the coreference system relies on cisnormative assumptions, then errors may disproportionately affect those who do not fit in the gender binary. This can lead to allocative harms (for hiring) as well as erasure harms.
- *Many stakeholders* If a machine translation system needs to use discourse context to gener-

ate appropriate pronouns or gendered morphological inflections in a target language, then errors can result in directly misgendering[3] subjects of the document being translated.

To address these (and other) potential harms in more detail, as well as where and how they arise, we need to (a) complicate what "gender" means and (b) uncover how harms can enter into natural language processing (NLP) systems. Toward (a), we begin with a unifying analysis (section 3.3) of how gender is socially constructed, and how social conditions in the world impose expectations around people's gender. Of particular interest is how gender is reflected in language, and how that both matches and potentially mismatches the way people experience their gender in the world. This reflection is highlighted, for instance, in folk notions such as an implicitly assumed one-to-one mapping between a gender and pronouns. Then, in order to understand social biases around gender, we find it necessary to consider the different ways in which gender can be realized linguistically, breaking down what previously have been considered "gendered words" in NLP papers into finer-grained categories of lexical, referential, grammatical, and social gender. Through this deconstruction (well-established in socio-linguistics), we can begin to interrogate what forms of gender stereotyping are prevalent in coreference resolution.

Toward (b), we ground our analysis by adapting Vaughan and Wallach's [2019] framework of how a prototypical machine learning lifecycle operates[4]. We analyze forms of gender bias in coreference resolution in six of the eight stages of the lifecycle in Figure 3.1. We conduct much of our analysis around task definition (subsection 3.4.1), bias in underlying text (subsection 3.4.2), model definition (subsection 3.4.4), and evaluation methodologies (subsection 3.4.5)

[3] According to Clements [2018], misgendering occurs when you intentionally or unintentionally refer to a person, relate to a person, or use language to describe a person that doesn't align with their affirmed gender.

[4] Vaughan and Wallach [2019] do not distinguish data collection and data annotation (and call the combined step "Dataset Construction"); for us, the separation is natural and useful for analysis. We also replace "Test Model" and "Deploy Model" with "Test System" and "Deploy System", where system is more inclusive with training data, trained model, and etc.

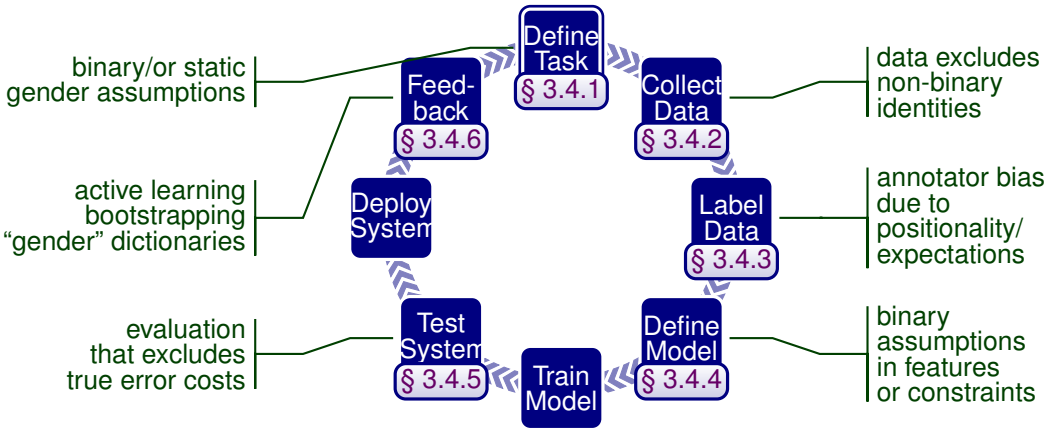

Figure 3.1: Machine learning lifecycle, with pointers to sections in this paper and some possible sources of bias listed for each state of the lifecycle; adapted from Vaughan and Wallach [2019].

by evaluating prior coreference datasets, their corresponding annotation guidelines, and through a critical read of "gender" discussions in natural language processing papers. For our analysis of bias in annotations due to annotator positionality (subsection 3.4.3), and our analysis of model definition (subsection 3.4.4), we construct two new coreference datasets: MAP (a similar dataset to GAP [Webster et al., 2018] but without binary gender constraints on which we can perform counterfactual manipulations; see subsection 3.4.2) and GICoref (a fully annotated coreference resolution dataset annotated on written by and/or about trans people; see subsubsection 3.4.4.3).[5] In all cases, we focus largely on harms due to over- and under-representation [Kay et al., 2015], replicating stereotypes [Sweeney, 2013, Caliskan et al., 2017a] (particular those that are cis-normative and/or heteronormative), and quality of service differentials Buolamwini and Gebru [2018].

The primary contributions of this chapter are:

- Analyzing gender bias in the entire coreference resolution lifecycle, with a particular focus on how coreference resolution may fail to adequately process text involving binary and

[5]Both datasets are released under a BSD license at github.com/TristaCao/into_inclusivecoref with corresponding datasheets [Gebru et al., 2018a].

non-binary trans referents (section 3.4).

- Developing an ablation technique for measuring gender bias in coreference resolution annotations, focusing on the *human* biases that can enter into annotation tasks (subsection 3.4.3).
- Constructing a new dataset, the Gender Inclusive Coreference dataset (GICOREF), for testing performance of coreference resolution systems on texts that discuss non-binary and binary transgender people (subsubsection 3.4.4.3).
- Connecting existing work on gender bias in natural language processing to sociological and sociolinguistic conceptions of gender to provide a scaffolding for future work on analyzing "gender bias in NLP" (section 3.3).

We conclude (section 3.5) with a discussion of how the natural language processing community can move forward in this task in particular, and also how this case study can be generalized to other language settings. Our goal is to highlight issues in previous instantiations of coreference resolution in order to improve tomorrow's instantiations, continuing the lifecycle of coreference resolutions' various task definition updates from MUC7 in 2001 through ACE in the mid 2000s and up to today.

3.2 Other Related Work

There are four related papers that consider gender bias in coreference resolution systems. Rudinger et al. [2018] evaluates coreference systems for evidence of *occupational stereotyping*, by constructing Winograd-esque [Levesque et al., 2012] test examples. They find that humans can reliably resolve these examples, but systems largely fail at them, typically in a gender-

stereotypical way. In contemporaneous work, Zhao et al. [2018a] proposed a very similar, also Winograd-esque scheme, also for measuring gender-based occupational stereotypes. In addition to reaching similar conclusions to Rudinger et al. [2018], this work also used a similar "counterfactual" data process as we use in subsubsection 3.4.3.1 in order to provide additional training data to a coreference resolution system. Webster et al. [2018] produced the GAP dataset for evaluating coreference systems, by specifically seeking examples where "gender" (left underspecified) could *not* be used to help coreference. They found that coreference systems struggle in these cases, also pointing to the fact that some success of current coreference systems is due to reliance on (binary) gender stereotypes. Finally, Ackerman [2019] presents an alternative breakdown of gender than we use (section 3.3), and proposes matching criteria for modeling coreference resolution linguistically, taking a trans-inclusive perspective on gender.

Gender bias in NLP has been considered more broadly than just in coreference resolution, including, for instance, natural language inference [Rudinger et al., 2017], word embeddings [e.g., Bolukbasi et al., 2016a, Romanov et al., 2019, Gonen and Goldberg, 2019], sentiment analysis [Kiritchenko and Mohammad, 2018], machine translation [Font and Costa-jussà, 2019, Prates et al., 2019], among many others [Blodgett et al., 2020, inter alia]. Gender is also an object of study in gender recognition systems [Hamidi et al., 2018]. Much of this work has focused on gender bias with a (usually implicit) binary lens, an issue which was also called out recently by Larson [2017b].

Outside of NLP, there have been many studies looking at how gender information (particularly in languages with grammatical gender) are processed by *people*, using either psycholinguistic or neurolinguistic studies. For instance, Garnham et al. [1995] and Carreiras et al. [1996] use reading speed tests for gender-ambiguous contexts, and observe faster reading when the ref-

erence was "obvious" in Spanish. Relatedly, Esaulova et al. [2014] and Reali et al. [2015] conduct eye movement studies around anaphor resolution in German, corresponding to stereotypical gender roles. In neurolinguistic studies, Osterhout and Mobley [1995] and Hagoort and Brown [1999] looked at event-related potential (ERP) violations for reflexive pronouns and antecedent in English, finding similar effects to violations of number agreement, but different effects from semantic violations. Osterhout et al. [1997] found ERP violations of the P600 type for violations of social gender stereotypes.

Issues of ambiguity in gender are also well documented in the translation studies literature, some of which have been discussed in the machine translation setting. For example, when translating from a language that can drop pronouns in subject position—the vast majority of the world's languages [Dryer, 2013]—to a language like English that (mostly) requires pronominal subjects, a system is usually forced to infer some pronoun, significantly running the risk of misgendering. Frank et al. [2004] observe that human translators may be able to use more global context to resolve gender ambiguities than a machine translation system that does not take into account discourse context. However, in some cases using more context may be insufficient, either because the context simply does not contain the answer[6], or because different languages mark for gender in different ways: e.g., Hindi verbs agree with the gender of their objects, and Russian verbal forms sometimes inflect differently depending on the gender of the speaker, the addressee, or the person being discussed [Doleschal and Schmid, 2001].

[6]For instance, the gender of the *chef de cuisine* in Daphne du Maurier's *Rebecca* is never referenced, and different human translators have selected different genders when translating that book into languages with grammatical gender [Wandruszka, 1969, Nissen, 2002].

3.3 Background: Linguistic & Social Gender

The concept of gender is complex and contested, covering (at least) aspects of a person's internal experience, how they express this to the world, how social conditions in the world impose expectations on them (including expectations around their sexuality), and how they are perceived and accepted (or not). When this complex concept is realized in language, the situation becomes even more complex: linguistic categories of gender do not even remotely map one-to-one to social categories. In order to properly discuss the role that "gender" plays in NLP systems in general (and coreference in particular), we first must work to disentangle these concepts. For without disentangling them (as few previous NLP papers have; see subsection 3.4.1), we can end up conflating concepts that are fundamentally different and, in doing so, rendering ourselves unable to recognize certain forms of bias. As observed by Bucholtz [1999]:

> *"Attempts to read linguistic structure directly for information about social gender are often misguided."*

For instance, when working in a language like English which formally marks gender on pronouns, it is all too easy to equate "recognizing the pronoun that corefers with this name" with "recognizing the real-world gender of referent of that name." Thus, possibly without even wishing to do so, we may effectively assume that "she" is equivalent to "female", "he" is equivalent to "male", and no other options are possible. This assumption can leak further, for instance by leading to an incorrect assumption that a single person cannot be referred to as both "she" and "he" (which can happen because a person's gender is contextual), nor by neither of those (which can happen when a person's gender does not align well with either of those English pronouns).

Furthermore, despite the impossibility of a perfect alignment with linguistic gender, it is generally clear that an incorrectly gendered reference to a person (whether through pronominalization or otherwise) can be highly problematic. This process of *misgendering* is problematic for both trans and cis individuals (the latter, for instance, in the all too common case of all computer science professors receiving "Dear Sir" emails), to the extent that transgender historian Stryker [2008] commented that:

> *"[o]ne's gender identity could perhaps best be described as how one feels about being referred to by a particular pronoun."*

In what follows, we first discuss how gender is analyzed sociologically (subsection 3.3.1), then how gender is reflected in language (subsection 3.3.2), and finally how these two converge or diverge (subsection 3.3.3). Only by carefully examining these two constructs, and their complicated relationship, will we be able to tease apart different forms of gender bias in NLP systems.

3.3.1 Sociological Gender

Many modern trans-inclusive models of gender recognize that *gender* encompasses many different aspects. These aspects include the experience that one has of gender (or lack thereof), the way that one expresses one's gender to the world, and the way that normative social conditions impose gender norms, typically as a dichotomy between masculine and feminine roles or traits [Kramarae and Treichler, 1985]. The latter two notions are captured by the "doing gender" model from social constructionism, which views gender as something that one *does* and to which one is socially *accountable* [West and Zimmerman, 1987, Butler, 1989, Risman, 2009]. However, viewing gender *purely* through the lens of expression and accountability does not capture

the first aspect: one's experience of one's own gender [Serano, 2007].

Such trans-inclusive views deconflate anatomical and biological traits and the sex that a person had assigned to them at birth from one's gendered position in society; this includes intersex people, whose anatomical/biological factors do not match the usual designational criteria for either sex. Trans-inclusive views further typically recognize that gender exists beyond the regressive "female"/"male" binary[7]; additionally, that one's gender may shift by time or context (often "genderfluid"), and some people do not experience gender at all (often "agender") [Kessler and McKenna, 1978, Schilt and Westbrook, 2009, Darwin, 2017, Richards et al., 2017]. These models of gender contrast with "folk" views (that are prevalent both in linguistics, sociology, and science more broadly, as well as many societies at large) which assume that ones gender is defined by one's anatomy (and/or chromosomes), that gender is binary between "male" and "female," and that one's gender is immutable. . . all of which are inconsistent with reality as it has been known for at least two thousand years.[8] In subsection 3.4.1 we will analyze the degree to which NLP papers make assumptions that are trans-inclusive or trans-exclusive.

Social gender[9] refers to the imposition of gender roles or traits based on normative social conditions [Kramarae and Treichler, 1985], which often includes imposing a dichotomy between feminine and masculine (in behavior, dress, speech, occupation, societal roles, etc.). Taking gender role as an example, upon learning that a nurse is coming to their hospital room, a patient may form expectations that this person is likely to be "female," which in turn may generate

[7]Some authors use female/male for sex and woman/man for gender; we do not need this distinction (which is itself contestable) and use female/male for gender.

[8]As identified by Keyes [2018], references appear as early as CE 189 in the Mishnah [HaNasi, 189]. Similar references (with various interpretations) also appear in the Kama Sutra [Burton, 1883, Chapter IX], which dates sometime between BCE 400 and CE 300. Archaeological and linguistic evidence also depicts the lives of trans individuals around 500 BCE in North America [Bruhns, 2006] and around 2000 BCE in Assyria [Neill, 2008].

[9] Ackerman [2019] highlights a highly overlapping concept, "bio-social gender", which consists of gender role, gender expression, and gender identity.

expectations around how their face or body may look, how they are likely to be dressed, how and where hair may appear, how to refer to them, and so on. This process, often referred to as *gendering* [Serano, 2007] occurs both in real world interactions, as well as in purely linguistic settings (e.g., reading a newspaper), in which readers may use social gender clues to assign gender(s) to the real world people being discussed. For instance, it is social gender that may cause an inference that my cousin is female in "My cousin is a librarian" or "My cousin is beautiful."

3.3.2 Linguistic Gender

Our discussion of linguistic gender largely follows [Corbett, 1991, Ochs, 1992, Craig, 1994, Corbett, 2013, Hellinger and Motschenbacher, 2015], departing from earlier characterizations that postulate a direct mapping from language to gender [Lakoff, 1975, Silverstein, 1979]. Here, it is useful to distinguish multiple ways in which gender is realized linguistically (see also [Fuertes-Olivera, 2007] for a similar overview). Our taxonomy is related but not identical to [Ackerman, 2019], which we discussed in section 3.2.

Grammatical gender, similarly defined in Ackerman [2019], is nothing more than a classification of nouns based on a principle of *grammatical agreement*. It is useful to distinguish between "gender languages" and "noun class languages." The former have two or three grammatical genders that have, for animate or personal references, considerable correspondence between a FEM (resp. MASC) grammatical gender and referents with female- (resp. male-)[10] social gender. In comparison, "noun class languages" have no such correspondence, and typically have many

[10]One difficulty in this discussion is that linguistic gender and social gender use the terms "feminine" and "masculine" differently; to avoid confusion, when referring to the linguistic properties, we use FEM and MASC.

more gender classes. Some languages have no grammatical gender at all; English is generally seen as one (viewing that referential agreement of personal pronouns does not count as a form of grammatical agreement, a view which we follow, but one that is contested [Nissen, 2002, Baron, 1971, Bjorkman, 2017]).

Referential gender (similar, but not identical to Ackerman's (2019) "conceptual gender") relates linguistic expressions to extra-linguistic reality, typically identifying referents as "female," "male," or "gender-indefinite." Fundamentally, referential gender *only* exists when there is an entity being referred to, and their gender (or sex) is realized linguistically. The most obvious examples in English are gendered third person pronouns (SHE, HE), including neopronouns (ZE, EM) and singular THEY[11], but also includes cases like "policeman" when the intended referent of this noun has social gender "male". (Note that this is *different* from the case when "policeman" is used exclusively non-referentially, as in "every policeman needs to hold others accountable", in which setting this is a case of *lexical gender*, as follows).

Lexical gender refers to extra-linguistic properties of female-ness or male-ness in a *non-referential* way, as in terms like "mother" or "uncle" as well as gendered terms of address like "Mrs." and "Sir." Importantly, lexical gender is a property of the linguistic unit, *not* a property of its referent in the real world, which may or may not exist. For instance, in "Every son loves his parents", there is no real world referent of "son" (and therefore no *referential* gender), yet it still (likely) takes HIS as a pronoun anaphor because "son" has lexical gender MASC.

We will make use of this taxonomy of linguistic gender in our ablation of annotation biases in subsection 3.4.3, but first need to discuss ways in which notions of linguistic gender match (or

[11]People's mental acceptability of singular THEY is still relatively low even with its increased usage [Prasad and Morris, 2020], and depends on context [Conrod, 2018a].

mismatch) from notions of social gender.

3.3.3 Interplays between Social and Linguistic Gender

The inter-relationship between all these types of gender is complex, and none is one-to-one. An individuals' gender identity may mismatch with their gender expression (at a given point in time). The referential gender of an individual (e.g., pronouns in the case of English) may or may not match either their gender identity or expression, and this may change by context. This can happen in the case of people whose everyday life experience of their gender fluctuates over time (at any interval), as well as in the case of drag performers (e.g., some men who perform drag are addressed as SHE while performing, and HE when not [Anonymous, 2017, Butler, 1989]).

The other linguistic forms of gender (grammatical, lexical) also need not match each other, nor match referential gender. For instance, a common example is the German term "Mädchen", meaning "girl" [e.g., Hellinger and Motschenbacher, 2015]. This term is grammatically neuter (due to the diminutive "-chen" suffix), has lexical gender as "female", and generally (but not exclusively) has female referential gender (by being used to refer to people whose gender is female). The idiom "Mädchen für alles" ("girl for everything", somewhat like "handyman") allows for male referents, sometimes with a derogatory connotation and sometimes with a connotation of appreciation.[12]

Social gender (societal expectations, in particular) captures the observation that upon hearing "My cousin is a librarian", many speakers will infer "female" for "cousin", because of either an entailment of "librarian" or some sort of probabilistic inference [Lyons, 1977], but not based on either grammatical gender (which does not exist anyway in English) or lexical gender. Such

[12] Dahl [2000] provides several several complications of this analysis.

inferences can also happen due to interplays between social gender and heteronormativity. This can happen in cases like "X's husband", in which some listeners may infer female social gender for "X", as well as in ambiguous cases like "X's spouse", in which some listeners may infer "opposite" genders for "X" and their spouse (the inference of "opposite" additionally implies a gender binary assumption).

In this chapter, we focus exclusively on English, which has no grammatical gender, but does have lexical gender (e.g., in kinship terms like "mother" and forms of address like "Mrs."). English also marks referential gender on singular third person pronouns.

English THEY, in particular, is tricky, because it can be used to refer to:

- plural non-humans (e.g., a set of boxes),
- plural humans (e.g., a group of scientists),
- a quantified human of unknown or irrelevant gender ("Every student loves their grade"),
- an indefinite human of unknown or irrelevant gender ("A student forgot their backpack"),
- a definite specific human of unknown gender, or one of non-binary gender ("Parker saw themself in the mirror").[13]

This ontology is due to Conrod [2018b], who also investigates the degree to which these are judged grammatical by native English speakers, and which we will use to quantify data bias (subsection 3.4.3).

Below, we use this more nuanced notion of different types of gender to inspect where in the machine learning lifecycle for English coreference resolution different types of bias play out. These biases may arise in the context of any of these notions of gender, and we encourage future

[13]The use of singular they to denote referents of unknown gender dates back to the late 1300s, while the non-binary use of they dates back at least to the 1950s [Merriam-Webster, 2016].

MUC7	Message Understanding Conference 7	2001T02
Zh-PB3	Chinese Propbank 3.0	2003T13
ACE04	Automatic Content Extraction 2004	2005T09
BBN	BBN Pronoun Coreference Corpus	2005T33
ACE05	Automatic Content Extraction 2005	2006T06
LUAC	Language Understanding Annotation Corpus	2009T10
NXT	NXT Switchboard Annotations	2009T26
MASC3	Manually Annotated Sub-Corpus	2013T12
Onto5	Ontonotes Version 5	2013T19
ACE07	Automatic Content Extraction 2007	2014T18
AMR2	Abstract Meaning Representation v 2	2017T10
GAP	Gendered Ambiguous Pronouns	Webster et al. [2018]
QB	QuizBowl Coreferences	Guha et al. [2015]

Table 3.1: Corpora analyzed in this paper.

work to extend care over what notions of gender are being utilized and when.

3.4 Sources of Bias

In this section, we analyze several ways in which harmful biases can and do enter into the machine learning lifecycle of coreference resolution systems (per Figure 3.1). Two stages discussed by Vaughan and Wallach [2019] that we exclude are Training Process and Deployment. It is rare (as they observed as well) for training processes (especially in batch learning settings) to lead to bias, and the same appears to be the case here. We do not consider the "Deployment" phase, because we are not aware of deployed coreference resolution systems to test; except, perhaps, those embedded in other systems, which we discuss in the context of testing (subsection 3.4.5).

				THEY			
	SHE	HE	NEO	SP	QI	PL	NH
MUC7		7			2	1	
Zh-PB3		4			3		
ACE04	2	6			1		
LUAC	1	2					
Onto5	1						
AMR2	5	17					
QB		1					
Total	9	37	0	0	6	1	0

Table 3.2: Frequency counts of different pronouns in example annotations given in annotation instructions for seven of the datasets that provide examples. Zeros are omitted from the table. *No datasets contain examples using neopronouns, nor any examples using "they" to refer to a singular specific entity (and only older datasets included any examples of quantified usages of "they").*

3.4.1 Bias in: Task Definition

Task definitions for linguistic annotations (like coreference) tend, in NLP, to be described in annotation guidelines (or, more recently, in datasheets or data statements Gebru et al. [2018a], Bender and Friedman [2018]). These guidelines naturally change over the years as the community understands more and more about both the task and the annotation process (this is part of what makes the lifecycle a *cycle*, rather than a pipeline). Getting annotation guidelines "right" is *difficult*, particularly in balancing informativeness with ability to achieve inter-annotator agreement, and *important* because poorly defined tasks lead to a substantial amount of wasted research effort.

For the purposes of this study, we consider here (and elsewhere in this paper) thirteen datasets on which coreference or anaphora are annotated in English (Table 3.1); eleven of these are corpora distributed by the Linguistic Data Consortium (LDC)[14], and two are not. According to

[14] See https://catalog.ldc.upenn.edu/{LDC-ID}.

the authors of the QB dataset (personal communication), it was annotated under the OntoNotes guidelines, with the exception that singleton mentions were also annotated. The final dataset, GAP, did not explicitly annotate full coreference, but rather annotated a binary choice of which of two names a pronoun refers to (as described in the associated paper). None of the annotation guidelines (or papers) give explicit guidance about what personal pronouns are to be considered (with the exception of GAP which explicitly limits to HE/SHE pronouns), or otherwise what information a human annotator should use to resolve ambiguous situations. However, many of them do provide running *examples*, which we can analyze. Although the examples in annotation guidelines (or papers) are insufficient to fully tell what annotations were intended by the authors, they do provide a sense of what may have been top of mind in dataset construction.

To assess task definition bias, we count, for each of the annotation guidelines, how many examples use different pronominal forms. For examples that use THEY, we separate four different subtypes, following Conrod's [2018b] categorization (see subsection 3.3.2):

NH: Plural non-human group – "The knives are put away in their carrier."

PL: Plural group of humans – "The children are friendly, and they are happy."

QI: Quantified/indefinite – "Most chefs harshly critique their own dishes."

SP: Specific singular referent – "Jun enjoys teaching their students."

The results are shown in Table 3.2 (datasets for which no relevant examples were provided are not listed). Overall, we see that in total across these seven datasets, examples with HE occur more than twice as frequently as all others combined. Furthermore, THEY is never used in a specific setting and, somewhat interestingly, is only used as an example for quantification in *older* datasets (2005 and before). Moreover, none of the annotation guidelines have examples using neopronouns. This lack does nothing to counterbalance a general societal bias that tends

to erase non-binary identities. In the case of GAP, it is explicitly mentioned that that *only* SHE and HE examples are considered (and only in cases where the gender of two possible referents "matches"—though it is unspecified what type of gender this is and how it is determined). Even on the binary spectrum, there is also an obvious gender bias between HE and SHE examples.

Tasks defined in these thirteen datasets only consider binary gender and are mostly male-dominated. Systems built along with such task definitions can hardly function for non-binary and female users. See subsection 3.4.5 for detailed analysis of system performances on data with both binary and non-binary pronouns.

3.4.2 Bias in: Data Input

In coreference resolution, as in most NLP data collection settings, one typically first collects raw text and then has human annotators label that text. Here, we consider biases that arise due to the selection of what texts to have annotated. As an example, if a data curator chooses newswire text from certain sources as source material, xey are unlikely to observe singular uses of THEY which, for instance, was only added to the Washington Post style guide in late 2015 [Walsh, 2015] and by the Associated Press Stylebook in early 2017 [Andrews, 2017]. If the raw data does not contain certain phenomena, this fundamentally limits all further stages in the machine learning lifecycle (a system which has never seen "hir" is unlikely to even know it's a pronoun, much less how to link it; indeed the off-the-shelf tokenizer we used often failed to separate "xey're" into two tokens, as it does for "they're").

To analyze the possible impact of input data, we consider our thirteen coreference datasets, and count how many instances of different types of pronouns are used in the raw data. We focus

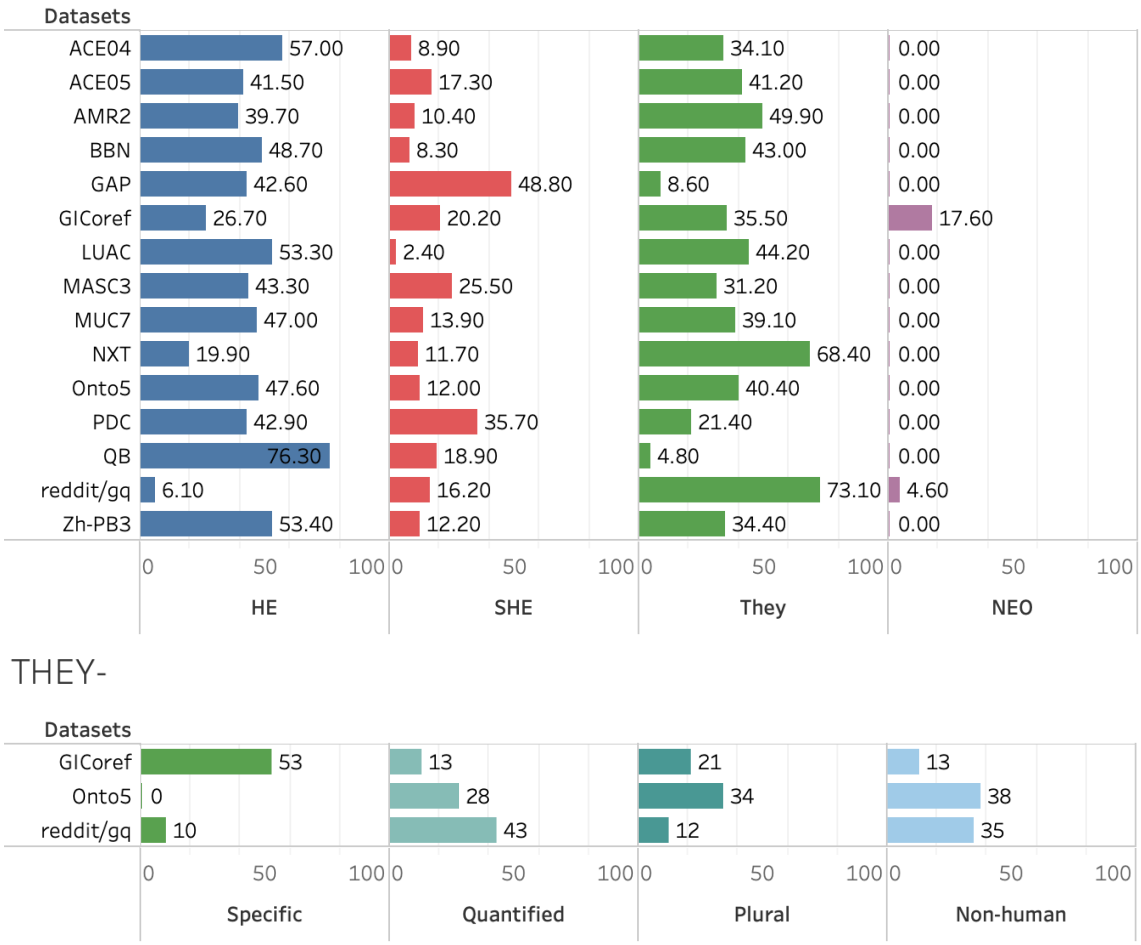

Figure 3.2: (Top) For each dataset under consideration, the fraction of pronouns with different forms. Only our dataset (GICoref) and the genderqueer subreddit include neopronouns. (In the case of GAP, there *are* some occurrences of THEY, but they are never considered targets for coreference and so we exclude them from these counts.) (Bottom) For three datasets, the fraction (out of 100 annotated) of each *they* into one of the four usage cases.

on SHE, HE, and THEY pronouns (in all their morphological forms); we additionally counted several neopronoun forms (HIR, XEY and EY) and found no occurrences nor their morphological variances.[15] In the case of THEY, we again distinguish between its four usage cases: plural, singular, quantified, and non-human. To achieve this, we annotated 100 examples uniformly at random by hand from OntoNotes Weischedel et al. [2011]. Furthermore, we compare to the raw counts in a 2015 dump from some of Reddit discussion forum[16], and also limited to the `genderqueer` subforum. We additionally include a new dataset for this study, GICoref. This new dataset is collected to evaluate current coreference resolution systems on gender-inclusive and naturally-occuring texts. Details if the dataset are described in subsubsection 3.4.4.3.

The results of this analysis are in Figure 3.2. Overall, the examples used in the documentation of each of these datasets focuses entirely on binary gendered pronouns, generally with many more HE examples than SHE examples. Only the older datasets (MUC7 and Zh-PB3) include any examples of THEY, some of which are in a quantified form.

Systems trained from these datasets never see non-binary pronouns during training. Thus when generalizing, system performance for non-binary users on singular THEY or neopronouns surely will be worse.

3.4.3 Bias in: Data Annotation

A significant possible source of bias comes from annotations themselves, arising from a combination of (possibly) underspecified annotations guidelines and the positionality of annota-

[15]There was one instance of "hir" but that was a almost certainly typo for "his" (given the context), and several instances of "ey" used a contractions for plural THEY in transcripts of spoken English.

[16]It is a dataset with publicly available comments from Reddit. The dataset has about 1.7 billion comments with their related fields, such as score, author, subreddits, etc. Here is the link to the dataset www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

tors themselves. Ackerman [2019] analyzes how human cognitively encode gender in resolving coreferences through a Broad Matching Criterion, which posits "matching gender requires at least one level of the mental representation of gender [e.g. the occupation of the person or the environment context, etc.] to be identical to the candidate antecedent in order to match." In this section, we delve into the linguistic notions of gender and study how different aspects of linguistic notions impact an annotator's judgments of anaphora.

Our study can be seen as evaluating which conceptual properties of gender are most salient in human annotation judgments. We start with natural text in which we can cast the coreference task as a binary classification problem ("which of these two names does this pronoun refer to?") inspired by Webster et al. [2018]. We then generate "counterfactual augmentations" of this dataset by ablating the various notions of linguistic gender described in subsection 3.3.2, similar to Zmigrod et al. [2019] and Zhao et al. [2018a]. We finally evaluate the impact of these ablations on human annotation behavior to answer the question: which forms of linguistic knowledge are most essential for human annotators to make consistent judgments.

As motivation, consider (1) below, in which an annotator is likely to determine that "her" refers to "Mary" and not "John" due to assumptions on likely ways that names may map to pronouns (or possibly by not considering that SHE pronouns could refer to someone named "John"). While in (2), an annotator is likely to have difficulty making a determination because both "Sue" and "Mary" suggest "her". In (3), an annotator lacking knowledge of name stereotypes on typical Chinese and Indian names (plus the fact that given names in Chinese—especially when romanized—generally do not signal gender strongly), respectively, will likewise have difficulty.

(1) John and Mary visited *her* mother.

(2) Sue and Mary visited *her* mother.

(3) Liang and Aditya visited *her* mother.

In all these cases, the plausible rough inference is that a reader takes a name, uses it to infer the sociological gender of the extra-linguistic referent. Later the reader sees the SHE pronoun, infers the referential gender of that pronoun, and checks to see if they match.

An equivalent inference happens not just for names, but also for lexical gender references (both gendered nouns (4) and terms of address (5)), grammatical gender references (in gender languages like Arabic (6)), and social gender references (7). The last of these ((7)) is the case in which the correct referent is likely to be least clear to most annotators, and also the case studied by Rudinger et al. [2018] and Zhao et al. [2018a].

(4) My brother and niece visited *her* mother.

(5) Mr. Hashimoto and Mrs. Iwu visited *her* mother.

(6) المطرب و الممثلة شاهدا والدتها

walidatu *-ha* shahadaa al-momathela w almutarab

mother *-her* saw actor[FEM] and singer[MASC]

The singer[MASC] and actor[FEM] saw *her* mother.

(7) The nurse and the actor visited *her* mother.

3.4.3.1 Ablation Methodology

In order to determine *which* cues annotators are using and the *degree* to which they use them, we construct an ablation study in which we hide various aspects of gender and evaluate how this impacts annotators' judgments of anaphoricity. To make the task easier for crowdsourcing, we follow the methodology of Webster et al.'s [2018] GAP dataset for studying ambiguous binary gendered pronouns. In particular, we construct binary classification examples taken from Wikipedia pages, in which a single pronoun is selected, and two possible antecedent names are given, and the annotator must select which one. We cannot use the GAP dataset directly, because their data is constrained that the "gender" of the two possible antecedents is "the same"[17]; for us, we are specifically interested in how annotators make decisions even when additional gender information is available. Thus, we construct a dataset called *Maybe Ambiguous Pronoun* (MAP), which is similar to the GAP dataset, but where we do not restrict the two names to match gender so that we can measure the influence of different gender cues.

In ablating gender information, one challenge is that removing social gender cues (e.g., "nurse" tending female) is not possible because they can exist anywhere. Likewise, it is not possible to remove syntactic cues in a non-circular manner. For example in (8), syntactic structure strongly suggests the antecedent of "herself" is "Liang", making it less likely that "He" corefers with Liang later (though it is possible, and such cases exist in natural data due either to gender-fluidity or misgendering).

(8) Liang saw herself in the mirror. . . *He* . . .

Fortunately, it *is* possible to enumerate a high coverage list of English terms that signal lexical

[17] It is unclear from the GAP dataset what notion of "gender" is used, nor how it was determined to be "the same."

~~Mrs.~~ $\xrightarrow{(d)}$ ∅ ~~Rebekah Johnson Bobbitt~~ $\xrightarrow{(b)}$ **M. Booth** was the younger ~~sister~~ $\xrightarrow{(c)}$ **sibling** of ~~Lyndon B. Johnson~~ $\xrightarrow{(b)}$ **T. Schneider**, 36th President of the United States. Born in 1910 in Stonewall, Texas, ~~she~~ $\xrightarrow{(a)}$ **they** worked in the cataloging department of the Library of Congress in the 1930s before ~~her~~ $\xrightarrow{(a)}$ **their** ~~brother~~ $\xrightarrow{(c)}$ **sibling** entered politics.

Figure 3.3: Example of applying *all* ablation substitutions for an example context in the MAP corpus. Each substitution type is marked over the arrow and separately color-coded.

gender: terms of address (Mrs., Mr.) and semantically gendered nouns (mother).[18] We assembled a list by taking many online lists (mostly targeted at English language learners), merging them, and manual filtering. The assembling process and the final list is published with the MAP dataset and its datasheet.

To execute the "hiding" of various aspects of gender, we use the following substitutions:

(a) ¬PRO: Replace all third person singular pronouns with a gender neutral program (THEY, XEY, ZE).

(b) ¬NAME: Replace all names (e.g., "Aditya Modi") by a random name with only a first initial and last name (e.g., "B. Hernandez").

(c) ¬SEM: Replace all semantically gendered nouns (e.g., "mother") with a gender-indefinite variant (e.g., "parent").

(d) ¬ADDR: Remove all terms of address (e.g., "Mrs.", "Sir")[19].

See Figure 3.3 for an example of all substitutions.

We perform two sets of experiments, one following a "forward selection" type ablation (start with everything removed and add each back in one-at-a-time) and one following "backward selection" (remove each separately). Forward selection is necessary in order to de-conflate

[18] These are, however, sometimes complex. For instance, "actress" signals *lexical* gender of female, while "actor" may signal *social* gender of male and, in certain varieties of English, may also signal *lexical* gender of male.

[19] An alternative suggested by Cassidy Henry that we did not explore would be to replace all with Mx. or Dr.

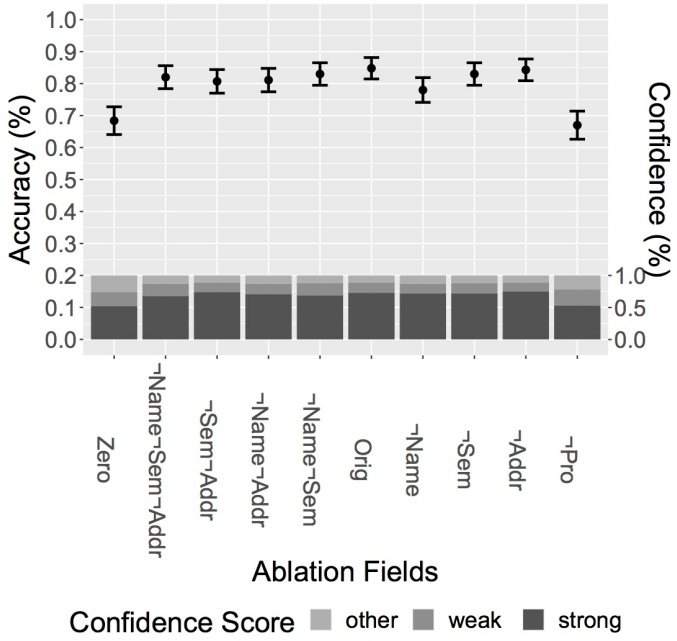

Figure 3.4: Human annotation results for the ablation study on MAP dataset. Each column is a different ablation, and the y-axis is the degree of *accuracy* with 95% significance intervals. Bottom bar plots are annotator certainties as how sure they are in their choices.

syntactic cues from stereotypes; while backward selection gives a sense of how much impact each type of gender cue has in the context of all the others.

We begin with ZERO, in which we apply all four substitutions. Since this also removes gender cues from the pronouns themselves, an annotator cannot substantially rely on social gender to perform these resolutions. We next consider adding back in the original pronouns (always HE or SHE here), yielding ¬NAME ¬SEM ¬ADDR. Any difference in annotation behavior between ZERO and ¬NAME ¬SEM ¬ADDR can only be due to social gender stereotypes.

To see why, consider the example from Figure 3.3. In this case, the only difference between the Zero setting and the ¬NAME ¬SEM ¬ADDR is whether the pronouns SHE/HER are substituted with THEY/THEIR—all other substitutions are applied in both cases. In the ZERO case, there are no gender cues at all to help with the resolution, precisely because gender has been removed even from the pronoun. So even if there were gendered information in the rest of the text, that logically

cannot help with the resolution of either of the pronouns. In the ¬NAME ¬SEM ¬ADDR case, all lexical and referential gender information *except* that on the pronoun have been removed (as English has no grammatical gender). If one accepts that the taxonomy of gender from subsection 3.3.2 is complete, then this means that the only gender information that exists in the rest of the this example is social gender. (And indeed, there is social gender—even in a fictitious world in which someone named T. Schneider was the 36th President of the U.S., social gender roles suggest that this person is relatively unlikely to be the referent of *she*). Thus, it is likely that in this latter case, readers and annotators can and will use the gender information on the pronoun to decide that SHE does not refer to Schneider and therefore likely refers to Booth. On the other hand, in the ZERO case, there is no such information available, and a reader must rely, perhaps, on parallel syntactic structure or centering [Joshi and Weinstein, 1981, Grosz et al., 1983, Poesio et al., 2004] if ey are to correctly identify that the referent is Booth.

The next setting, ¬SEM ¬ADDR removes both forms of lexical gender (semantically gendered nouns and terms of address); differences between ¬SEM ¬ADDR and ¬NAME ¬SEM ¬ADDR show how much names are relied on for annotation. Similarly, ¬NAME ¬ADDR removes names and terms of address, showing the impact of semantically gendered nouns, and ¬NAME ¬SEM removes names and semantically gendered nouns, showing the impact of terms of address.

In the backward selection case, we begin with ORIG, which is the unmodified original text. To this, we can apply the pronoun filter to get ¬PRO; differences in annotation between ORIG and ¬PRO give a measure of how much *any* sort of gender-based inference is used. Similarly, we get ¬NAME by only removing names, which gives a measure of how much names are used (in the context of all other cues); we get ¬SEM by only removing semantically gendered words; and

¬ADDR by only removing terms of address.

3.4.3.2 Annotation Results

We construct examples using the methodology defined above. We then conduct annotation experiments using crowdworkers on Amazon Mechanical Turk following the methodology by which the original GAP corpus was created[20]. Because we wanted to also capture uncertainty, we ask the crowdworkers how sure they are in their choices, between "definitely" sure, "probably" sure and "unsure."[21]

Figure 3.4 shows the human annotation results as binary classification accuracy for resolving the pronoun to the antecedent. We can see that removing pronouns leads to significant drop in accuracy. This indicates that gender-based inferences, especially social gender stereotypes, play the most significant role when annotators resolve coreferences. This confirms the findings of Rudinger et al. [2018] and Zhao et al. [2018a] that human annotated data incorporates bias from stereotypes.

Moreover, if we compare ORIG with columns left to it, we see that name is another significant cue for annotator judgments, while lexical gender cues do not have significant impacts on human annotation accuracies. This is likely in part due to the low appearance frequency of lexical gender cues in our dataset. Every example has pronouns and names, whereas 49% of

[20]This study was approved by the Microsoft Research Ethics Board (#427). We recruit workers from countries with large native English speaking populations (Australia, Canada, New Zealand, United Kingdom, and United States), and who have greater than 80% HIT approval rate and more than 100 HITs approved. Workers were paid $1 to annotate ten contexts (the average annotation time was seven minutes). Crowdworkers were informed as part of the instructions and examples that they should expect to see both singular THEY and neopronouns, with examples of each.

[21]In some of the examples, a crowdworker may apply knowledge of the situation or entities involved, for instance "President of the United States" has, to date, always been referred to using HE pronouns. To capture this, we additionally asked the crowdworkers if they recognized any of the entities or the situation involved. Note, however, that even though this is true in the real world, it is not hard to imagine fictional contexts in which a human annotator would have no difficulty finding "President of the United States" to corefer with SHE or THEY pronouns.

the examples have semantically gendered nouns but only 3% of the examples include terms of address. We also note that if we compare ¬NAME ¬SEM ¬ADDR to ¬SEM ¬ADDR and ¬NAME ¬ADDR, accuracy drops when removing gender cues. Though the differences are not statistically significant, we did not expect the accuracy drop.

Finally, we find annotators' confidence follow the same trend as the accuracy: annotators have a reasonable sense of when they are unsure. We also note that accuracy score are essentially the same for ZERO and ¬PRO, which suggests that once explicit binary gender is gone from pronouns, the impact of any other form of linguistic gender in annotator's decisions is also removed.

Overall, we can see that annotators may make unlicenced inferences of various gender cues by conflating gender concepts. Thus systems trained by treating these annotator judgments as ground truth can be problematic for both binary and non-binary people.

3.4.3.3 Limitations of (approximate) counterfactual text manipulation.

Any text manipulation—like we have done in this section—runs the risk of missing out on how a human author might truly have written that text under the presumed counterfactual. For example, a speaker uttering (1) may assume that aer interlocutor shares, or at least recognizes, social biases that lead one to assume that the person named "John" is likely referred to as HE and "Mary" as SHE. This speaker may use this assumption of the listener to determine that "her" is sufficiently unambiguous in this case as to be an acceptable reference (trading off brevity and specificity; see, for instance [Arnold, 2008, Frank and Goodman, 2012, Orita et al., 2015a]). However, if we "counterfactually" replaced the names "John" and "Mary" to "H. Martinez" and

"R. Modi" (respectively), it is unlikely that the supposed speaker would make the same decision. In this case, the speaker may well have said "Modi's mother" or some other reference that would have been sufficiently specific to resolve, even at the cost of being more wordy. That is to say, the counterfactual replacements here and their effect on human annotation agreement should be taken as a sort of *upper bound* on the effect one would expect in a truly counterfactual setting.

Moreover, although we studied crowdworkers on Mechanical Turk (because they are often employed as annotators for NLP resources), if other populations are used for annotation, it becomes important to consider their positionality and how that may impact annotations. This echoes a related finding in annotation of hate-speech that annotator positionality matters [Olteanu et al., 2019].

3.4.4 Bias in: Model Definition

Bias in machine learning systems can also come from how models are structured, for instance what features they use, and what baked-in decisions are made. For instance, some models may simply fail to recognize anything other than a dictionary of fixed pronouns as possible entities. Others may use external resources, such as lists that map names to guesses of "gender," that bake in stereotypes around naming.

In this section, we analyze prior work in systems for coreference resolution in three ways. First, we do a literature study to quantify how NLP papers discuss gender, broadly. Second, similar to Zhao et al. [2018a] and Rudinger et al. [2018], we evaluate a handful of freely available systems on the ablated data from subsection 3.4.3. Third, we evaluate these systems on the dataset we created: Gender Inclusive Coreference (GICOREF).

	All Papers		Coref Papers	
Paper Discusses Linguistic Gender?	52.6%	(79/150)	95.4%	(21/22)
Paper Discusses Social Gender?	58.0%	(87/150)	86.3%	(19/22)
Paper Distinguishes Linguistic from Social Gender?	11.1%	(3/27)	5.5%	(1/18)
Paper assumes Social Gender is Binary?	92.8%	(78/84)	94.4%	(17/18)
Paper Assumes Social Gender is Immutable?	94.5%	(70/74)	100.0%	(14/14)
Paper Allows for Neopronouns/THEY-SP?	3.5%	(2/56)	7.1%	(1/14)

Table 3.3: Analysis of a corpus of 150 NLP papers that mention "gender" along the lines of what assumptions around gender are implicitly or explicitly made in the work.

3.4.4.1 Cis-normativity in published NLP papers

In our first study, we adapt the approach Keyes [2018] took for analyzing the degree to which computer vision papers encoded trans-exclusive models of gender. In particular, we began with a random sample of 150 papers from the ACL anthology that mention the word "gender" and coded them according to the following questions:

- Does the paper discuss coreference or anaphora resolution?
- Does the paper study English (and possibly other languages)?
- Does the paper deal with linguistic gender (i.e., grammatical gender or gendered pronouns)?
- Does the paper deal with social gender?
- (If yes to the previous two:) Does the paper explicitly distinguish linguistic from social gender?
- (If yes to social gender:) Does the paper explicitly recognize that social gender is not binary?

- (If yes to social gender:) Does the paper explicitly or implicitly assume social gender is immutable?[22]
- (If yes to social gender and to English:) Does the paper explicitly consider uses of definite singular "they" or neopronouns?

The results of this coding are in Table 3.3 and the list of the full set of annotations is in section A.1. Here, we see out of the 22 coreference papers analyzed, the vast majority conform to a "folk" theory of language:

◇ Only 5.5% distinguish social from linguistic gender (despite it being relevant);

◇ Only 5.6% explicitly model gender as inclusive of non-binary identities;

◇ No papers treat gender as anything other than completely immutable;

◇ Only 7.1% (one paper) considers neopronouns and/or specific singular THEY.

The situation for papers not specifically about coreference is similar (the majority of these papers are either purely linguistic papers about grammatical gender in languages other than English, or papers that do "gender recognition" of authors based on their writing; May [2019] discusses the (re)production of gender in automated gender recognition in NLP in much more detail). Overall, the situation more broadly is equally troubling, and generally also fails to escape from the folk theory of gender. In particular, none of the differences between papers and papers about coreference are significant at a $p < 0.05$ level except for the first two questions, due to the small sample size (according to an $n-1$ chi-squared test). The result of this analysis is that although we do not knew exactly what decisions are baked in to all models, the vast majority in our study (including two papers by one of the authors Daumé and Marcu [2005], Orita et al. [2015b]) come with

[22]The most common ways in which papers implicitly assume that social gender is immutable is either 1) by relying on external knowledge bases that map names to "gender"; or 2) by scraping a history of a user's social media posts or emails and assuming that their "gender" today matches the gender of that historical record.

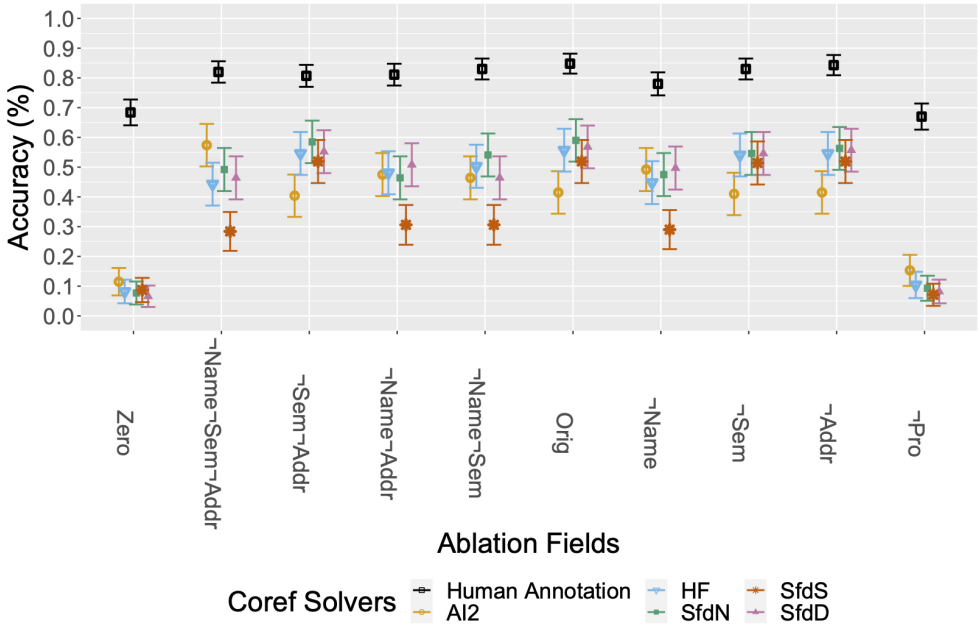

Figure 3.5: Coreference resolution systems results for the ablation study on MAP dataset. The y-axis is the degree of *accuracy* with 95% significance intervals.

strong gender binary assumptions, and exist within a broader sphere of literature which erases non-binary identities.

3.4.4.2 Coreference system performance on MAP

Next, we analyze the effect that our different ablation mechanisms have on existing coreference resolutions systems. In particular, we run five coreference resolution systems on our ablated data: the AI2 system [AI2; Gardner et al., 2017], hugging face [HF; Wolf, 2017], which is a neural system based on spacy [Honnibal and Montani, 2017], and the Stanford deterministic [SfdD; Raghunathan et al., 2010], statistical [SfdS; Clark and Manning, 2015] and neural [SfdN; Clark and Manning, 2016] systems. Figure 3.5 shows the results. We can see that the system accuracies mostly follow the same pattern as human accuracy scores, though all are significantly lower than human results. Accuracy scores for systems drop dramatically when we ablate out ref-

erential gender in pronouns. This reveals that those coreference resolution systems rely heavily on gender-based inferences. In terms of each system, HF and SfdN systems have similar results and outperform other systems in most cases. SfdD accuracy drops significantly once names are ablated.

These results echo and extend previous observations made by Zhao et al. [2018a], who focus on detecting stereotypes within occupations. They detect gender bias by checking if the system accuracies are the same for cases that can be resolved by syntactic cues and cases that cannot, with original data and reversed-gender data. Similarly, Rudinger et al. [2018] focus on detecting stereotypes within occupations as well. They construct dataset without any gender cues other than stereotypes, and check how systems perform with different pronouns – THEY, SHE, HE. Ideally, they should all perform the same because there is not any gender cues in the sentence. However, they find that systems do not work on "they" and perform better on "he" than "she". Our analysis breaks this stereotyping down further to detect which aspects of gender signals are most leveraged by current systems.

3.4.4.3 Coreference system performance on GICoref

We introduce a new dataset, GICOREF. for the purpose to evaluate current coreference resolution systems in the contexts where a broader range of gender identities are reflected, where linguistic examples of genderfluidity are encountered, where non-binary pronouns are used, and where misgendering happens. In comparison to [Zhao et al., 2018a] and Rudinger et al. [2018] (as well as in contrast to our MAP dataset), we focused on *naturally* occurring data, but sampled specifically to surface more gender-related phenomena than may be found in, say, the Wall Street

Journal.

The GICOREF dataset consists of 95 documents from three types of sources: articles on English Wikipedia about people with non-binary gender identities, articles from LGBTQ periodicals, and fan-fiction stories from Archive Of Our Own[23] (with the respective author's permission). Each author of this paper manually annotated each of these documents and then we jointly adjudicated the results[24]. To reduce annotation time, any article that was substantially longer than 1000 words (pre-tokenization) was trimmed at the 1000th word.[25] This data includes many examples of people who use pronouns other than SHE or HE, people who are genderfluid and whose name or pronouns changes through the article, people who are misgendered, and people in relationships that are not heteronormative. One annotation decision we made was around the specific case of people who perform drag. Following Butler [1989], in our annotation we considered drag performance as a form of genderfluidity; as such, we annotate the performer and the drag persona as coreferent with each other (as well as the relevant pronouns), akin to how we believed a reasonable model for handling stage names (e.g., Christopher Wallace / Notorious B.I.G.) would mark them as coreferent, while famous roles played by multiple actors (e.g., Carrie, played by Sissy Spacek, Angela Bettis, and Chloë Grace-Moretz), would be marked as non-coreferent. In addition, *incorrect references* (misgendering and deadnaming[26]) are explicitly annotated.[27] Two example annotated documents, one from Wikipedia, and one from Archive of

[23] See https://archiveofourown.org; thanks to Os Keyes for this suggestion.

[24] We evaluate inter-annotator agreement by treating one annotation as gold standard and the other as system output and computing the LEA metric; the resulting F1-score is 92%. During the adjudication process we found that most of the disagreement are due to one of the authors missing/overlooking mentions, and rarely due to true "disagreement."

[25] There are four documents we accidentally trimmed at the 2000th word and so we keep the longer version of them with 2000 words in the dataset.

[26] According to Clements [2017] deadnaming occurs when someone, intentionally or not, refers to a person who's transgender by the name they used before they transitioned.

[27] Thanks to an anonymous reader of a draft version of this paper for this suggestion.

Our Own, are provided in section A.2 and section A.3 respectively.

Although the majority of the examples in the dataset are set in a Western context, we endeavored to have a broader range of experiences represented. We included articles about people who are gender non-conforming, but where sociological notions of gender mismatch the general sex/gender/sexuality taxonomy of the West. This includes people who identify as *hijra* (Indian subcontinent), *phuying* (Thailand, sometimes referred to as *kathoey*), *muxe* (Oaxaca), *two-spirit* (Americas), *fa'afafine* (Samoa) and *māhū* (Hawaii) individuals.

We run the same systems as before on this dataset. Table 3.4 reports results according to the standard coreference resolution evaluation metric LEA Moosavi and Strube [2016]. It is not clear how systems or evaluation metrics should handle incorrect references (misgendering and deadnaming). Taking (9) as an example, should the misgendering entities and pronouns (cluster c) be included as a coreference to the person (cluster a) or not. If the person is a real human, including the misgendering reference as a ground truth may be potentially harmful to the person. Since no systems are implemented to explicitly mark incorrect references, and no current evaluation metrics address this case, we perform the same evaluation twice. One with incorrect references included as regular references in the ground truth (cluster a and cluster c are the same cluster); and other with incorrect references excluded (cluster a and cluster c are separate clusters). Due to the limited number of incorrect references (0.6% of total references of people) in the dataset, the difference of the results are not significant – the difference is less than 0.2% for each entry. Nonetheless, although these are rare, they constitute a significant potential harms. Here we only report the result for latter.

(9) Frisk$_A$ sat in the back of the classroom, silently praying that their$_A$ teacher wouldn't call on them$_A$. They$_A$ were having a bad day and didn't think they$_A$ could be misgendered today. But just their$_A$

LEA	Precision	Recall	F1
AI2	40.4%	**29.2%**	33.9%
HF	**68.8%**	22.3%	33.6%
SfdD	50.8%	23.9%	32.5%
SfdS	59.8%	24.1%	**34.3%**
SfdN	59.4%	24.0%	34.2%

Recall	HE/SHE	THEY	NEO
AI2	96.4%	93.0%	25.4%
HF	95.4%	85.0%	21.1%
SfdD	97.6%	96.6%	0.0%
SfdS	96.2%	86.8%	20.4%
SfdN	96.6%	90.1%	0.0%

Table 3.4: (Left) LEA scores on GICOREF dataset with various coreference resolution systems. Rows are different systems while columns are precision, recall, and F1 scores. When evaluate, we only count exact matches or pronouns and name entities. (Right) Recall scores of coreference resolution systems for detecting binary pronouns, THEY (of any type), and neopronouns.

> luck, their$_A$ teacher$_B$ was staring straight at them$_A$. "Felix$_C$? Do you know the answer?" ... Chara$_D$'s hand shot up. "Ms. Richards$_B$, their$_A$ name is Frisk$_A$, remember?" "Christine$_E$," Ms. Richards$_B$ sighed, ignoring Chara$_D$'s flinch. "His$_C$ name is whatever is on the sheet, the same way yours is. We$_F$ have had this discussion, remember?"

The first observation is that there is still plenty room for coreference systems to improve; the best performing system achieves as F1 score of 34%, but the Stanford neural system's F1 score on CoNLL-2012 test set reaches 60% [Moosavi, 2020]. Here are some examples where the huggingface and the Stanford deterministic system output erroneous resolutions: (10), (11), and (12). As demonstrated, even when there are clear syntactic cues and declaration of preferred pronouns, both systems fail to resolve the coreferences correctly due to various internal biases.

(10) **HF:** The artwork$_B$ consisted of Sulkowics$_A$, who uses they$_B$/them$_B$ pronouns, carrying a mattress wherever they$_B$ went on campus.

SfdD: The artwork consisted of Sulkowics$_A$, who uses they/them pronouns$_B$, carrying a mattress wherever they$_C$ went on campus.

(11) **HF & SfdD:** As the son of a military father$_{CB}$, she$_A$ faced many challenges to be accepted.

(12) **HF:** Soon after Vasold$_A$'s win was announced, ze$_B$ spoke with the school newspaper about zir surprise at winning.

SfdD: Soon after Vasold$_A$'s win was announced, ze spoke with the school newspaper about zir surprise$_B$ at winning.

Additionally, we can see system precision dominates recall. This is likely partially due to poor recall of pronouns other than HE and SHE. To analyze this, we compute the *recall* of each system for finding referential pronouns at all, regardless of whether they are correctly linked to their antecedents. Results are shown in Table 3.4. We find that all systems achieve a recall of at least 95% for binary pronouns, a recall of around 90% on average for THEY, and a recall of around a paltry 13% for neopronouns (two systems—Stanford deterministic and Stanford neural—never identify any neopronouns at all).

Overall, we have shown current coreference resolution systems fail to escape from the folk theory of gender and rely heavily on gender-based inferences. Therefore when deployed, these systems can easily make biased inferences that will lead to both direct and indirect harms to binary and non-binary users.

3.4.5 Bias in: System Testing

Bias can also show up at testing time, due either to data or metrics. For instance, if one evaluates on highly biased data, it will be difficult to capture disparities (akin to the over-representation of light skinned men in computer vision datasets [Buolamwini and Gebru, 2018]). Alternatively, evaluation metrics may weight different errors in a way that is incongruous with their harm. For example, depending on the use case, corefering someone's name with an incorrectly gendered pronoun may produce a harm askin to misgendering, potentially leading to a high cost social error [Stryker, 2008]; evaluation metrics may or may not reflect the true cost of such mistakes.

In terms of data, most coreference resolution systems are evaluated intrinsically, by testing them against gold standard annotations using a variety of metrics; in this case, all the observations on data bias (subsection 3.4.2 and subsection 3.4.3) apply. Sometimes coreference resolution is used as part of a larger system. For instance, information retrieval can use coreference resolution to help accurately rank documents by up-weighting the importance of entities that are referred to frequently [Du and Liddy, 1990, Pirkola and Järvelin, 1996, Edens et al., 2003]. In machine translation, producing correct gendered forms in gender languages often requires coreference to have been solved [Mitkov, 1999, Hardmeier and Federico, 2010, Guillou, 2012, Hardmeier and Guillou, 2018]. In the translation case, this then raises the question: which data are being used and how biased are they? It turns out, "quite biased." Even limiting to just SHE and HE pronouns, the bias is significant: four times as many HE than SHE in Europarl [Koehn, 2005] and the Common Crawl [Smith et al., 2013], six times as many in News Commentaries [Tiedemann, 2012], and fifty times as many in Hong Kong Laws corpus.

In terms of metrics, most intrinsic evaluation is carried out using metrics like MUC [Vilain et al., 1995], ACE [Mitchell et al., 2005], B^3 [Bagga and Baldwin, 1998, Stoyanov et al., 2009], or CEAF [Luo, 2005] (see also [Cai and Strube, 2010] for additional discussion and variants). As observed by Agarwal et al. [2019], most of these metrics are rather insensitive to arguably large errors, like the inability to link pronouns to names; to address this, they introduce a new metric to focus specifically on this named entity coreference task. These metrics also generally treat all errors similarly, regardless of whether the error compounds societal injustices (e.g., ignoring an instance of XYR) or not (e.g., ignore an instance of HE), despite the fact that these have vastly different implications from the perspective of justice [e.g., Fraser, 2008].

For extrinsic evaluation, the metrics used are those that are appropriate for the downstream

task (e.g., machine translation). In the case of machine translation, one can ask whether standard evaluation metrics like Bleu [Papineni et al., 2002] are sensitive to misgendering. To quantify this, we use the sacreBLEU toolkit [Post, 2018] and compute Bleu scores between the ground truth reference outputs and those same references where all SHE pronouns were replaced with morphologically equivalent HE forms (none of these datasets contain neopronouns and analysis of a small sample did not find any singular specific uses of THEY). From wmt08–wmt18 and iwslt17 test sets, the average percentage *drop* in Bleu score from this error is 0.67% (± 0.22), which is barely statistically significant according to a bootstrap test at sensitivity 0.05. Evaluated only on the $\approx 17\%$ of the sentences in these datasets containing either HE or SHE, the degradation is about 3.1%. While this degradation is noticeable, it perhaps does not reflect the real cost of such translation errors due to the high, and asymmetric, societal cost of misgendering.

3.4.6 Bias in: Feedback Loops

The final sources of bias we consider are feedback loops: essentially, when the bias from a coreference system feeds back on itself, or onto other coreference systems.

The most straightforward way in which this can happen is through coreference resolution systems that engage in *statistically biased* active learning (or bootstrapping) techniques.[28] Active learning for coreference has been popular since the early 2000s, perhaps largely because coreference annotation is quite costly. Considering the approaches used in dominant papers, the active learning algorithms used are not statistically unbiased [Ng and Cardie, 2003, Laws et al., 2012, Miller et al., 2012, Sachan et al., 2015, Guha et al., 2015].

[28] Here, the overloading of "bias" is unfortunate. By "statistically biased," we mean bias in the technical sense that of a learning system that even in the limit of infinite data does not infer the optimal model, a fundamentally different concept from the sort of "bias" we consider in the rest of this paper.

Another example is the use of external dictionaries that encode world knowledge that is potentially useful to coreference resolution systems. The earliest example we know of that uses such knowledge sources is the end-to-end machine learning approach of Daumé and Marcu [2005], which found substantial benefit by using mined mappings between names and professions to help resolve named entities like "Bill Clinton" to nominals like "president" (later examples include that of Rahman and Ng [2011] and Bansal and Klein [2012], who found less benefit from a similar approach).

More frequent is the almost ubiquitous use of "name lists" that map names (either full names or simply given names) to "gender." And the most frequently used of these is the resource developed by Bergsma and Lin [2006] (henceforth, B+L), in which a large quantity of text was processed with "high precision" anaphora resolution links to associate names with "genders." The process specifically mapped names to *pronouns*, from which gender (presumably an approximation of referential gender) was inferred. This leads to a resource that pairs a full name or name substring (like "Bill Clinton") counts for identified coreference with HE (8150, 97.7%), SHE (70, 0.8%), IT (42, 0.5%) and THEY (82, 1%); these are referred to, respectively, as "male", "female", "neuter" and "plural," and seemingly largely used as such in work that leverages this resource. We focus on this resource only because it has become ubiquitous, both in coreference resolution and in gender analysis in NLP more broadly.

The first question we ask is: what happens when this "gender" inference data is used to infer the gender of prominent non-binary individuals. To this end, we took the names of 104 non-binary people referenced on Wikipedia[29] and queried the B+L data with them. In almost all

[29] https://en.wikipedia.org/wiki/List_of_people_with_non-binary_gender_identities, accessed Jan 5, 2019.

	"Gender" % Likelihood					
Pronoun	**Masc**	**Fem**	**Neut**	**Plur**	**Unk**	**count**
she	27	41	18	0	14	22
he	64	7	7	0	22	14
they	39	32	12	5	12	41
ze	22	33	33	11	0	9
they∨she	25	50	25	0	0	4
s/he	0	0	50	0	50	2
they∧she	50	0	0	0	50	2
all/any	50	50	0	0	0	2
xe	50	0	0	0	50	2
ey	50	0	0	0	50	2
v	100	0	0	0	0	1
ae	0	100	0	0	0	1
ne	100	0	0	0	0	1
ze∧she	0	0	100	0	0	1

Table 3.5: For non-binary individuals in our Wikipedia sample (for whom Wikipedia attests current pronouns), a confusion matrix between the pronoun(s) they use (rows) and the inferred gender of their name (based on [Bergsma and Lin, 2006]) in columns (where Masc="he", Fem="she", Neut="it" and Plur="they"; "Unk" means the name was not found). The final column is the total count. The semantics of "they∨she" is that the person accepts both "they" and "she" pronouns, while "they∧she" indicates that the person uses "they" or "she" depending on context (for instance, "she" while performing drag and "they" otherwise).

cases, the full name was unknown in the B+L data (or had counts less than five), and in such cases we backed off to simply querying on the given name. We cross-tabulated the correct (according to Wikipedia) pronouns for these 90 people with the "gender" inferred by the B+L data.

The results are shown in Table 3.5, where we can see that of those who use a pronoun other than SHE or HE (exclusively) are, essentially always, misgendered. Even on binary pronouns SHE and HE, the accuracy is only 50%. For the case of people who use THEY pronouns, one might ask what the ideal behavior would be given the framing of this resource—given that "Plural" is interpreted to be "coreferent with 'they'", we might hope that (aside from the naming issue), people who use THEY are considered Plural: this only happens in 5% of cases, though. Expanding on this, one might hope that people who use THEY $\vee$SHE or THEY $\wedge$SHE are mapped to either Fem or Plural, but this only happens in $2/6$ cases (and always to Fem in those cases). For the neopronoun cases, the manner in which the resource is constructed nearly excludes any reasonable behavior (aside from, perhaps, a "least bad" option of simply abstaining with an output—which only happens in $3/19$ cases). This approach actively misgenders individuals, is harmful, and demonstrates that assigning gender to "names" does not work: anybody can have any combination of names and pronouns.

3.5 Discussion and Limitations

We found varying amounts of bias entering in task definitions (including, in particular, strong assumptions around binary and immutable gender), data collection and annotation (in particular how sources of data impact that sorts of linguistic gender phenomena observed), testing and feedback. In order to do so, we made substantial use of sociological and sociolinguistic

notions of gender, in order to separate out different types of bias.

To run many of these studies, we additionally created—and released—two datasets for studying gender inclusion in coreference resolution. The MAP dataset we created counterfactually (and therefore it is subject to general concerns about counterfactual data construction), which allowed us to very precisely control how different types gender information. The GICOREF dataset we created by targetting specific linguistic phenomena (searching for uses of neopronouns in LGBTQ periodicals) or social aspects (Wikipedia articles and fan fiction about people with non-binary gender). Both datasets show significant gaps in system performance, but perhaps moreso, show that taking crowdworker judgments as "gold standard" can be problematic, especially when the annotators are judging referents of singular THEY or neopronouns. It may be the case that to truly build gender inclusive datasets and systems, we need to hire or consult experiential experts [Patton et al., 2019, Young et al., 2019].

Moreover, we realized that both human and coreference systems rely heavily on gender cues in resolving coreferences. Though it is natural for human, we want to emphasize that both human and systems should not overrely on the risky cues such as names, semantically gendered nouns, and terms of address, compared to relatively safe cues like syntax. In annotating the GICOREF dataset, we only had about three ambiguous coreferences where both annotators agreed either reference was possible, thus demonstrating that people are able to resolve coreferences without relying extensively on the riskier cues. One cue that we explored in detail is that around names, and it is worth pointing out recent work by Agarwal et al. [2020] in the context of named entity recognition. In this paper, the authors found that state of the art systems perform poorly on documents from non-U.S. contexts, due in large part to systems' unfamiliarity with non-Western names. We expect similar results would hold in the coreference case, where it would be

particularly interesting to evaluation in the context of name-gender lists.

When building a coreference system, a developer must make decisions about what features to include or exclude, and therefore what grammatical or social notions of gender are incorporated. Our view is *not* that "risky" features must be excluded in order to build an inclusive system, but rather that developers should be aware of the risks when such features are included. After all, in a speaker-listener model of language understanding [Frank and Goodman, 2012, Bard and Aylett, 2004], it is rational for a human speaker to assume that *outside of additional context*, a listener will resolve "his" to "Tom" in "Tom and Mary went to his house." However, human speakers know how to adjust the context when default expectations cannot be used, as in examples (10), (11), and (12) in subsubsection 3.4.4.3. Recall that there and example (13) here, we found that even given very explicit cues systems are unable to override their internal biases. If the goal is to understand human communication, have a system that can understand speaker intent is highly important.

(13) **HF & SfdD:** Tom$_{A}$ and Mary$_{B}$$_{C}$ are at home. Tom$_{A}$ regards herself$_{B}$ in the mirror.

This analysis potentially changes if such a model is "flipped" and used, e.g., as a method for performing referring expression generation [Krahmer and van Deemter, 2012]. Depending on a developer's normative stance, ze may need to make a decision about whether hir system will conform to, or challenge, hegemonic language usage, particularly around gender binaries, even though that may produce text that reads as unusual to some (or many) readers. For example, along masculine-as-default lines, does a system generate "engineer" or "man engineer" (when the referent is known to be male), and along non-trans-as-default lines, does a system generate "he/him" or "she/her" in the previous sentence, or "ze/hir"? What a system "should" do in such cases is highly contextual, and perhaps varies even depending on the population expected to use

the system. What does not change is that these questions should be addressed head on, so that explicit decisions can be made and consequences understood, rather than being surprised later.

More broadly than in coreference resolution, we found that natural language processing papers also tend to make strong, binary assumptions around *gender* (typically implicitly), a practice that we hope to see change in the future. In more recent papers, we begin to see footnotes that acknowledge that the discussion omits questions around trans or non-binary, issues. We hope to see these be promoted from footnotes to objects of study in future work; mentioning the existence of non-binary people in a footnote does little to minimize the harms a system may cause them. Much inspiration here may come from third wave feminism and queer theory [De Lauretis, 1990, Jagose, 1996], and perhaps more closely the recent movement within Human Computer Interaction toward Queering HCI Light [2011] and Feminist HCI Bardzell and Churchill [2011]. The goal that queer theory has of deconstructing social norms and associated taxonomies is particularly important as NLP technology addresses more and more socially relevant issues, including but not limited to issues around gender, sex and sexuality.

Limitations The primary limitation is that our experiments and analysis are largely limited to English: our consideration of task definition in subsection 3.4.1 discusses other languages, but all the data and models we consider are for English. This is particularly limiting because English lacks a grammatical gender system (discussion in subsection 3.3.2), and some extensions to languages with grammatical gender are non-trivial. We also emphasize that while we endeavored to be inclusive, in particular in the construction of our datasets, our own positionality has undoubtedly led to other biases. One in particular is a largely Western bias, both in terms of what models of gender we use (e.g., the division of sex, gender, and sexuality along a Western

frame; see section 3.3) and also in terms of the data we annotated (subsubsection 3.4.4.3). We have attempted to partially compensate for this latter bias by intentionally including documents with non-Western binary and non-binary trans expressions of gender in the GICoref dataset, but the compensation is incomplete.

Additionally, our ability to collect *naturally occurring* data was limited because many sources simply do not yet permit (or have only recently permitted) the use of gender inclusive language in their articles (discussion in subsection 3.4.2). This led us to counterfactual text manipulation in subsection 3.4.3, which, while useful, is essentially impossible to do flawlessly (additional discussion in subsubsection 3.4.3.3). Finally, because the social construct of gender is fundamentally contested (discussion in subsection 3.3.1), some of our results may apply only under some frameworks.

3.6 Conclusion

Our goal in this chapter was to take a singular task—coreference resolution—and identify how different sources of bias enter into machine learning-based systems for that task. These biases can lead to systems that are misaligned with the needs and values of a broad spectrum of users, particularly impacting marginalized groups. The investigation extends beyond the commonly acknowledged sources of bias in training datasets and model architecture. We have demonstrated that biases can infiltrate every aspect of the machine learning workflow, including the initial task definition, system testing phases, and feedback loops, particularly when the development process neglects the inclusion of significant stakeholder groups. This finding accentuates the imperative for a closer integration of AI system development with its stakeholders and advocates

for a more human-centric methodology in the advancement of artificial intelligence.

Chapter 4: Alignment with Users' Needs through Specialized User Contexts

Building on the insights from Chapter 3, which examined how biases can enter AI systems when developers fail to consider real-life stakeholders, this chapter delves into the challenges of aligning AI systems with the needs of specific user groups. We explore two distinct application contexts: content moderation assistance for volunteer moderators and visual question answering for blind and visually impaired users. In both cases, we identify gaps between the capabilities of existing AI technologies and the unique requirements of these specialized user groups. By engaging with users directly and analyzing real-world data, we uncover valuable insights that can guide the development of AI systems to better serve the needs of their users. This chapter highlights the importance of a human-centered approach in creating AI solutions that effectively empower and collaborate with users across various domains.

4.1 Content Moderation Assistance for Volunteer Content Moderators

Joint work with Lovely-Frances Domingo, Sarah Ann Gilbert, Michelle Mazurek, Katie Shilton, Hal Daumé III. Under Submission

4.1.1 Overview

Content moderation guards online forums against hostility and extremism while maintaining community norms, ensuring the forums remain healthy and open to all participants. While many platforms pay for this service, others, such as Reddit, Discord, Facebook, and Twitch, use a hybrid model, relying on the labor of volunteers. Yet, behind the screen, being a volunteer content moderator is time- and emotionally-draining work. Moderators frequently deal with abusive language, sensitive posts, and unpleasant interactions with users Seering et al. [2019], Gilbert [2020], Dosono and Semaan [2019], Wohn [2019], Jiang et al. [2019], often doing this work in addition to full-time jobs. To support these volunteers, efforts have been made to develop models, such as Google Perspective API[1] and OpenAI undesired content detection Markov et al. [2023], that can automatically identify content for removal in order to alleviate moderators' workload.

Although these systems have shown great success in detecting "undesired" content, they primarily focus on toxic content. Yet, content moderation encompasses more than toxicity detection,[2] particularly in platforms that leverage volunteer moderation within smaller communities hosted by the site. For example, Reddit is a platform consisting of various communities, known as "subreddits," focused on a diverse set of topics, and each subreddit has its own moderation rules. Fiesler et al. [2018] conducted a study to explore various subreddit rules, consolidating similar ones, and arrived at 25 distinct rule types. Hence, in order to support moderators in detecting potential rule-violating content, content moderation tools need to support much more than just toxicity detection.

We thus aim to assess to what extent current natural language processing (NLP) models

[1] https://perspectiveapi.com/

[2] We use "toxicity detection" as an umbrella term for hate speech detection, incivility detection, etc.

can serve the wide spectrum of moderation rules that exist. First, to understand the functions previous automated content moderation models have focused on, we conduct a *model review* on Hugging Face (HF) with rules from three subreddits as exemplars. This allows us to gauge the progress of past model developments in covering various moderation rules. We use model review as opposed to the more common literature review to gain a technical understanding of the existing models' functions. In addition to examining models that are built to handle specific tasks, we also assess so-called "general-purposed" large language models' (LLMs') capability in covering various moderation rules. We evaluate GPT-4 and Llama-2 on a new evaluation dataset that we collected, consisting of moderation decisions from *r/AskHistorians* and covering a wide range of rules. Finally, using the models' performance on this new dataset as an empirical grounding, we conduct a survey study with *r/AskHistorians* moderators ($N = 11$) to evaluate the models' performance and gain insights on performance requirements for a useful model for different rules.

We find a substantial gap in both existing NLP models and LLMs' performance when to cover the diverse set of moderation rules that subreddits employ. Our analysis shows the majority of moderation rules from the three subreddits (approximately 80%) are unrelated to toxicity detection, with nearly 70% lacking an huggingface model designed for their resolution. While one might hope that general-purpose LLMs could fill this gap, our experiments with GPT-4 and Llama-2 show that LLMs fall short in their ability to detect violations of many rules. Specifically, both GPT-4 and Llama-2 exhibited moderate to low precision and/or recall ($< 70\%$) for half of the rules from *r/AskHistorians*. Findings from our survey study also indicate that neither LLM has good enough performance to be useful for 6 of 23 *r/AskHistorians* rules (26%). Finally, our survey study shows that moderators are excited about an assistant model but have

complex needs in terms of model precision and recall for different kinds of rules. We highlight the necessity of, and provides a guide for, future content moderation work to expand its focus beyond toxicity detection, encompassing a broader spectrum of moderation rules to meet the needs of moderators seeking automation assistance.

4.1.2 Related Work

Volunteer moderators have been a staple of online content regulation from the earliest days of the internet, tackling issues such as spam Seering [2020], trolling Binns [2012], and abuse Dibbell [1994]. Currently, volunteers contribute moderation efforts on nearly all major platforms. For example, volunteers are responsible for moderation within Facebook's Groups, Discord's servers, Twitch's streams, Reddit's subreddits, and through X's (formally known as Twitter) Community Notes. These efforts bring significant value to platforms, with one study estimating that, at the baseline, volunteer moderators on Reddit collectively work 466 hours a day, amounting to a value of about 3.4 million USD per year Li et al. [2022]. With the workload and continuous exposure to online harassment Gilbert [2020], Wohn [2019], Dosono and Semaan [2019] and toxic content, such as hateful language and violent or upsetting imagery, this moderation work can easily result in burnout.

A common way to mitigate burnout is through the development of tools that support moderation labor, particularly through automation. For example, prior study has found that automation can help at scale, supporting moderators of large communities with high levels of activity Kiene and Hill [2020], Seering et al. [2018]. Automation is also helpful when communities experience unprecedented or unexpected growth, such in cases where communities receive an influx of new

users Kiene et al. [2016] and in cases where communities are subject to sustained brigades of bad actors spamming hateful content Han et al. [2023]. Mostly, automated tools are developed and maintained by moderators themselves. For example, the most commonly used tool on Reddit, Automod, was originally designed by a moderator before the company took over responsibility for its development and maintenance Jhaver et al. [2019] and moderators on Twitch have developed bots on the fly in response to hate raids Han et al. [2023]. While automation can help support moderation work, it may also add to it in different ways—like needing to build and maintain a bot—and requires skill sets not all moderators have Jhaver et al. [2019].

Meanwhile, the NLP field has a long line of work on developing automated models to detect toxic content, which can be used to support moderation work [e.g. Jigsaw, 2019a, Pavlopoulos et al., 2020, Park and Fung, 2017, Zampieri et al., 2019]. Beyond that, Lees et al. [2022] extended beyond English content and proposed the multilingual Perspective API to detect offensive and hateful content from a diverse range of languages. Moreover, some study delves into various types of toxic content, trying to improve the nuance of detection models Markov et al. [2023], Price et al. [2020]. However, these studies primarily focus on detecting toxic content, which is merely a subset of moderation work. Therefore, we use Reddit rules as an example to identify gaps in NLP moderation models within a wide spectrum of rules.

Recently, Kumar et al. [2023] tested LLMs' performance on rule-based moderation with rules from 95 subreddits. They evaluated GPT models (3, 3.5, 4) on a dataset consisting of removed Reddit posts as violating and not violating content. They conclude that the GPT models are effective in conducting moderation on subreddits like *r/Movies* but struggle with other subreddits such as *r/askscience* and *r/AskHistorians*. We build on previous results by studying gaps in both purpose-built moderation models and LLMs. Rather than focusing on the broad binary

questions of removing or keeping posts, we focus on coverage for specific moderation rules. In addition, we include the perspectives of human moderators to understand how existing models address the needs of real-world moderation.

4.1.3 Hugging Face Model Review

One traditional approach to understanding a scientific context is to perform a literature review. Because we are particularly interested in deployable models, we opt for a *model review*, where instead of reviewing papers, we review publicly available models that can be adapted to assist content moderators in making moderation decisions, with the focus on Reddit rules. We conduct the model review via Hugging Face (HF)[3], an open-source platform that provides pre-trained models for various NLP tasks, to understand the progress of model development from existing research in covering various moderation rules. To conduct the model review, we gather rules from three subreddits and manually investigate for each rule whether there exists an HF model suitable for detecting such rule violations.

4.1.3.1 Method

Chandrasekharan et al. [2018] clustered 100 subreddits into six Meso groups plus four Micro groups based on similarity of the removed posts. Since Micro groups are subreddits, whose rules are more specific to the particular subreddit, we focus on the Meso groups, whose rules are shared across multiple subreddits within the group. We pick three subreddits – *r/Atheism, r/Movies, r/AskHistorian*, from three of the major Meso subreddit clusters and use their rules as representatives for the study. We verify the coverage of the three subreddits' rules by cross-

[3] https://huggingface.co/models

referencing them with the list of 25 rule types from Fiesler et al. [2018], ensuring that all text-based rule types are covered by at least one of the rules from the three subreddits. See appendix Table B.1 for our list of rules and its cross-referencing with the list of rule types.

Procedure For each rule, we first crawl its rule description from Reddit and answer the following two questions based on the description.

► **RULE-BASED?**: Can detection of violations on this rule be done by rule-based approaches, e.g. by regular expressions (yes or no)? We first mark out the rules that can easily be handled by rule-based approaches, such as detecting reposts. If the answer is yes, we do not consider this rule any further; all other rules are fully analyzed using following steps.

► **TOXIC?**: Is this rule covered by toxicity detection (yes or no)? To understand what portion of the policies are related to toxicity detection, we mark the rules that can be handled by toxicity detectors, such as detecting harassment.

Then, we search on HF for a *matching model* to each rule. Based on the rule descriptions and our knowledge of previous NLP task names, we identify keywords for model searching. For example, for the rule `Personal Attacks or Flaming`: *"Keep things civil. Avoid fighting words and personal attacks. (applies to all forms of user content, including the user's name.)"*, we use keywords `doxxing` and `cyberbullying`.

We then search on HF by matching keywords with model names or model cards to find the most relevant model to each moderation rule. We traverse the top 20 search results to find the most relevant model, which we call the *matching model*. If more than one model is relevant, we select the model with the highest number of likes on HF. If nothing relevant is found, we declare there is no matching model for this moderation rule. We also skip any model that does not contain

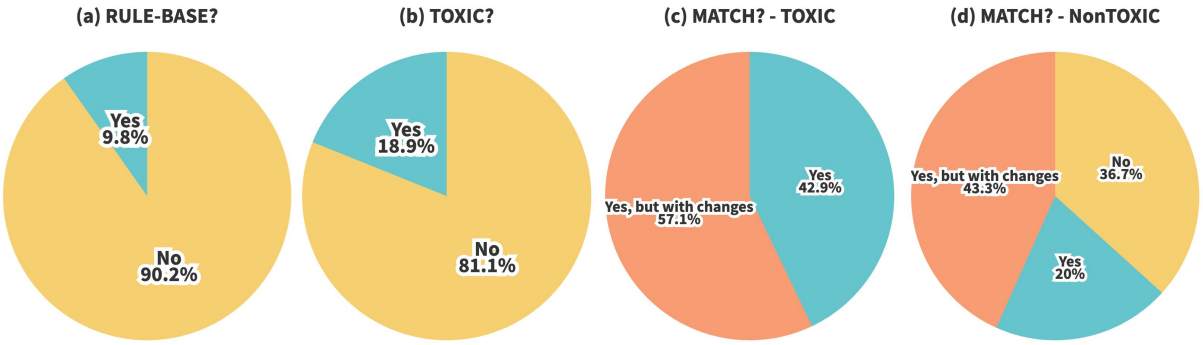

Figure 4.1: Model review annotation results. Figures (a) and (b) are the results of annotation questions **RULE-BASED?** and **TOXIC?**, respectively. Figures (c) and (d) are the results of question **MATCH?** for rules related to toxicity detection and not related, respectively.

a model card (i.e., no model description). See Tables B.3, B.4, and B.5 in the appendix for the keywords we used and the matching models for each rule.

With the matching model, we then answer the following questions for each moderation rule:

► **MATCH?**: Is there an HF model that can be used to detect violations of this rule? Possible answers are *yes*; *yes, but with changes*; or *no*. We mark *yes, but with changes* when the model is intended to be used for tasks similar to detecting violations of this rule, but some adjustment are needed for the model to be useful in this case. The adjustments needed are noted in the last question.

► **PURPOSE?**: What is the original purpose of this matching model? We note the intended use case for the model as claimed in the model card.

► **GAP?**: In order to adapt this model to detect violations of this rule, what needs to be adjusted? We write the changes, if needed, for the model to be used for this rule, such as domain adaptation.

4.1.3.2 Model Review Results

As shown in Figure 4.1, among the 41 moderation rules, 37 (90%) of the rules cannot be handled by rule-based approaches. Among these, only 7 (19%) are toxicity-related. This reemphasizes that models solely designed for detecting toxicity fall short in meeting the practical needs of content moderators.

Among the 7 toxicity-related rules, we see that 4 (57%) require model modifications. Most of these modifications are customization for the rule, which we will describe later. Even toxicity-related rules, despite having many developed models, do not have perfectly matching models.

Among the 37 non-toxicity-related rules, only 6 (20%) rules have matching HF models that can be applied directly. 13 (43%) of the rules require some model modifications in order for the matching model to be adapted for the rule, whereas 11 (37%) rules have no matching model at all. These rules include, for example, `low-effort posts` and `proselytizing` from *r/Atheism*, and `no homework` and `no "soapboxing" or loaded questions` from *r/AskHistorians*.

Overall, many rules lack a matching model. For the rules that have one, most require non-trivial model modifications to be covered. These results indicate a substantial gap between the functionalities of previously developed models and those needed for Reddit moderation.

For the **GAP?** question, we primarily identify four types of gaps.

1. The most common adjustment is domain adaptation to the rule-related topic or to Reddit texts. For example, the `No Ambiguous/Misleading/Inaccurate Information` rule in *r/Movies* can be handled with a misinformation detection model, but the model needs to be adapted to title-like texts and information about movie releases.

2. Some models are developed with a limited scope of training data and thus may require extended training to be used for moderation. For instance, we found a plagiarism model that was trained on the Machine Paraphrase Corpus Wahle et al. [2022] consisting of $\sim 200k$ examples of original and paraphrases using two online paraphrasing tools. The plagiarism cases on Reddit may be more complicated than paraphrasing, and the potential plagiarism source is larger. Hence, the model needs an extension of scope in order to be useful.
3. Models for some rules, especially toxicity-related rules, require customization. For example, *r/Atheism* has the rule `Harassment or Bigotry` that prohibits harassment but permits curse words. A toxicity detection model can be used here but needs modification.
4. Some models can be applied to certain rules, but with a degree of stretching. For instance, violations of the rule `Off Topic` can potentially be caught by a topic classifier model, but the model requires some major changes to adapt to the subreddit and its related topics.

To see detailed annotations for each question, please refer to Tables B.3, B.4, and B.5 in appendix.

4.1.4 LLM Performance on Subreddit Rules

Finding a suitable NLP model for handling a moderation rule, let alone executing the necessary modifications to the model, is not trivial, and some rules simply lack a matching model to address their specific issues. It requires familiarity with NLP techniques and tasks, which may not be within the scope of a content moderator's abilities. Question-answering-based LLM applications, on the other hand, may be more practical for content moderators to employ in assisting their jobs. Therefore, we also evaluate LLMs' performance in catching rule violations. For this evaluation, we focus on the *r/AskHistorians* subreddit, which has 23 moderation rules, and two

LLMs — Llama-2 and GPT-4.

4.1.4.1 Method

Evaluation Dataset The evaluation dataset was created and annotated by a volunteer moderator for *r/AskHistorians* who has expertise in both moderation and qualitative analysis. For each rule, 11–13 violating posts or comments were selected, save for posts violating the `Illegal artifacts` rule, which, due to the rareness of violations, was only represented 3 times in the dataset. Most content was selected from *r/AskHistory*, a Reddit community that is thematically similar to *r/AskHistorians* but with different moderation rules. Some additional content was taken directly from *r/AskHistorians*. We exclude all borderline cases and only select posts and comments that clearly violate the rules. In total, the dataset includes 101 questions and 134 comments.[4] Because content often violates more than one rule, questions and comments included in the dataset often reflect multiple violations; however, annotations represent the most relevant or severe violation. For example, a comment containing only a joke would be annotated as violating the rule prohibiting joke responses, even though it would also violate rules on comprehensiveness or depth. The dataset also included 11 questions and 10 comments that do not violate any *r/AskHistorians* rules. We use these data as our evaluation dataset.

Model Testing and Prompts To assess the performance of LLMs, we test GPT-4 and Llama-2-13b models. We conduct a pilot experiment with a small number of randomly selected examples from our dataset to determine the prompt for model assessment. Since moderators are mostly unfamiliar with prompt tuning and have limited time to learn this skill, we aim to test the models

[4] In *r/AskHistorians*, posts are questions users have asked and comments are responses to questions. Comments included in the dataset are not necessarily responses to the questions included in the dataset.

without extensively experimenting different wordings or examples of the prompt, roughly as they might be used by moderators.

First, as suggested in the prompting guidelines Shieh [2023], we find that using rule descriptions in the prompt is helpful for the model to detect violations of the rule. Hence, we crawl rule descriptions from *r/askHistorians* rule explanation web page. As the descriptions from the subreddit web page are long and include many constructive suggestions for post-writers, we manually select useful information for moderation to include in the prompt. See appendix Table B.2 for the list of the rules and their descriptions.

Moreover, we experiment with both zero-shot and few-shot settings. In the few-shot setting, we provide the model with three examples that violated the rule and three examples that did not for question contents. For comment contents, we provide one example for each due to the model's input length limit. We realize that few-shot examples are helpful for the Llama-2 model to better understand the task, whereas the GPT-4 model does not have significant improvement from incorporating the examples. Therefore, for the main experiment, we test GPT-4 with the zero-shot setting and Llama-2 with the few-shot setting, but we also report the full set of experiments in the appendix.

Finally, following the suggestions from Shieh [2023], we set the model's role as a moderator assistant and put the main question at the beginning of the prompt. In the few-shot setting, we also repeat the main question at the end of the prompt. See following for the exact prompts we used.

Prompt for Testing Llama-2 with the Few-shot Setting

```
<s>[INST] «SYS»
```

```
You are a moderator assistant.  There is a moderation policy named {rule} that
says {rule description}.  Does the following content violate the {rule} policy?
Simply give a yes or no answer.
«/SYS»
{violating example content}
Does the following content violate the {rule} policy?
Simply answer yes or no.[/INST] Yes </s><s>[INST]
{not violating example content}
Does the following content violate the {rule} policy?
Simply answer yes or no.[/INST] No </s><s>[INST]
{test content}
Does the following content violate the {rule} policy?
Simply answer yes or no.[/INST]
```

Prompt for Testing GPT-44 with the Zero-shot Setting

```
"role":  "system", "content":  "You are a moderator assistant.",
"role":  "user", "content":  "There is a moderation policy named {rule} that says
{rule description}.  Does the following content violate the {rule} policy?  Simply
give a yes or no answer.",
"role":  "user", "content":  {test content},
```

4.1.4.2 Moderator Evaluation

To understand volunteer moderators' perspectives on the relative performance of LLMs, we also conducted an evaluation survey with active moderators from *r/AskHistorians*.[5] The sur-

[5]Approved by our institutional IRB, #1704882-8.

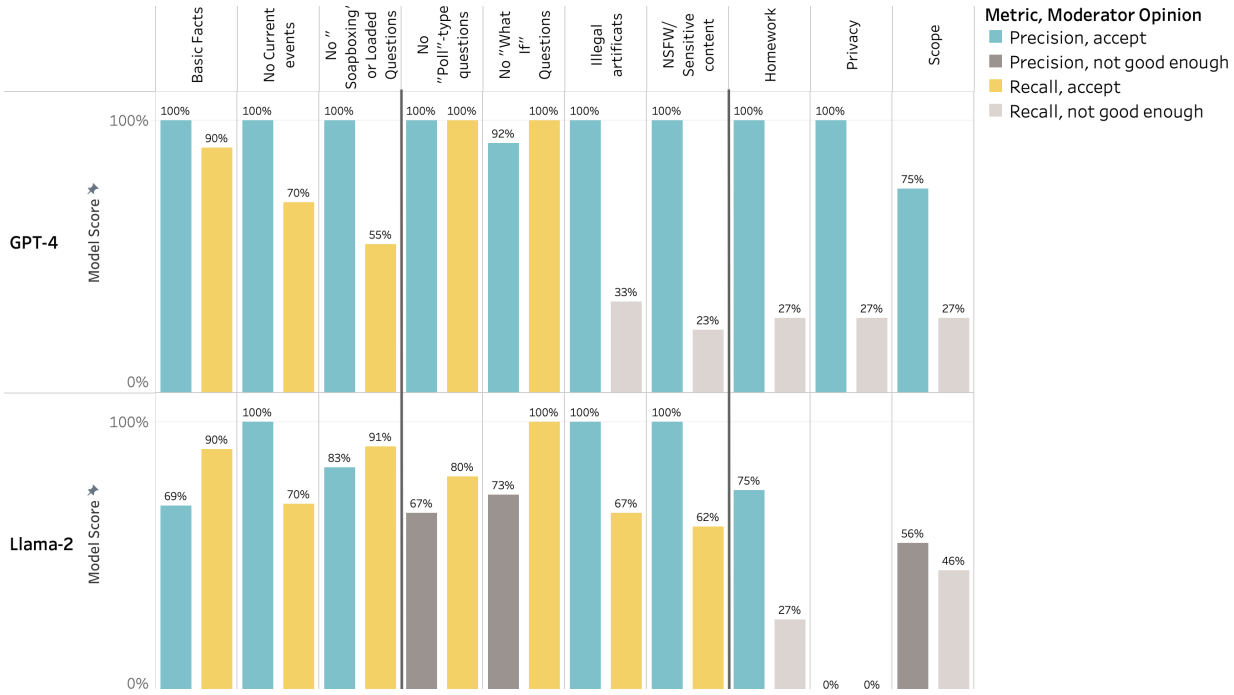

Figure 4.2: Model performance on detecting question posts that violate moderation rules with GPT-4 (top) and Llama-2 (bottom) models. The x-axis is the moderation rules for question posts. Each bar pair is the precision (left) and recall (right) scores on the specific rule. The ones marked grey are the model scores that moderators would not consider useful as a moderation assistant tool. The left-most rules have good performance from both of the LLMs; the rules in the middle have good performance from at least one of the LLMs; the right-most rules do not have good enough performance from either of the LLMs.

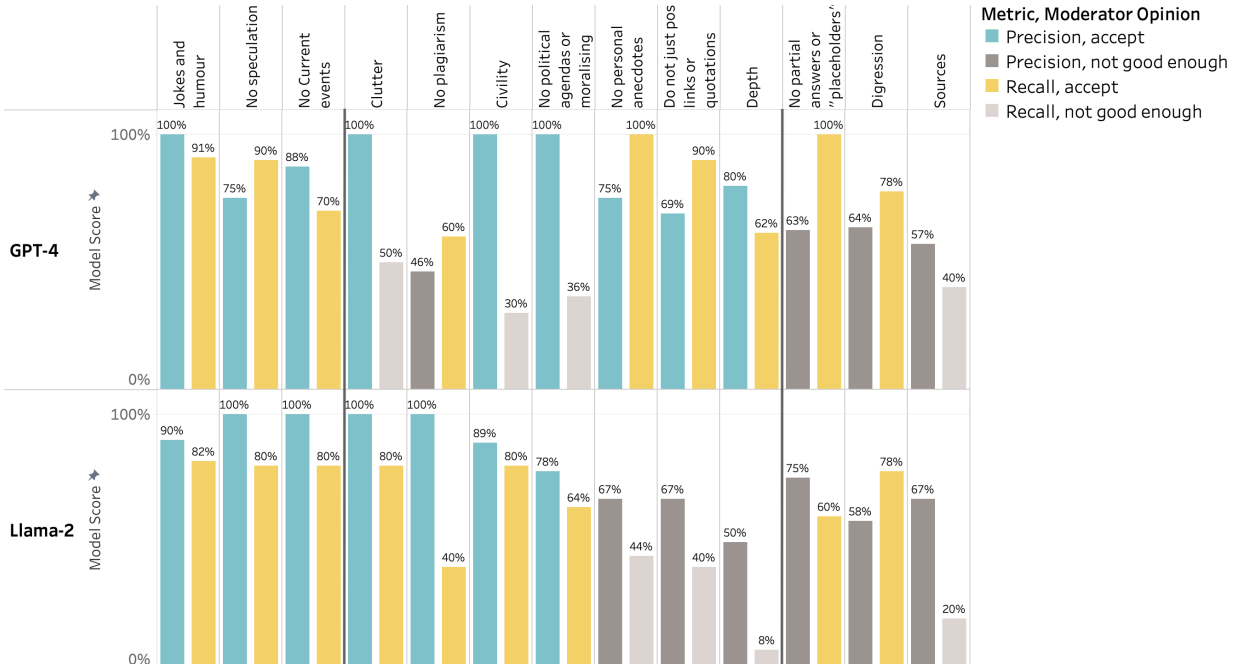

Figure 4.3: Model performance on detecting comment posts that violate moderation rules with GPT-4 (top) and Llama-2 (bottom) models. The x-axis is the moderation rules for comment posts. Each bar pair is the precision (left) and recall (right) scores on the specific rule. The ones marked grey are the model scores that moderators would not consider useful as a moderation assistant tool. The left-most rules have good performance from both of the LLMs; the rules in the middle have good performance from at least one of the LLMs; the right-most rules do not have good enough performance from either of the LLMs.

vey aimed to capture 1) whether moderators would use the LLM to help with their moderation work on each rule, given the LLM's performance (precision and recall) identifying violations for this rule, and 2) for each rule, what kind of model performance is needed to effectively assist moderators.

To answer the first question, we asked participants to evaluate each model's performance. We showed participants the LLMs' precision and recall scores for each rule from our experiment and asked them to choose among: 1). *performance for this rule is good enough that I would use the tool*, 2). *I would use this tool for this rule if it could catch more violations* (need higher recall), 3). *I would use this tool for this rule if it could be correct more often* (need higher precision), or 4). *both measures would have to improve for me to use this tool.*

For the second question, we asked participants to cluster moderation rules based on how important it is for the model to have high precision (or high recall) for each rule, compared to other rules. Options included: *most important, less important,* and *would not use a model for this rule*. We also asked for participants' comments on using LLMs to identify violating posts in general. See appendix Figures B.3 and B.4 for the exact questions we asked.

Overall, by invitation, we recruited 11 out of 36 total active moderators from *r/AskHistorians*. We obtained their consent at the beginning of the survey. Our participants spent an average of 25 minutes completing the survey. In return, participants received compensation in the amount of $25 USD upon completion of the survey in full.

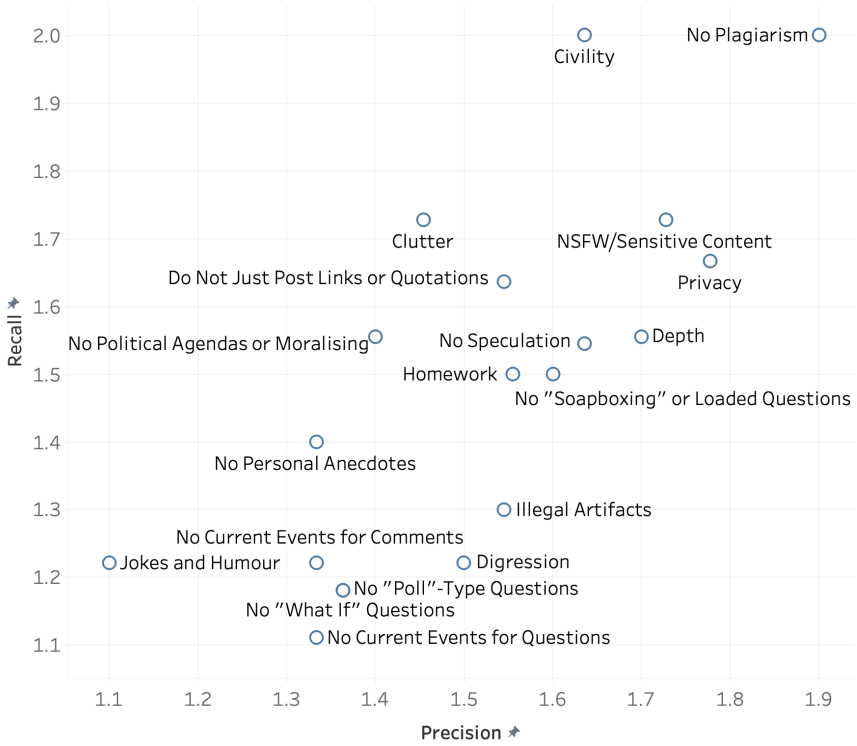

Figure 4.4: Clustering results from the survey study. Each point is a moderation rule from *r/AskHistorians*. The x-axis shows the precision importance score, or how important it is for a model to have high precision for this rule. Similarly, the y-axis shows the recall importance score. Note that we removed rules for which at least three participants stated they would not use a model.

4.1.4.3 Model Performance – Results

We tested GPT-4 and Llama-2 on the evaluation dataset as we described above. We use the first occurrence of "Yes" or "No" in the generated text as the model's output. The results are shown in Figure 4.2 and Figure 4.3 (See appendix Figure B.1 and Figure B.2 for the results of the models in both few-shot and zero-shot settings). The results are mixed – some rules have both high precision and high recall from at least one model, while many of the rules have moderate to low precision, recall, or both. Neither of the two models is clearly better than the other in identifying violating contents for all the rules. The results indicate that certain rules remain unresolved, even for state-of-the-art LLMs.

Our survey results indicate that different moderators have different perceptions of what constitutes an adequate model: for the evaluation questions, participants exhibited only fair agreement (Fleiss' Kappa 0.34). Mirroring policy decision-making standards in the subreddit, we use majority vote to summarize participants' evaluation of the models' performance. The results are shown in Figure 4.2 and Figure 4.3. In most cases, moderators would consider models with $> 70\%$ precision and recall to be useful. Among the 23 rules, both models are considered useful for three question rules and three comment rules. For some rules, such as `NSFW/Sensitive content`, `No plagiarism`, `No political agendas or moralising`, and `Depth`, moderators can accept low recall; for some, such as `Do not just post links or quotations`, moderators can accept low precision. Importantly, there are six rules (26% of all rules) for which neither model is considered useful. This emphasizes that, while current LLMs provide possibilities for supporting content moderation work, further improvement is needed to support useful moderation tools.

Most of the moderators were excited about the idea of having even imperfect assistant models to support their work. Two participants expressed willingness to accept a model with moderate levels of precision and recall, because *"any help is better than no help."* One stated *"A tool flags a lot of things but isn't right too often? Well, it's still flagging things for my attention. A tool that doesn't flag a lot but is very frequently right? Great, I don't have to think about whether or not to remove what it's found."* This response emphasizes the importance of model transparency. Model users often can adapt to a flawed model and make it useful as long as they are informed of its performance limitations.

Some moderators value precision more than recall because a model that mis-flagged many posts would make moderators' workload unnecessarily large. On the other hand, some moderators consider recall to be more important to catch all violating posts missed by moderators, especially for urgent or sensitive content, such as for the rule `Civility`. Moreover, some participants state that the need for the model varies according to the rules. One participant specified, *"A tool would be useful for assessing and filtering breaches of more complex rules like plagiarism (especially Chat GPT use) or poor sourcing, but given the time needed to assess these claims it would need to be correct the overwhelming majority of the times it flagged something. For simpler rules like clutter, depth, or no jokes then I would like to see a tool which is able to flag the majority of comments which break the rules."* The same participant also pointed out that they would not use a model for rules that are inherently subjective between moderators, such as `Basic facts`. Another participant further indicated the need for model explanations for complex rules.

To inform future improvement on moderation assistant models, we asked moderators to cluster rules according to their needs. The results, as in Figure 4.4, provide quantified insight

into moderators' needs for model precision and recall. We aggregate participants' answers by averaging their choice of cluster for each rule. Each time a participant assigned a rule to *most important*, it is scored as two points; assignments to *less important* are scored as one point. Rules for which a participant selected *would not use* are not included in the score calculation but reported separately. Among the 23 rules, ten require both high precision and high recall (> 1.5). For three rules, at least three participants claimed they would not use a model — `Basic facts`, `Scope`, and `Sources`. We hope our results can serve as a guide for future moderation assistant models to better align with moderators' needs.

4.1.5 Discussion and Limitations

Overall, our findings suggest potential directions for future research on moderation assistants. First, there remains ample opportunity for model improvements. There are ten rules that moderators require both high precision and high recall, as in Figure 4.4. For three of them, neither GPT-4 nor Llama-2 are considered useful by moderators: `Privacy` and `Homework` with low recall (27%), and `No partial answer or "placeholders"` with low precision (63%). For four of the ten rules, although moderators mark one model as acceptable, there is still room for improvement in precision or recall: `Plagiarism`, `NSFW/Sensitive content`, and `Depth` with recall $40 - 62\%$, and `Do not just post links or quotations` with precision 69%. Besides, the rule `Digression`, which requires high precision, has low precision from both models (58% and 64%). Though some rules are distinct for *r/AskHistorians* (e.g., `Homework`), many rules are also enforced in other subreddit communities. For example, among the 523 subreddits studied by Fiesler et al. [2018], 39% have rules on off-topic content (similar to

the `Digression` rule in our study) and 27% have rules on NSFW content. Hence, our findings underscore the existence of plenty of rules from online communities beyond toxicity detection that require exploration.

Despite the models not performing as well as one might hope, volunteer moderators express significant enthusiasm and flexibility about having moderation assistants. People are used to working with tools that may not be perfect, and they are skilled at finding ways to make the most of them. In our findings, moderators explained several personal approaches they take in using a model with moderately low precision or with moderately low recall. However, in order to exercise these strategies, models' abilities must be transparent to enable users to adapt accordingly.

Furthermore, we observed varying requirements among users and rules. Even for the same rule, different moderators expressed distinct preferences for high precision or recall. This suggests that in situations where a trade-off between precision and recall is necessary, moderators may lean towards a model offering more user control, thus allowing for an adjustable trade-off. This way, they can customize the model to align with their preferences. In addition, for complex rules (e.g. `Plagiarism` and `Digression`) or subjective rules (e.g. `Basic facts` and `Scope`), moderators prefer model explanations over mere flagging of violations.

Finally, this study showcases the importance of involving direct stakeholders (e.g. moderators) in the design and development loop for AI tools. They not only ease the alignment of model construction with users' needs but also provide insights model developers might otherwise miss.

Limitation This study has several limitations. We are primarily studying Reddit; other platforms may have different sets of rules. We are also focusing on English and text-only posts; violations in other languages or in other forms, such as images and videos, may face different

challenges. Furthermore, our survey study is limited to volunteer content moderators from the *r/AskHistorians* community. While we anticipate that the insights gained from the moderation experience might have broader applicability across communities and for other moderators, the extent of generalization remains uncertain without further study.

Moreover, this study and findings are grounded in the set of Reddit moderation rules rather than the frequency of their violation. Consequently, we lack information regarding which rules moderators encounter most frequently and for which rules they want additional assistance.

In addition, our model testing is limited to one-shot and few-shot contexts, as opposed to other strategies, such as interactive teaching and Chain-of-Thoughts.

4.2 Visual Question Answering Systems for Blind and Visually Impaired People

Joint work with Kyle Seelman[1], Kyungjun Lee[1], Hal Daumé III. Appeared at AACL 2022

In this section, we explore how machine systems can empower humans with disabilities and, thus, have unique requirements. Much research has focused on evaluating and pushing the boundary of machine "understanding" – can machines achieve high scores on tasks thought to require human-like comprehension, including image tagging and captioning [e.g., Lin et al., 2014], and various forms of reasoning [e.g., Wang et al., 2018, Sap et al., 2020]. In recent years, with the advancement of deep learning, we saw great improvements in machines' capabilities in accomplishing these tasks, raising the possibility for deploying machine systems for assisting

[1]Equal contribution.

human in real-life tasks. In this section, we delve into the visual question answering (VQA) task to investigate opportunities and challenges of adapting machine understanding VQA systems to assist visually-impaired people.

4.2.1 Overview

Visual question answering (VQA) is a task that requires a model to answer natural language questions based on images. This idea dates back to at least to the 1960s in the form of answering questions about pictorial inputs [Coles, 1968, Theune et al., 2007, i.a.], and builds on "intelligence" tests like the total Turing test [Harnad, 1990]. Over the past few years, the task was re-popularized with new modeling techniques and datasets [e.g. Malinowski and Fritz, 2014, Marino et al., 2019]. However, besides the purpose of testing a models' multi-modal understanding, VQA systems could be potentially beneficial for visually impaired people in answering their questions about the visual world in real-time. For simplicity, we call the former view *machine understanding VQA* (henceforth omitting the scare quotes) and the latter *accessibility VQA*. The majority of research in VQA (subsection 4.2.2) focuses on the machine understanding view. As a result, it is not clear whether VQA model architectures developed and evaluated on machine understanding datasets can be easily adapted to the accessibility setting, as the distribution of images, questions, and answers might be—and, as we show, are—quite different as shown in Figure 4.5.

We aim to investigate the gap between the machine understanding VQA and the accessibility VQA by uncovering the challenges of adapting machine understanding VQA model architectures on an accessibility VQA dataset. Here, we focus on English VQA systems and datasets; for

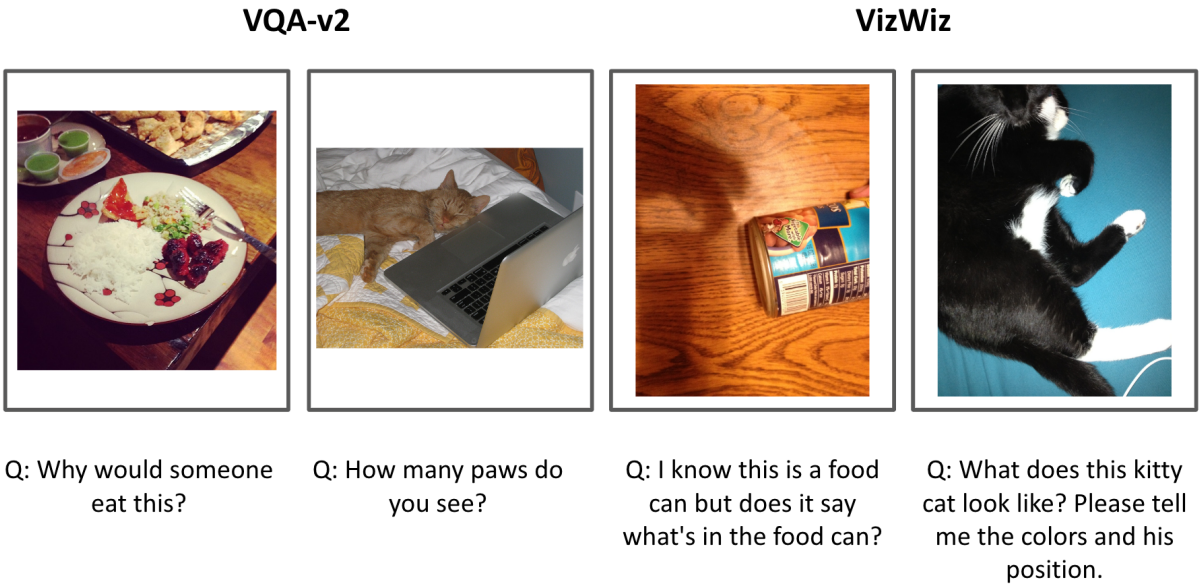

Figure 4.5: Given similar image content (left: food, right: cat), questions in the machine "understanding" dataset VQA-v2 and the accessibility dataset VizWiz are substantially different. In the VizWiz examples, we see questions that are significantly more specific (with one question even explicitly stating that it's already obvious that this is a can of food), more verbal, and significantly less artificial (as in the cat examples).

machine understanding VQA, we use the VQA-v2 dataset Agrawal et al. [2017], while for accessibility VQA, we use the VizWiz dataset Gurari et al. [2018] (subsubsection 4.2.3.1). Through performance assessments of seven machine understanding VQA model architectures that span 2017–2021 (subsubsection 4.2.3.3), we find that model architecture advancements on machine understanding VQA also improve the performance on the accessibility task, but that the gap of the model performance between the two is still significant and is *increasing* (subsubsection 4.2.4.1). This increasing gap in accuracy indicates that adapting model architectures that were developed for machine understanding to assist visually impaired people is challenging, and that model development in this area may indicate architectural overfitting.

We then further investigate what types of questions in the accessibility dataset remain hard for the state-of-the-art (SOTA) VQA model architecture (subsubsection 4.2.4.2). We adopt the

data challenge taxonomies from Bhattacharya et al. [2019] and Zeng et al. [2020] to perform both quantitative and qualitative error analysis based on these challenge classes. We find some particularly challenging classes within the accessibility dataset for the VQA models as a direction for future work to improve on. Additionally, we observe that many of the questions on which state-of-the-art models perform poorly are not due to the model not learning, but rather due to a need for higher quality annotations and evaluation metrics.

4.2.2 Related Work

Brady et al. [2013] conduct a thorough study on the types of questions people with visual impairments would like answered, and provide a taxonomy for the types of questions asked and the features of such questions. It was a significant step in understanding the need in people with visual impairments for VQA systems. In combination with our study, this gives a more complete picture of what kinds of questions not only contribute to better model performance, but actually help individuals with visual impairments. Additionally, Zeng et al. [2020] seek to understand the task of answering questions about images from people with visual impairments (i.e., VizWiz) and those from sighted people (i.e., VQA-v2). The authors identified the common vision skills needed for both scenarios and quantified the difficulty of these skills for both humans and computers on both datasets.

Gurari et al. [2018], who published a very first visual question answering (VQA) dataset, "VizWiz" containing images and questions from people with visual impairments, pointed out the artificial setting of other VQA datasets that include questions that are artificially created by random sighted people. The VizWiz challenge is based on real-world data and directs researchers

working on VQA problems toward real-world VQA problems. This dataset was built on data collected with a crowdsourcing app, where users with visual impairments share an image and a question with a sighted crowdworker who answer the question for them Bigham et al. [2010]. Other existing datasets, such as VQA Antol et al. [2015], DAQUAR Malinowski and Fritz [2014], and OK-VQA Marino et al. [2019], are different in that their questions were not provided by those who took images. Instead, the images were first extracted from web searches, and then questions were later provided by sighted crowdworkers who viewed and imagined questions to ask about those images. Here, we see that people with visual impairments can benefit the most from VQA technology but most of the existing VQA datasets do not involve people with visual impairments.

Some prior studies has investigated VQA datasets further, focusing on assessing diversity in answers to visual questions. For instance, Yang et al. [2018] looked at answers to visual questions created by blind people and sighted people and worked on anticipating the distribution of such answers. Predicting the distribution of answers asked, they helped crowdworkers create as many unique answers as possible for answer diversity. Bhattacharya et al. [2019] tackle the same issue by looking at images of VQA. They proposed a taxonomy of nine reasons that cause differences in answers and developed a model predicting potential reasons that can lead to differences in answers. However, little work explores discrepancies between questions from actual users of VQA applications (i.e., users with visual impairments) and contributors who helped develop data for VQA applications.

We aims to understand this gap by assessing the discrepancies between the dataset containing artificially created data and the dataset containing real-world application data present across differnet VQA models. More specifically, we assess the performance of VQA models that were proposed in different times and delve into the old model and the state-of-the-art model with in-

dividual datapoints to identify patterns where the models perform poorly for the accessibility dataset.

4.2.3 Experiment Setup

To evaluate how existing VQA models' performance on machine understanding dataset align with performances on the accessibility dataset, we select two VQA datasets and seven VQA models. One of the dataset, VQA-v2, was proposed for machine understanding, whereas the other dataset, VizWiz, was collected to improve accessibility for visually-impaired people. The seven VQA models, selected from the VQA-v2 leaderboard[6], include MFB Yu et al. [2017], MFH Yu et al. [2018], BAN Kim et al. [2018], BUTD Anderson et al. [2018], MCAN Yu et al. [2019], Pythia Jiang et al. [2018], and ALBEF Li et al. [2021]. We assess all seven models on both of the datasets to investigate and understand the model progress across the machine understanding and accessibility datasets.

4.2.3.1 Datasets

As a representative of machine understanding VQA, we take the **VQA-v2** dataset Agrawal et al. [2017], which includes around 204,000 images from the COCO dataset Lin et al. [2014] with around one million questions. The images are collected through Flickr by amateur photographers. Thus the images are from sighted people rather than visually-impaired people. In addition, questions in VQA-v2 are collected in a post-hoc manner — given a image, sighted crowdworkers are asked to create potential questions that could be asked for the image. Finally, given the image-question pairs, a new set of annotators are asked to answer the questions based on the

[6] https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-test-dev

image information. For each image-question pair, ten annotations are collected as ground-truth.

As a representative of accessibility VQA, we take the **VizWiz** dataset Gurari et al. [2018], which includes around 32,000 images and question pairs from people with visual impairments. This dataset was built on data collected with a crowdsourcing-based app Bigham et al. [2010] where users with visual impairments asks questions by uploading an image with a recording of spoken question. The VizWiz dataset uses the image-question pairs from the data collected through the app and asks crowdworkers to annotate answers. Similarly, ten ground-truth answers are provided for each image-question pair. Note that in VizWiz each image-question pair is provided simultaneously by the same person, which is different from how the VQA-v2 dataset was curated.

Our evaluation also uses a smaller subset of VQA-v2's training set, which we call *VQA-v2-sm*, limited in size to match that of VizWiz's training set. This dataset is created to evaluate the effects of dataset size in VQA models' performance.

4.2.3.2 Evaluation Metric

We evaluate the seven models on the VQA-v2 and the VizWiz datasets with the standard "accuracy" evaluation metric for VQA. Since different annotators may provide different but valid answers, the metric does not penalize for the predicted answer not matching all the ground truth answers. For each question, given the ten ground-truth from human annotators, we compute the model answer accuracy as in Equation 4.1. If the model accurately predicts an answer that matches at least three ground-truth answers, it receives a maximal score of 1.0. Otherwise, the

accuracy score is the number of ground-truth answers matched, divided by three:

$$\text{accuracy} = \min\left\{1, \frac{\text{\# matches}}{3}\right\} \tag{4.1}$$

4.2.3.3 Models

All of the following models approach the problem as a classification task by aggregating possible answers from the training and validation dataset as the answer space.

MFB & MFH: The multi-modal factorized bilinear & multi-modal factorized high-order pooling models [Yu et al., 2017, 2018] are built upon the multi-modal factorized bilinear pooling that combines image features and text features as well as a co-attention module that jointly learns to generate attention maps from these multi-modal features. The MFB model is a simplified version of the MFH model.

BAN: The bilinear attention network model Kim et al. [2018] utilizes bilinear attention distributions to represent given vision-language information seamlessly. BAN considers bilinear interactions among two groups of input channels, while low-rank bilinear pooling extracts the joint representations for each pair of channels.

BUTD: The bottom-up and top-down attention model Anderson et al. [2018] goes beyond top-down attention mechanism and proposes the addition of a bottom-ups attention that finds image regions, each with an associated feature vector. Thus, creating a bottom-up and top-down approach that can calculate at the level of objects and other salient image regions.

MCAN: The modular co-attention network model Yu et al. [2019] follows the co-attention approach of the previously mentioned models, but cascades modular co-attention layers at depth,

to create an effective deep co-attention model where each MCA layer models the self-attention of questions and images.

Pythia: Pythia is an extension of the BUTD model, utilizing both data augmentation and ensembling to significantly improve VQA performance Jiang et al. [2018].

ALBEF: The align before fusing model Li et al. [2021] builds upon existing methods that employ a transformer-based multimodal encoder to jointly model visual tokens and word tokens, by aligning the image and text representations and fusing them through cross-model attention.

For all the models, the answer space of the VQA-v2 dataset is $3,129$, while the answer space of the VizWiz dataset is $7,371$, which is provided by Pythia Jiang et al. [2018].

Implementation details. We use three different code bases for our evaluation: OpenVQA[a], Pythia[b], and ALBEF[7]. On the OpenVQA platform, four VQA models—MFB, BAN, BUTD, and MCAN—are already implemented. Pythia supports both of the VQA-v2 and Vizwiz datasets, but the OpenVQA and ALBEF only support the VQA-v2 dataset. Thus, we implement the support of the VizWiz dataset on OpenVQA (i.e., for MFB, BAN, BUTD, and MCAN) and ALBEF. Their default hyperparameters are used to train models on VQA-v2 and VizWiz, respectively. For OpenVQA and ALBEF on which we implement the VizWiz support, the default hyperparameters for VQA-v2 are used to train models on VizWiz as well. We fix the default accuracy metric implemented in OpenVQA, which is silently incompatible with the VizWiz data format, consistently underscoring predictions.

[7] https://github.com/MILVLG/openvqa,
https://github.com/allenai/pythia,
https://github.com/salesforce/ALBEF.

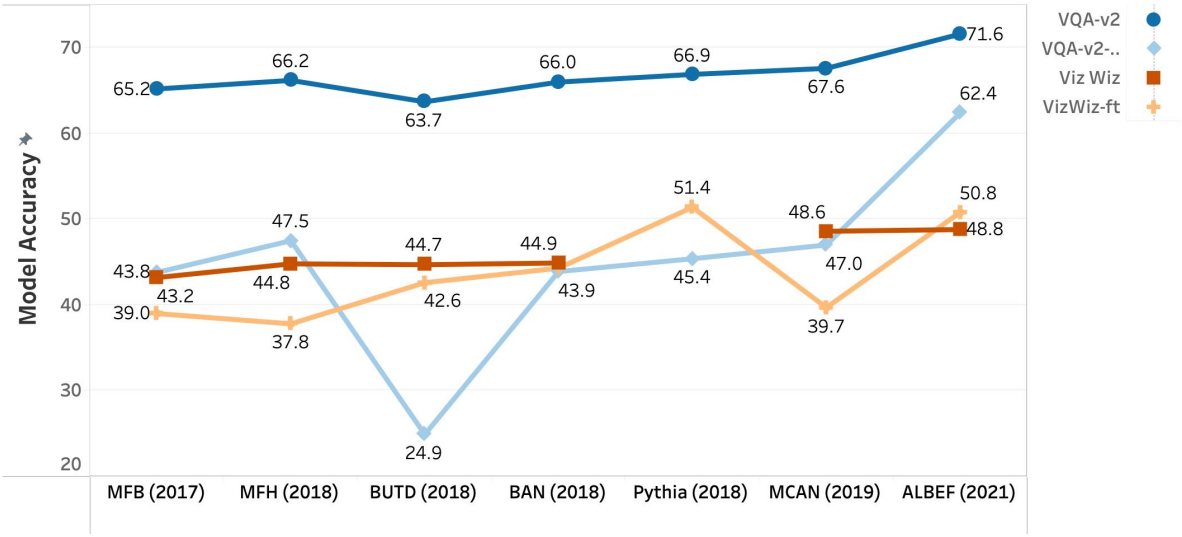

Figure 4.6: Model accuracy on the VQA-v2 dataset and the VizWiz dataset. The models are ordered by the time they were proposed.

4.2.4 Findings

Our objective in this section is to investigate challenges of the VQA task on two different datasets. We assess the performance progress of VQA models and delve into errors. Then, we discuss research directions that future work could take.

4.2.4.1 Model Performance Progress

First, we examine whether the progress of VQA model architectures on the machine understanding dataset (VQA-v2) also apply to the accessibility dataset (VizWiz). For VizWiz, we report testing results on both trained from scratch with VizWiz (*VizWiz*) and trained on VQA-v2 and finetuned with VizWiz (*VizWiz-ft*). As mentioned in Section 4.2.3.1, we randomly sampled the same number of datapoints from the train set of VQA-v2 as that in VizWiz to form *VQA-v2-sm* to understand the effect of dataset size in the VQA performance.

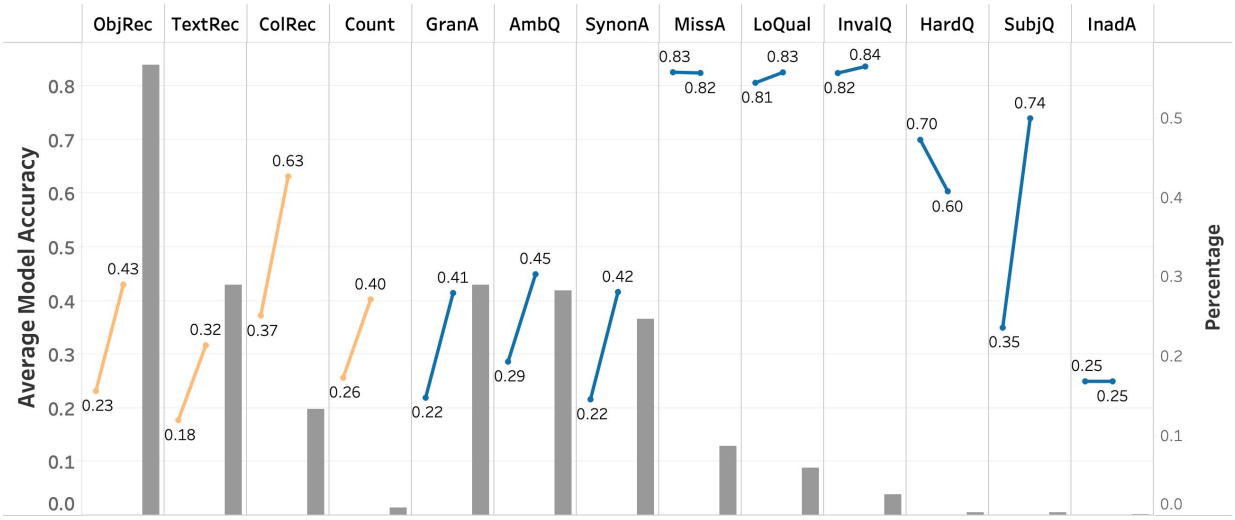

Figure 4.7: The model accuracy progress from MFB (left point) to ALBEF (right point) on each challenge class in Zeng et al. [2020] (orange) and Bhattacharya et al. [2019] (blue) for the VizWiz dataset. The grey bar represents the percentage of data points labeled with such challenge class.

The results are shown in Figure 4.6.[8] Overall, we observe that along with the advancement of model structures based on the VQA-v2 dataset, the model accuracy also improves on the VizWiz dataset. We observe that, from 2018 through 2021, performance on VQA-v2 improved 10% (from 65.2% to 71.6% accuracy), resulting in a similar improvement of 11% (43.8% to 48.8%) on VizWiz without fine-tuning and 30% (39% to 50.8%) on VizWiz with fine-tuning. The models fine-tuned from VQA-2 to VizWiz (i.e., VizWiz-ft) do not significantly outperform those trained on VizWiz from scratch. Gurari et al. [2018] also reported a similar pattern but pointed out the gap between model performance and human performance. These results show that improvements on VQA-v2 *have* translated into improvements on VizWiz, whereas the performance gap between the two datasets are still significant.

However, when controlling for dataset size, we see an improvement of 42% (43.8% to

[8]We do not include results of Pythia trained from scratch on VizWiz because their code expected to train VizWiz from a VQAv2.0 checkpoint, not from scratch.

	Label	%	Definition
Vision	ObjRec	56.7%	object recognition
	TextRec	29.0%	text recognition
	ColRec	13.3%	color recognition
	Count	1.0%	counting
Image-Question	GranA	29.0%	answers at different granularities
	AmbQ	28.3%	ambiguous question w/ > 1 valid answer
	SynonA	24.7%	different wordings of same answer
	MissA	8.7%	answer not present given image
	LoQual	6.0%	low quality image
	InvalQ	2.7%	invalid question
	HardQ	0.4%	hard question requiring expertise
	SubjQ	0.4%	subjective question
	InadA	0.0%	inadequate answers

Table 4.1: VQA challenge taxonomies and the number of data examples in the VizWiz dataset labeled with each challenge class.

62.4%) on VQA-v2-sm, where the training data is capped at the size of VizWiz, a substantially larger improvement than the 11% seen on VizWiz (the result on VizWiz with fine-tuning is not comparable here, because it is fine-tuned from the full VQA-v2 dataset). This appears to demonstrate a significant "overfitting" effect, as both VQA-v2-sm and VizWiz start at almost exactly the same accuracy (43.8% and 43.2%) but performance on VQA-v2-sm improves significantly more than on VizWiz.

4.2.4.2 Error Analysis

We perform both quantitative and qualitative error analysis to better understand which types of data will be useful to improve accessibility VQA for future dataset collection and model improvement. In this section we discuss the overall patterns found for models evaluated on VizWiz and what type of questions specifically, these model fail on.

Figure 4.8: The performance improvements from MFB to ALBEF on the VizWiz dataset with respect to the reduced number of data examples with 0 accuracy score. Red color represents the number of data examples ALBEF got 0 score on while red plus blue color represents the number of data examples MFB got 0 score on.

VQA Challenge Datasets In our first set of experiments, we aim to better understand precisely what that models have improved on between 2017 and 2021 that has led to an overall accuracy improvement on VizWiz-ft from 39.0% (MFB) to 50.8% (ALBEF). To do this, we make use of two meta-data annotations of a subset of the VizWiz validation dataset ($3,143$ data examples): one labels each example with the *vision skills* required to answer that question Zeng et al. [2020], the second labels each with aspects of the *image-question* combination that are challenging Bhattacharya et al. [2019]. Both of these papers investigate the challenges for *annotators*; here, we use these annotations to evaluate models. Table 4.1 shows the taxonomies and the percent of VizWiz validation examples that are labeled with the challenge class according to majority vote over five annotations.

Given this taxonomy, we assess the performance progress between MFB and ALBEF in the VizWiz-ft setting across each VQA challenge class. The results are reported in Figure 4.7. In comparing to MFB, ALBEF improves on every class of challenges except HardQ—hard questions that may require domain expertise, special skills, or too much effort to answer—though HardQ is also one of the rarest categories. (It is somewhat surprising the high performance of the

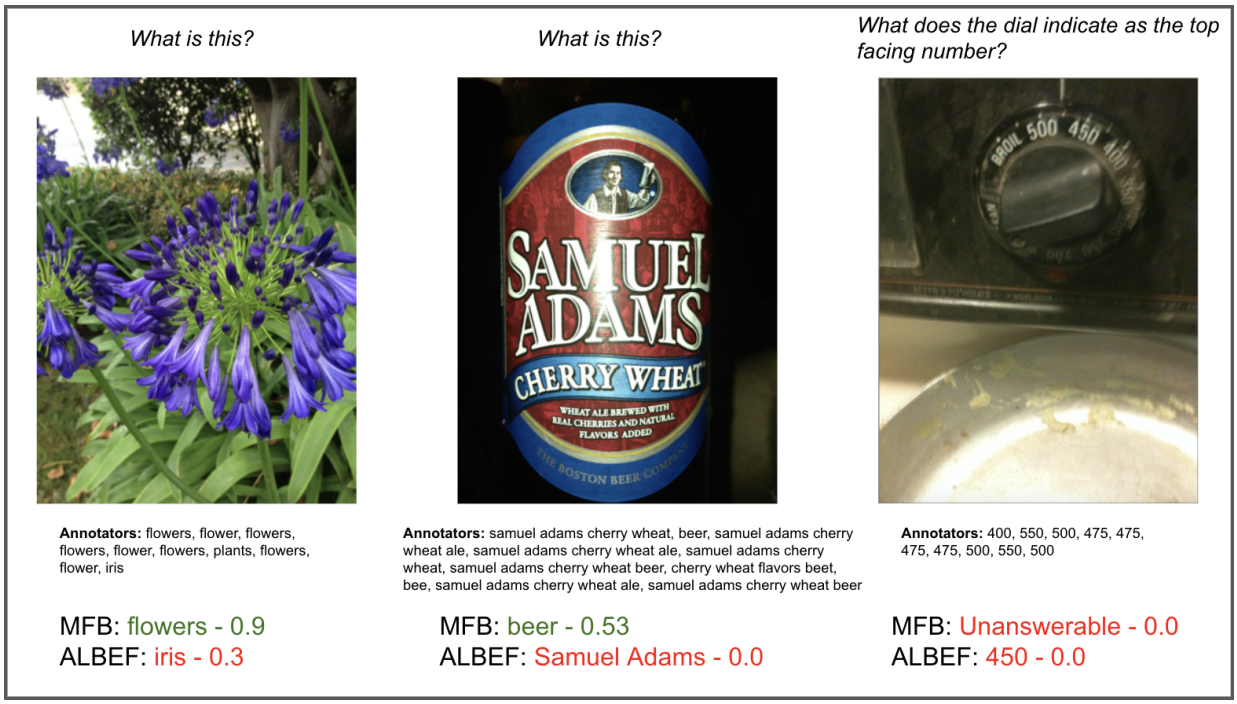

Figure 4.9: Examples of low-performance ALBEF image/question pairs, that should be correct, together with the accuracy scores. (Left) ALBEF gives a more detailed answer of *Iris*[9], but since most annotators put *flowers*, performance score is low. (Middle) ALBEF correctly names the beer, but once again does not match the annotators, so MFB appears to perform better. (Right) ALBEF gives a number answer that is close to correct (and which is not much different from the set of ground truth answers), where MFB does not make an attempt.

models on these "hard" questions.) We observe that among the *vision skill* challenge classes, the models struggles the most on recognizing texts. Among the *image-question* challenges, models have low accuracy on almost all the challenge classes related to the answers — ground-truth answers with different granularities, wordings, and inadequate answers. This indicates a potential problem in evaluating models on the VizWiz dataset, which is further explored in our qualitative analysis in paragraph 4.2.4.2. For the questions, models struggle the most with handling ambiguous or subjective questions, which we will discuss more in the next section. Overall, the results point out the challenges that models have most difficulty on, which we hope can bring insights for future work to improve accessibility VQA systems.

Where the Models Fail To further understand the data examples that the models fine-tuned on VizWiz perform poorly on, we manually investigate the validation examples on which models achieve 0% accuracy: matching none of the ten human-provided answers. We measure how many data examples that have 0% accuracy on MFB got improved by the ALBEF model for each challenge class, shown in Figure 4.8. We see that model improvement is greatest on color recognition (65%) and low quality images (63%). On the other hand, object recognition, text recognition, color recognition, and ambiguous questions are the challenge classes which a current state-of-the-art model has the most difficulty.

When taking a closer look at the individual examples that ALBEF has 0% accuracy on, it turns out the issue is often with the *evaluation measure* and not with the ALBEF model itself. The most frequent issues are:

Answerable Questions Marked Unanswerable. The biggest difference (and what we deem an improvement) between ALBEF and MFB has to do with "unanswerable" questions. 27% of the questions in the validation data are deemed "unanswerable" by at least three annotators—making "unanswerable" a prediction that would achieve perfect accuracy. For 56% of the questions that were not of type *"unanswerable"*, MFB still answered *"unanswerable"*, while ALBEF did this only 30% of the time. This skew helps MFB on the evaluation metric, but ALBEF's answers for many of these questions are at least as good—and therefore useful to a user—as saying "unanswerable." For example, the *number* question type, MFB only answered with a number 2.2% of the time, whereas ALBEF answered with a number 56% of the time and, in those cases, the answers are often very close to the correct answer (see Figure 4.9).

[9] As identified by my committee members, this flower is not iris, it may be Lily of the Nile.

Overly Generic Ground Truth. It is often the case that ALBEF provides a correct answer that is simply more specific than that provided by the ground truth annotation. For example, a common question in VizWiz is *"What is this?"*. When comparing ALBEF and MFB models, by accuracy alone, ALBEF outperforms MFB in 28.8% of such cases, and MFB outperforms ALBEF in 12.6%. However, in the majority of these examples, ALBEF gives a correct, but more detailed response than the ground truth, thus earning it 0% accuracy. So while, based on the annotation, ALBEF is wrong, the model is actually correctly answering the question and performs worse than the MFB model only 2.6% of the time. Furthermore, we found that both MFB and ALBEF models are both challenged by *yes/no* question types, but that these questions were often subjective of ambiguous.

Annotator Disagreement. Questions such as *"Is this cat cute?"* or *"Are these bad for me?"* arguably make for poor questions when evaluating model performance: highly subjective yes/no questions often have annotations where at least three annotators state *"yes"*, and at least three state *"no"*. Therefore, per the evaluation metric, either answer achieves an accuracy of 1. For example, for the question *"Do these socks match?"* ALBEF had an accuracy score of 60% for an answer of *no* and MFB had an accuracy score of 83% for an answer of *yes*, even though either is arguably correct.

4.2.5 Discussion and Limitations

We have shown that, overall, performance improvements on machine "understanding" VQA *have* translated into performance improvements on the real-world task of accessibility VQA. However, we have also shown evidence that there may be a significant overfitting ef-

fect, where significant model improvements on machine "understanding" VQA translate only into modest improvements in accessibility VQA. This suggests that if the research community continues to only hill-climb on challenge datasets like VQA-v2, we run the risk of ceasing to make any process on a pressing human-centered application of this technology, and, in the worst case, could degrade performance.

We have also shown that along with the overall model improvement, the accessibility VQA system have performance growth on almost all of the challenge classes though some challenges remain difficult. In general, we observe the models struggle most on questions that require text recognition skill as well as ambiguous questions. Future work thus may wish to pay more attention on these questions in both data collection and model design.

Finally, we have seen that we are likely reaching the limit of the usefulness of the standard VQA accuracy metric, and that more research is needed to develop automated evaluation protocols that are robust and accurately capture performance improvements. On top of this, VQA systems are reaching impressive levels of performance, suggesting that human evaluation of their performance in ecologically valid settings is becoming increasingly possible. As ecological validity would require conducting such an evaluation with blind or low-vision users, research is needed to ensure that such evaluation paradigms are conducted ethically and minimize potential harms to system users.

Limitation First, while we analyzed many models across several years of VQA research, our analysis is limited to two datasets. Moreover, as will be discussed in paragraph 4.2.4.2, the "ground truth" in these datasets, especially when combined with the standard evaluation metric, is not always reliable. Second, our analysis is limited to English, and may not generalize directly

to other languages. Finally, blind and low-vision users are not a monolithic group, and the photos taken and questions asked in the VizWiz dataset are representative only of those who used the mobile app, likely a small, unrepresentative subset of the population.

4.3 Conclusion

In addressing the content moderation task, our study has highlighted gaps between the capabilities of existing NLP models and the specific requirements needed to comply with Reddit's moderation policies. Volunteer content moderators have shown considerable enthusiasm and adaptability toward utilizing moderation tools. They can accommodate imperfections in these tools, provided the tools are transparent of their limitations, enabling moderators to make informed adjustments. Moderators have also offered valuable insights in directing future enhancements of the models. Moderators highlighted the need for tools to address complex moderation rules that require additional read time or specialized domain knowledge. Furthermore, their needs varied in terms of models' precision and recall for different moderation rules, with a preference for handling more subjective rules on their own.

Regarding the visual question answering (VQA) task, we have also observed a significant gap in adapting existing model architectures, primarily designed for machine "understanding" of VQA, to the accessibility VQA dataset. Alarmingly, this gap appears to be widening, suggesting that incremental improvements on the challenge datasets might not effectively translate to real-world applications of this technology. Additionally, we have identified shortcomings in the standard VQA evaluation metrics, which fail to align with user preferences, potentially leading to misguided directions in model improvement.

In both applications, we have identified significant gaps between the capabilities of current AI systems and the unique requirements of these users. We have also demonstrated that specified users can bring practical insights for system improvements as they are the domain experts of their own experiences. These findings emphasize the need for developers to actively collaborate with target users throughout the design, development, and evaluation processes to create AI solutions that effectively address their specific challenges and preferences.

Chapter 5: Alignment with Diverse Stakeholders' Values in General Large Language Models

In the previous chapter, we explored the importance of aligning AI systems with the needs of specific user groups, highlighting the value of user-centered design and stakeholder engagement. This chapter addresses the critical issue of aligning AI systems with the values of diverse stakeholders. We focus on the challenge of stereotyping in large language models, which are increasingly being deployed across a wide range of applications and have the potential to impact users from all walks of life. To systematically evaluate stereotypical associations within these models, we propose a theory-grounded methodology that considers the unique experiences and perspectives of different social groups, including those with intersectional identities. We then extend our analysis to investigate stereotype leakage across languages in multilingual models, highlighting the risk of harmful stereotypes being transferred across cultural boundaries. By examining the prevalence and nature of stereotypes in these widely-used AI systems, we aim to contribute to the development of more inclusive and equitable technologies that respect and uphold the values of diverse stakeholders.

5.1 Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models

Joint work with Anna Sotnikova[1], Rachel Rudinger, Linda Zou, Hal Daumé III. Appeared at NAACL 2022

5.1.1 Overview

Stereotypes are abstract and over-generalized pictures in people's minds that capture attributes about groups of people in the complex social world [Lippmann, 1965]. They influence people's thoughts and behaviors, and allow people to make predictions beyond their personal experience or information given [Bruner et al., 1957, Wheeler and Petty, 2001]. Stereotypes are also entwined with the production of prejudice, discrimination, and in-group favoritism [Stangor, 2014, Jackson, 2011]. A long line of research in social psychology has established models of generic dimensions that estimate people's stereotypes of social groups [Koch et al., 2016, Fiske et al., 2002, i.a.]. We build on the Agency Beliefs Communion (ABC) model, which measures stereotypes toward a social group with respect to 16 traits in three dimensions: Agency/ Socioeconomic Success, Conservative–Progressive Beliefs, and Communion (subsection 5.1.2); an analysis of the group "man" across 32 traits (16 opposing dyads) is shown in Figure 5.1.

Pre-trained language models (LMs) encode correlations between social groups and traits, like associating the group "Muslim" with the trait `threatening`, or "man" with `confident` [e.g., Bender et al., 2021, Nozza et al., 2021, Hovy and Yang, 2021]. We conduct a system-

[1]Equal contribution.

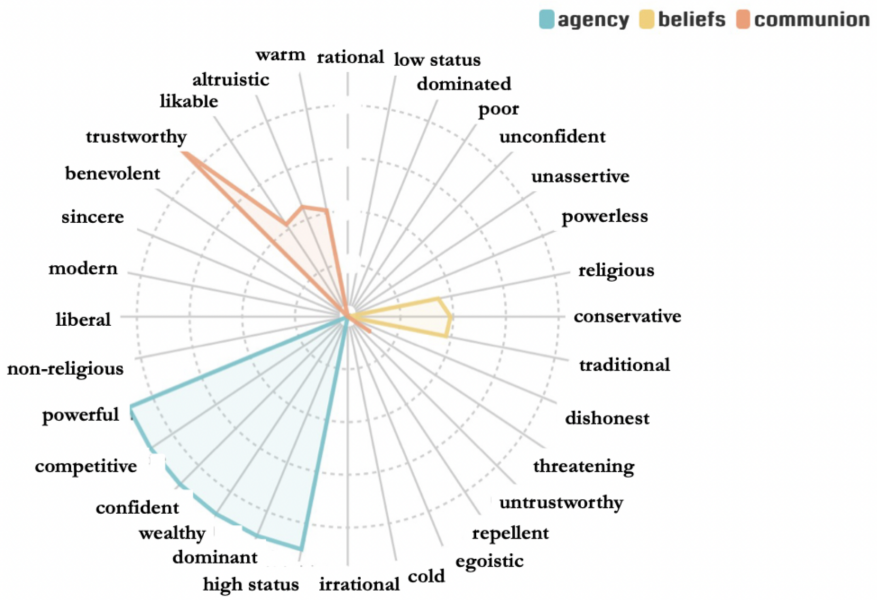

Figure 5.1: Crowdsourced analysis of the social group "man" under the ABC model [Koch et al., 2016].

atic study of social stereotypes in contextualized English masked LMs, grounded in group-trait associations from the ABC model. To capture the group-trait associations in the LM, we first assess two previously proposed word association tests and also propose a new measurement: the sensitivity test (SeT) (subsection 5.1.3).

To evaluate the degree to which two LMs—BERT [Devlin et al., 2019] and RoBERTa [Liu et al., 2019b]—align with human stereotype judgments, we design a human study for collecting group-trait judgments (subsection 5.1.5). We show that our measure, SeT, best aligns with human judgements on group-trait associations and find that, in general, the association from language models have moderate alignment with human judgements.

Finally, with the best-aligned association measurement, we extend the ABC approach to study LM stereotypes on intersectional groups (subsubsection 5.1.6.2). Due largely to the difficulty of extending current approaches for measuring stereotypes in LMs to large numbers of

Agency		Beliefs		Communion	
powerless	powerful	religious	science-oriented	untrustworthy	trustworthy
low status	high status	conventional	alternative	dishonest	sincere
dominated	dominating	conservative	liberal	cold	warm
poor	wealthy	traditional	modern	benevolent	threatening
unconfident	confident			repellent	likable
unassertive	competitive			egotistic	altruistic

Table 5.1: List of stereotype dimensions and corresponding traits in the ABC model [Koch et al., 2016].

groups, most current approaches only study isolated groups, despite the fact that people's social identities are multifaceted [Ghavami and Peplau, 2013]. Because our approach is generalizable to unstudied groups, we take a step towards exploring stereotypes of intersectional identities, finding some correspondence between model behavior and the literature on intersectional stereotypes.

5.1.2 Related Work

Ours is far from the first work to assess stereotypes in language models, and has both advantages and disadvantages compared to previous approaches (see Table 5.2). Past work has generally taken one of two approaches. The first approach tests systems with hand-constructed templates like "The `[group]` is □", where `[group]` ranges over social groups (e.g., "woman" or "Hispanic"), and □ represents a "masked word" and ranges over occupations ("a professor" or "a nurse") [e.g., Bolukbasi et al., 2016b, May et al., 2019] or associations drawn from implicit association tests (IAT) (e.g., pleasant/unpleasant words or career/family-related words) [e.g., Caliskan et al., 2017b, Guo and Caliskan, 2021]. In Table 5.2 we refer to these as "unnatural" prompts. The second approach collects more natural sentences containing stereotypes, either by web crawling with crowdworkers annotations for social bias [Sap et al., 2019] or by having crowdworkers directly write stereotyping sentences [Nangia et al., 2020, Nadeem et al., 2020].

We take the first approach with traits from the ABC model, using prompts. The advan-

Measurement	Generalizes	Grounded	Exhaustive	Natural	Specificity
Debiasing (Bolukbasi et al.)	✓				✓
CrowS-Pairs (Nangia et al.)			✓	✓	✓
Stereoset (Nadeem et al.)			✓	✓	✓
S. Bias Frames (Sap et al.)			✓	✓✓	✓
CEAT (Guo and Caliskan)	✓	✓		✓✓	
This Work	✓	✓	✓✓		

Table 5.2: Comparison with previous work: Generalizes denotes approaches that naturally extend to previously unconsidered groups; Grounded approaches are those that are grounded in social science theory; Exhaustiveness refers to how well the traits cover the space of possible stereotypes; Naturalness is the degree to which the text input to the LM is natural (we consider naturally occurring web scraped data as "very natural" and crowdsourced sentences as "somewhat natural."). Specificity indicates whether the stereotype is specific or abstract.

tage of this approach is that the templates and the traits are completely controlled and are easy to extend to other social groups. The second approach is harder to control, which also leads to significant annotation challenges [Blodgett et al., 2021]. Using natural sentences limits generalizability, as it requires a unique collection of prompts (and embedded traits) for each social group; in contrast, the prompt-based approach easily generalizes to any plausible group, especially when based on a theoretically grounded framework like ABC or IAT.

An advantage of our approach is that the ABC traits are more exhaustive in stereotype coverage with verification from social psychological experiments. The ABC model covers three dimensions with 16 traits, which are consensual, spontaneous, and have been tested using expansive range of social groups [Koch et al., 2021]. They used a carefully designed data-driven approach to gather people's fundamental dimensions of social perceptions with as little sampling bias as possible. Thus the resulted 16 traits cover most stereotypes.

Nevertheless, the main trade-off of our approach is that the testing data are not as natural and specific as other approaches. Although we carefully pick and adjust the templates and the form of the social group terms so that the testing sentences are grammatically correct, they are

Domain	Groups
Gender/ sexuality	*man*, *woman*, *non-binary*, *trans*, *cis*, gay, lesbian
Race/ ethnicity	*Black*, *White*, *Hispanic*, *Asian*, *Native American*
Religion	*Jewish*, *Muslim*, *Christian*, *Buddhist*, *Mormon*, Catholic, Amish, Protestant, Atheist, Hindu
Socio-economic	*wealthy*, *working class*, *immigrant*, *veteran*, *unemployed*, refugee, doctor, mechanic
Age	*teenager*, *elderly*
Disability status	*blind*, *autistic*, *neurodivergent*, Deaf, person with a disability
Politics	Democrat, Republican
Nationality	Mexican, Chinese, Russian, Indian, Irish, Cuban, Italian, Japanese, German, French, British, Jamaican, American, Filipino

Table 5.3: Social groups domains and corresponding social groups used for the model experiments and human experiments. Single groups for human experiments are highlighted with italic font style.

likely not representative of sentences seen in the real world or in the training data of the language models. Further, while our approach has the benefit of near-exhaustive coverage of potential stereotypes, this comes at a cost: the traits we consider are much more high level (e.g., "repellent") than more fine-grained stereotypes collected by other means (e.g., the angry Black woman stereotype Collins [2002])—this approach therefore trades coverage for specificity.

5.1.3 Measuring Stereotypes in LMs

Our goal is to measure stereotypes in (masked) LMs, and compare them to stereotypes elicited from people. [1] In subsection 5.1.5 we describe our approach for eliciting human judgments of group-trait affinities; here we describe how we measure these in LMs. Previous work has proposed various ways to measure word associations in LMs, including increased log probability score (ILPS) and contextualized embedding association test (CEAT), both of which we have summarized in subsection 2.4.2. We also present a new measurement which we call the Sensitivity Test (SeT), which adapts concepts from active learning to the task of measuring a

[1] Both the code and the dataset, along with a datasheet [Gebru et al., 2018b], are available under a MIT licence at: https://github.com/TristaCao/U.S_Stereotypes.

LM's associations.

5.1.4 Measurements of Word Associations

Singular The/That/A `[group]` is □.	**Plural** Most/Many/All `[group]` are □. / `[Group]` are □.
Declarative `[Group]` are □.	**Interrogative** Why are `[group]` □?
Non-adverbial `[Group]` are □.	**Adverbial** `[Group]` are very/so/mostly □.
Fact `[Group]` are □.	**Belief** I/We/Everyone/People believe/expect/think/know(s) that `[group]` are □.
Fact `[Group]` are □.	**Social Expectation** `[Group]` are supposed to be/should be/are seen as/ought to be/are expected to be □.
Group-first `[Group]` are □.	**Trait-first** The □ people are `[group]`.
Non-comparative `[Group]` are □.	**Comparative** `[Group]` are more likely to be □ than others.

Table 5.4: Template Variations.

Sensitivity Test (SeT) is a new approach we propose to measure word association for social bias in language models, inspired by ideas from active learning Beygelzimer et al. [2008]. The intuition of SeT is that even though a model assigns the same probability to two different words, the robustness of those two probabilities may be different. For example, both $p(\texttt{kind}|\textit{“Men are □”})$ and $p(\texttt{competent}|\textit{“Blind people are □.”})$ might be low. However, the language model may well not have seen many examples of blind people, as opposed to the presumably very large number of examples of men. In this case, a small number of examples may be sufficient to alter the model's predictions about blind people, while a larger number would be required for men. SeT captures the model's confidence in a prediction by measuring how much the model weights would have to change in order to change that prediction. Specifically, SeT computes the minimal change to the last layer of the language model so that a given trait becomes the highest probability

trait (over the full vocabulary).

For example, consider the template "The `[group]` is □." with the group "woman" and the trait `incompetent`. Let $\boldsymbol{\ell}$ be the logits at □ when the input is "The woman is □.", and let `t` be the index of `incompetent` in $\boldsymbol{\ell}$ (so that $\ell_{\texttt{t}} = p(\texttt{incompetent} \mid \text{context})$). Let $\mathbf{h}$ be the last hidden layer before the logits, and let $\mathbf{A}$ be the matrix of the last linear layer so that $\boldsymbol{\ell} = \mathbf{A}\mathbf{h}$. SeT computes the minimal distance between $\mathbf{A}$ and some other matrix $\mathbf{A}'$ so that `t` is the top word among the new logits $\boldsymbol{\ell}' = \mathbf{A}'\mathbf{h}$. Formally:

$$\text{SeT}(\texttt{g}, \texttt{t}) = \log \frac{\Delta(\mathbf{A}, \mathbf{h}_{\texttt{g}}, \texttt{t})}{\Delta(\mathbf{A}, \mathbf{h}_{\square}, \texttt{t})} \tag{5.1}$$

where $\mathbf{h}_{\texttt{g}}$ is the penultimate layer on input `g`

$\mathbf{A}$ is the matrix before the logits

$$\Delta(\mathbf{A}, \mathbf{h}, \texttt{t}) = \min_{\mathbf{A}'} \mid \mathbf{A}' - \mathbf{A} \mid_2^2 \tag{5.2}$$

$$\text{s.t. } (\mathbf{A}'\mathbf{h})_{\texttt{t}} \geq (\mathbf{A}'\mathbf{h})_{\texttt{t}'} + \gamma, \forall \texttt{t}' \neq \texttt{t}$$

for a fixed margin $\gamma > 0$, which we set to 1. SeT returns the *negative distance* as measure of the association between the corresponding group and trait, normalized by a prior akin to ILPS. This optimization problem does not (to our knowledge) admit a closed form solution; we solve it iteratively using the column squishing algorithm [Bittorf et al., 2012, Daumé and Kumar, 2017].

5.1.4.1 Implementation details

We test all three measurements on both BERT and RoBERTa pretrained large models from an open-source HuggingFace[2] library.

Social groups. Table 5.3 lists all the individual social groups we cover in this work. We manually construct the list by combining and picking groups from the list of social groups from Sotnikova et al. [2021a] and Koch et al. [2016] and also adding social groups we think are stereotyped in U.S. culture.

Traits. We use the 32 adjectives of the 16 traits from the ABC model (Table 5.1). For each traits, we calculate the score of its left-side adjective from its right-side adjective: $S_{\texttt{powerless-powerful}}(\texttt{g}) = S(\texttt{g}, \texttt{powerful}) - S(\texttt{g}, \texttt{powerless})$, where S is one of the scores from subsection 5.1.4.[3]

Templates. ILPS and SeT both require templates in calculating scores. We thus carefully construct a list of templates (Table 5.4) that covers multiple grammatical and semantic variations, inspired by work investigating harmful search automatic suggestions [Hazen et al., 2020]. We find that different model structure requires different templates in order to bring up stereotypes that correlate with human data. See subsection 5.1.6 for evidence.

Subwords. Due to the nature of BERT and RoBERTa's tokenizers, some of the adjectives are divided into multiple subwords. This is problematic because all the measurements compute their

[2] https://huggingface.co/models

[3] In preliminary experiments, when calculating the score for each adjective, we considered including 1-3 additional adjectives by averaging their scores to improve robustness and mitigate ambiguity. The full list is in Appendix Table C.1. However, we found that this did not improve correlations, so we reverted to using the 32 adjectives from the ABC model.

scores at token level. Neither ILPS nor CEAT deals with subwords directly: in their released implementations, they either take the first or the last sub-token of the word. To remedy this, we adjust the ILPS measurement (denoted as ILPS*) to properly compute the probability of traits in context using the chain rule across subwords. For SeT, we calculate the sensitivity score for each subword individually and take the maximum SeT score as the SeT score for the word, which effectively computes a *lower-bound* on how much the model parameters would need to change. We did not modify CEAT's measurement as it is not clear what is the best way to compute comparable word embeddings for words that consist of multiple subwords.

5.1.5 Human Study

In the previous section, we describe how we compute associations between groups and traits in language models.[4] In this section, we assess stereotypes of social groups through groups-trait association, like in Figure 5.1. We adopt this approach because it is widely used to evaluate group stereotypes in social psychology field [Fiske et al., 2002, Koch et al., 2016]. It also aligns with Lippmann [1965]'s theory of stereotypes that they are abstract pictures in people's head. We broadly follow procedures from previous social psychology papers to collect human evaluation on social groups.

Survey Design. We recruit participants from Prolific[5]. Each participant is paid $\$2.00$ to rate 5 social groups on 16 pairs of traits and on average participants spend about 10 minutes on the survey. This results in a pay of $\$12.00$ per hour. Maryland's current minimum wage is $\$12.20$

[4] Approved by our institutional IRB, #1724519-1.
[5] https://www.prolific.co/

[6]. First, participants read the consent form, and if they agree to participate in the study, they see the survey's instructions. For each social group, participants read "As viewed by American society, (while my own opinions may differ), how `[e.g., powerless, dominant, poor]` versus `[e.g., powerful, dominated, wealthy]` are `<group>`?" They then rate each trait with a 0-100 slider scale where two sides are the two dimensions of the trait (e.g. `powerless` and `powerful`). Each annotated group is shown on a separate page, and participants cannot go back to previous pages. To avoid social-desirability bias, we explicitly write in the instruction that *"we are not interested in your personal beliefs, but rather how you think people in America view these groups."*

Participant Demographics. At the end of the survey we collect participants' demographic information, including gender, race, age, education level, type of living area, etc. Our participants represent 26 states, with 63.3% from California, New York, Texas, or Florida; the gender breakdown is 48.2% male, 49.6% female, and 2.2% genderqueer, agender, or questioning; and skew young, with over 96% at most 40 years old; and with racial demographics that approximately match the U.S. census. For more details on demographics, see subsection C.3.2.

Quality Assurance. Ensuring annotation quality in a highly subjective task is a challenge, and common approaches in NLP like having questions where we "know" the answer as tests, measuring interannotator agreement, and calibrating reviewers against each other [Paun et al., 2018] do not make sense here. Yet, it is still important to ensure the annotation quality. After much iteration, we include three test questions, and warn the participants at the beginning that there are test questions.

[6] https://www.minimum-wage.org/maryland

1. After the first group, participants must name the group they just scored.
2. After the second, participants must list one trait they just marked high and one marked low.
3. The fifth (final) group is a repetition of one of the four groups they previously scored.

We discard annotations with incorrect answers to either of the first two questions. For the third test, we compute intra-annotator (self) agreement and discard annotations with accuracy-to-self lower than 80%. For each group we collect 20 annotations that pass our quality threshold. In total, we collected annotations from 247 participants, with 133 passing the quality tests (suggesting that having such tests is important). The 114 annotations that did not pass tests were excluded from our dataset, but all 247 participants were paid.

Social groups and traits. The social groups we used for the human study are highlighted in Table 5.3. This table contains only single groups used for the model subsection 5.1.3 and human experiments. We collect annotations for 25 social groups within 5 domains, across all 16 pairs of traits.

5.1.6 Results

In this section we present results on correlations between human and model stereotypes for individual groups, comparing across different measurements, including our proposed measurement, SeT (subsubsection 5.1.6.1). Next, we analyze how model scores change for intersectional social groups. We consider several possible factors that may influence the score changes such as identity order, some domain domination, and consider emergent traits (subsubsection 5.1.6.2).

Measure	RoBERTa τ	RoBERTa Template(s)	BERT τ	BERT Template(s)
ILPS	0.280	That `[group]` is `[trait]`.	0.215	All `[group]` are `[trait]`. `[Group]` should be `[trait]`.
ILPS*	0.258	All `[group]` are `[trait]`. That `[group]` is `[trait]`.	0.123	We expect that `[group]` are `[trait]`. `[Group]` should be `[trait]`.
SeT	0.253	That `[group]` is `[trait]`.	0.214	All `[group]` are `[trait]`. `[Group]` should be `[trait]`.

Table 5.5: Best two templates for each measurement-model pair and corresponding correlations. Some have only one template because there is no combination of two templates that give higher correlation score than this one template.

	CEAT		ILPS		ILPS*		SeT	
	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT
Kendall's τ	0.019	0.111†	0.169†	0.094†	0.175†	0.015	**0.199**†	0.116
Precision at 3	0.500	0.587	0.620	0.533	**0.653**	0.560	**0.653**	0.613

Table 5.6: Overall alignment scores with human annotations. The highest scores are bold for each row. For correlation scores, we mark scores where the p-value is < 0.05 with †.

5.1.6.1 Correlation on Individual Groups

Before we answer the question of how language model stereotype scores align with human stereotypes across the measurements introduced in subsection 5.1.3, we first run a pilot experiment to select the best template(s) for each measurement-model pair from the set of templates in Table 5.4 (except for CEAT, which does not require templates). We randomly picked four social groups (Asian, Black, Hispanic, immigrant) and five annotations from each group for the pilot. Since our goal is to inspect the alignment between human and model stereotypes, we take the averaged score of the five annotations as "ground truth" and select templates that give the correlation score according to Kendall τ. We limit the selection to at most two templates to avoid overfitting on the pilot data, selected to maximize correlation for each measurement-model pair.

The selected templates and corresponding correlation scores are shown in appendix (Ta-

ble 5.5); the score range for weak correlation is 0.10 - 0.19, moderate 0.20 - 0.29, and strong 0.30 and above Botsch [2011]. For a fixed LM, the best templates tend to be similar across all measures: RoBERTa tends to achieve highest correlation with templates like "That `[group]` is `[trait]`." while for BERT the preferred templates tend to be "All `[group]` are `[trait]`." or "`[Group]` should be `[trait]`."

Given the best templates for each measurement-model pair, we measure to what degree language model stereotypes are aligned with human stereotypes with all annotations on 25 social groups. To quantify alignment, we both calculate the Kendall rank correlation coefficient (Kendall's τ) and the Precision at 3 (P@3). The former indicates the correlation between model and human scores on group-trait associations in terms of the number of swaps required to get the same order. The latter indicates the percentage of the model's top stereotypes which accord with human's judgements. For P@3, we also calculate at both the group level and overall with all groups. For each group, we compute its P@3 score by taking the average of the P@3 scores with the top 3 traits (top at one polarity) and the score with the bottom 3 (top at the other polarity) because each trait has two polar adjectives and the group-trait score is calculated with the difference of the two polarities. To calculate the P@3 scores, we binarize the human group-trait scores at a threshold of 50. The overall P@3 score is the average of the groups' individual P@3 scores.

The overall scores are in Table 5.6. We see that in general that RoBERTa contains group-trait associations that are more similar to human judgements than does BERT. Additionally, we see that both ILPS* and SeT have higher P@3 scores than CEAT and ILPS. The RoBERTa model with the SeT measurement approach yields outputs are the most aligned with human's judgements, with RoBERTa/ILPS* a close second. From its scores, we see that model's group-trait

associations have moderate correlation with human's judgements. Moreover, in general, two out of the three top ranked group-trait associations from the model agree with human data. See appendix Table C.13 for the overall scores of test groups only, where the four pilot groups are excluded, and section C.2 for group level alignment scores.

5.1.6.2 Intersectional Groups in LMs

Background. Intersectionality is a core concept in Black feminism, introduced in the Combahee River Collective Statement in 1977 [1977, 1983], considering the ways in which feminist theory and antiracism need to combine: "Because the intersectional experience is greater than the sum of racism and sexism, any analysis that does not take intersectionality into account cannot sufficiently address the particular manner in which Black women are subordinated." The concept was applied in law by Crenshaw [1989] to analyze the ways in which U.S. antidiscrimination law fails Black women.

The concept of intersectionality has broadened and, while its boundaries remain contested [e.g., Browne and Misra, 2003], there are a number of core principles that are central [Steinbugler et al., 2006, Zinn and Dill, 1996]: (1) social categories and hierarchies are historically contingent, (2) the experience at an intersection is more than the sum of its parts Collins [2002], King [1988], (3) intersections create both oppression and opportunity Bonilla-Silva [1997], (4) individuals may experience both advantage and disadvantage as a result of intersectionality, and (5) these hierarchies impact social structure and social interaction.

Goals and Research Questions. We aim to understand whether we can measure evidence of intersectional behavior in language models with respect to stereotyping. In particular, we are

interested in questions surrounding how language models stereotype people who simultaneously belong to multiple social groups. We will only use the term "intersectionality" when specifically considering cases where (per (3) above) the resulting experience (in this case, stereotyping) is more than the sum of its parts. For example, common U.S. stereotypes for Black women are as "welfare queens" (which may show up as low agency in our traits), while common stereotypes for Black men is as "criminal" (which may show up as low communion) [hooks, 1992, Collins, 2002]. To limit our scope, we will only consider pairs of social groups (e.g., cis men), and will refer to the the groups that make up a pair as the component identities (e.g., cis, or men). We aim to answer the following research questions:

1. When presented with a paired identity, is the language model sensitive to the order in which the component identities appear?
2. When paired, do certain social categories dominate others in a language model's predictions?
3. Can the language model detect stereotypes that belong to an intersectional group (but not to either of the components that make up the pair)?

To answer these questions, we use the SeT measurement with the RoBERTa model (the best performing pair on the single-group experiments) to compute group-trait associations on our paired groups, which are combinations of all the single groups in Table 5.3. We manually omit the groups that do not logically exist (e.g. "cis non-binary person", "teenage elderly person") or are grammatically awkward (e.g. "doctor elderly person", "immigrant blind person"). Note we include both orders of the single groups in the paired groups when possible (e.g. "Catholic teenager" and "teenage Catholic person"). We then conduct the analysis by computing the correlation between groups' list of trait scores with Kendall's τ.

Q1: Identity Order. Given an paired group with two identities, the language model may not be able to capture both of the identities and may predict stereotypes based only on one of the components. In fact, the average correlation score between a paired group and the most correlated of its components is 0.56, which is moderately high. We thus calculate the correlation of trait scores between the paired group and both its first and second component identities (when both orders are possible). In addition, we calculate the correlation of paired groups with reversed identity order (e.g. "Asian teenager" and "teenage Asian person"). The average correlation score between a paired group and its first component is 0.43; the correlation score to its second component is 0.46, which are quite close. Further, the average correlation score of intersectional groups with reversed identity is 0.69, which is moderately high. Taken together, these results indicate that (a) many paired groups have similar group-trait association scores with one of their component identities alone; (b) the order does not matter significantly, but the language model tends to focus slightly more on the second component. The implication of this is that we can expect that the language model *may* be able to capture intersectional stereotypes.

Q2: Dominant Domains. Stryker [1980] suggests that people tend to identify themselves with their race/ethnicity identity before other identities, though this is contested and, in some cases, thought to be antithetical to the idea of intersectionality [e.g., Collins, 2002]. Prompted by this debate, we ask if there is a hierarchy of the domains that language model picks up on for paired groups. To answer this question, for each identity domain pair, we compute the average correlation score between the paired groups with each of its two component identities, and take the difference of the averaged correlation scores of the two domains. For each domain, we count the domains it dominates (i.e. has score difference ≥ 0.1) and is dominated by.

These results show that age and political stance are dominant domains, which is expected as identities within these two domains have strong characteristics that may overwhelm domains they are paired with. On the other end, race and nationality are, generally, dominated domains. It is surprising that the race domain is majorly dominated, contrasting documented literature in human behavior. The full results are shown in appendix Table C.5 as well as detailed scores Table C.6.

Q3: Emergent Intersectional Stereotypes. Finally, we look into emergent stereotypes of paired groups, with the goal of finding intersectional behavior in the language model. To detect intersectional stereotypes, we need to operationalize the notion of the whole being greater than its parts. For a fixed paired group $\mathtt{g} = (\mathtt{g}_1, \mathtt{g}_2)$ (e.g., "trans Democrats"), and a given trait $\mathtt{t}$ (e.g., `warm`), we compute $S(\mathtt{g}, \mathtt{t}) - \max\{S(\mathtt{g}_1, \mathtt{t}), S(\mathtt{g}_2, \mathtt{t})\}$, where S is the score from the language model, capturing whether this trait is more associated with the paired group than the maximum of its association with the component identities. (We consider also the reverse, where we look for scores much less than the min.) We might hope to find some well attested intersectional identities from the literature, such as "Black women" `have an attitude` (low communion) and "White men" are `privileged` (high agency) [Ghavami and Peplau, 2013].

The top 50 emergent group-trait associations according to our measure are listed in appendix Table C.7. We also see some good examples are: the language model scores "Hispanic unemployed people" as more `egotistic` than people of the component identities, "Democrat teenagers" as more `altruistic`, "male doctors" as more `benevolent`, etc. However, there are also some unexpected patterns; for instance, almost all nationality identities combined with "mechanic" are `trustworthy` and `likeable`, and almost all nationality identities combined

with "autistic" are `egotistic`. Looking into the scores themselves, we find that both "mechanic" and "autistic" have low scores on the corresponding traits, and combining them with nationalities raises to about average levels.

Aside from analyzing face validity—which is mixed—we compare the results of our model to the traits that Ghavami and Peplau [2013] found when conducting human studies of race/gender pairs. To do this, we categorize the traits from Ghavami and Peplau [2013] to the ABC dimensions[7] and compare with our full list of emergent group-trait associations. Taking their group-trait matches as ground truth, our detection of traits for these race/gender intersectional groups achieves a precision 0.83 and recall 0.65—better than random guessing (precision 0.72, recall 0.50) but far from perfect.

5.1.7 Discussion and Limitations

In this section, we measured language model (LM) stereotypes by adopting the ABC stereotype model from social psychology. Comparing to previous work on detecting LM stereotypes, our approach is easy to extend to previously unconsidered groups, grounded in traits proven effective by social psychology, and exhaustively covering the space of possible stereotypes, at the cost of being more abstract than in other NLP work. This yields a different set of trade-offs than previous approaches to measuring stereotypes in LMs.

With the ABC model and data regarding human stereotypes from our human study, we assessed LM stereotypes using three different association measurements, including SeT, a metric we proposed. We showed that LM group-trait stereotypes in general have moderate correlation

[7] Ghavami and Peplau [2013] covers paired groups combined with race domain and binary genders. The traits they raised span the agency and communion dimensions.

with human judgements, and that SeT provides correlations that better align with human's. Based on these results, we extended our analysis to intersectional groups. We found that the LM *may* be able to capture intersectional stereotypes but is not particularly good on identifying emergent intersectional stereotypes. Our results also show that, in general, age and political stance are dominant domains in language models, whereas race and nationality are dominated domains.

Limitation First, our results are likely affected by reporting bias and by a defaulting effect where, when people annotate traits for "men", they may actually have in their head "cis straight white men", because the defaults go unremarked. This goes both for the human scores (how does a participant conceptualize "men"?) and language model scores (what do sentences containing the word "man" assume given that most language models have been trained on likely exhibits defaulting?).

Second, we only focus on assessing stereotypes within language models and not in any deployed system. Though stereotypes from language models may impact the outputs of downstream systems which are built upon these language models, it is not clear how exactly the stereotypes transfer [Cao et al., 2022a]. Additionally, this study is limited to English and U.S. social stereotypes.

Third, although we followed and built on best practices from social psychology in developing the human study, it nevertheless has some shortcomings. In particular, even after many iterations on wording, it was difficult to phrase the survey questions to encourage people to reporting their true impressions. There is tension between asking a participant what *they* think—which risks a counfounding potential social desirability bias [Latkin et al., 2017] (people's tendency to respond in socially acceptable ways)—and asking what they think *others* think—which led to

comments from a few participants that they felt unqualified to speak for others. Asking these questions of participants and collecting the data also raises the possibility of inadvertantly reinforcing stereotypes.

Finally, aggregating human judgements into a single number by averaging (or any other statistic) to compare to model predictions risks collapsing a significant amount of information down to a single number. This number cannot distinguish between a weakly held but common stereotype and a strongly held but rare one. Nor can it distinguish between traits where half of annotators say 0 and the other half say 100, from traits where all annotators say 50. These average judgments should be interpreted as not what any single person would say, but an average over people. This limitation is exacerbated by the defaulting effect, where some people may imagine a different prototype for a given group, and other people may imagine another.

5.2 Multilingual large language models leak human stereotypes across language boundaries

Joint work with Anna Sotnikova[1], Jieyu Zhao, Rachel Rudinger, Linda Zou, Hal Daumé III.

Under Submission

Given the approach to measure stereotypes, we extend our analysis to investigate the phenomenon of stereotype leakage across languages in multilingual language models. As these models become more language-agnostic and are trained on texts from diverse languages and cultures, there is a risk of harmful stereotypes from one culture being transferred to another. In this section,

[1]Equal contribution.

we first define the term "stereotype leakage" and propose a framework for its measurement. With this framework, we investigate how stereotypical associations leak across four languages: English, Russian, Chinese, and Hindi. Our findings show a noticeable leakage of positive, negative, and non-polar associations across all languages.

5.2.1 Overview

Cultural stereotypes about social groups can be transmitted based on how these social groups are represented, treated, and discussed within each culture [Martinez et al., 2021, Lamer et al., 2022, Rhodes et al., 2012]. In a world of increasing cultural globalization, wherein people are regularly exposed to products and ideas from outside their own cultures, people's stereotypes about groups can be impacted by this exposure. For instance, blackface is characterized as one of America's first cultural exports, as the performance of American minstrelsy shows in different countries popularized racist depictions of Black Americans within those other cultures [THELWELL, 2020].

Recently, the deployment of large language models has the potential to exacerbate the issue. Large language models are becoming increasingly language-agnostic. For instance, models like ChatGPT[8] and mBART [Lin et al., 2022] can operate without being restricted to a specific language, handling input and output in multiple languages simultaneously. This thus gives rising opportunities for what we refer to as stereotype leakage, or the transmission of stereotypes from one culture to another.

Stereotype leakage within large language models may export harmful stereotypes across

[8] https://openai.com/chatgpt

cultures and reinforce Anglocentricism[9]. Previous works [e.g., Goldstein et al., 2023, Weidinger et al., 2021] have highlighted the potential for language model outputs to change users' perceptions and behaviors. Stereotype leakage from large language models may further entrench existing stereotypes among model users, as well as create new stereotypes that have transferred from a different language. Therefore, in this section, we investigate the degree of stereotype leakage within multilingual large language models (MLLMs) as a step toward understanding and mitigating stereotype leakage for AI systems.

Large language models are currently the backbone of many natural language processing (NLP) models. MLLMs are language models pre-trained with a large amount of data from multiple languages so that they can process NLP tasks in various languages as well as cross-lingual tasks. Recent MLLMs, such as GPT models [Brown et al., 2020, OpenAI, 2023] designed for standalone applications and models such as mBERT [Müller et al., 2020], XLM [Lample and Conneau, 2019], mT5 [Xue et al., 2020], mBART [Lin et al., 2022], intended for use as back-end tools, show satisfactory performance on NLP tasks across around 100 languages. One major advantage of such models is that low-resource languages (languages with less training data) can benefit from high-resource languages through shared vocabulary [Lample and Conneau, 2019] and structural similarities (word-ordering or word-frequency) [K et al., 2020].

Large language models are trained on existing language data, and even monolingual language models have been demonstrated to replicate stereotypical associations present in the training data. [Nadeem et al., 2020, Nangia et al., 2020, Cao et al., 2022b]. Thus, with the shared knowledge between languages in MLLMs, it is likely that stereotypes may also leak between

[9]Anglocentrism is the practice of viewing and interpreting the world from an English-speaking perspective with the prioritization of English culture, language, and values. Anglocentrism can lead to biases and neglect of global perspectives and experiences.

languages. Stereotypes are abstract and over-generalized pictures drawn about people based on their group membership, and these perceptions can be specific to each culture. Though LLMs are trained on language-based data rather than culture-based data, languages reflect the stereotypes associated with the cultures they represent. Thus, for the purpose of studying stereotypes in MLLMs, we divide the world according to languages, with the understanding that a single language may reflect multiple cultures. Previously, many works have examined Western stereotypes in English language models [e.g. Nadeem et al., 2020, Nangia et al., 2020, Cao et al., 2022b], whereas limited works have attempted to assess stereotypes in multilingual language models [e.g. Kaneko et al., 2022, Levy et al., 2023, Câmara et al., 2022] due to the complexity of stereotypes manifested in various cultures, limited resources, and Anglocentric norms [Talat et al., 2022].

In this study, we aim to investigate the existence of *stereotype leakage* in MLLMs, which we define as the effect of stereotypical word associations in MLLMs of one language impacted by stereotypes from other languages. We conduct a human study to collect human stereotypes, adopt word association measurement approaches from previous works [Cao et al., 2022b, Kurita et al., 2019] to measure stereotypical associations in MLLMs and analyze the strength and nature of stereotype leakage across different languages both quantitatively and qualitatively.

To test our hypothesis that there are significant stereotypes leaked across languages in MLLMs, we sample four languages: English, Russian, Chinese, and Hindi. We pick languages that come from the Indo-European and Sino-Tibetan language families, ranging from high (English) to low-resource (Hindi) languages[10]. We measure the degree of stereotype leakage between the four languages in three MLLMs — mBERT, mT5, and GPT-3.5[11]. Both mBERT and mT5

[10]High-resource languages are languages that have more training data available, while low-resource languages have less.

[11]Both the code and the dataset, along with a datasheet [Gebru et al., 2018b], are available under a MIT licence at: https://github.com/AnnaSou/Stereotype_Leakage.

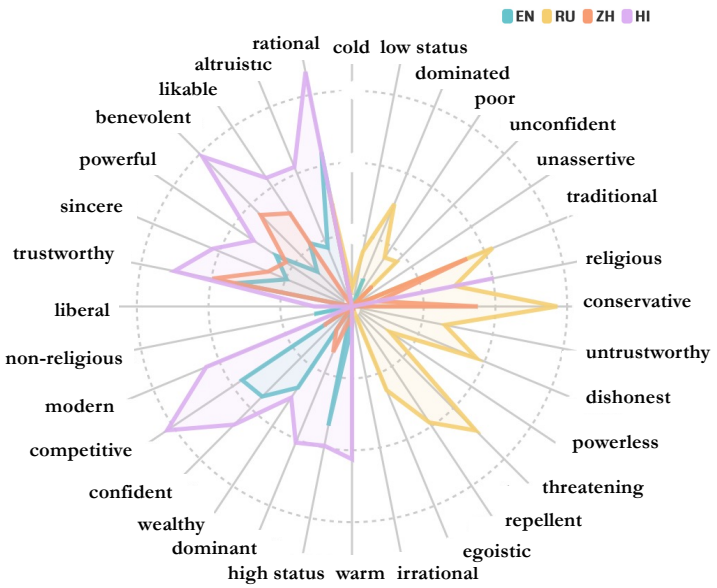

Figure 5.2: The figure shows results of human annotations in English (EN), Russian (RU), Chinese (ZH), and Hindi (HI) languages based on ABC model for the social group “Asian people”. It shows average scores across all annotators per language.

are back-end MLLMs. MT5 has better multilingual performance than mBERT, whereas mBERT has more comparable monolingual BERT models for the four languages. GPT-3.5 is one of the state-of-the-art MLLMs that has been popularly deployed to users. With these, we examine the impact of human stereotypes from different languages on stereotypical associations in MLLMs.

5.2.2 Measuring Stereotype Leakage in MLLMs

For each language, we aim to assess the degree of stereotype leakage from the other languages to this target language in MLLMs. Specifically, we measure the effect of human stereotypes from all four languages ($H_{en}, H_{ru}, H_{zh}, H_{hi}$) on the target language's MLLM stereotypical association ($MLLM_{tgt}$), as shown in Equation 5.3. We also control the impact of the stereotypical association from the target monolingual model (LM_{tgt}). However, since we can only find monolingual BERT model for all four languages, we use these as LM_{tgt} for all MLLMs. We use a mixed-effect model to fit the formula and calculate the effect. If the coefficient of a variable is positive and has a p-value of less than 0.05, then the variable has a significant effect on $MLLM_{tgt}$. If there are significant effects from the non-target language's human stereotypes, then there are potential stereotype leakages from this language to the target language.

$$MLLM_{tgt} = \alpha_{en}H_{en} + \alpha_{ru}H_{ru} + \alpha_{zh}H_{zh} + \alpha_{hi}H_{hi} + \beta LM_{tgt} + C \tag{5.3}$$

In the following section, we discuss how we measured each of the variables.

5.2.2.1 Stereotype Measurement

We measure stereotypes through group-trait associations as proposed in section 5.1. We use 16 trait pairs (each pair represents two polarities) from the Agency Beliefs Communion (ABC) model of stereotype content [Koch et al., 2020], as listed in Table 5.1. If a group (e.g. "immigrant", "Asian person") has a high degree of association with a trait (e.g. `religious`, `confident`), then we consider that trait a stereotype of the group. For example, Figure 5.2 is the stereotype map of the group "Asian people" collected from our human study across the four languages that we study.

For the groups, we pick 30 groups listed in Table 5.7: 10 *shared groups with shared stereotypes* (groups that are present in all four countries and are expected to be targeted by similar stereotypes), 8 *shared groups with non-shared stereotypes* (groups that are present in all four countries but expected to be targeted by dissimilar stereotypes), and 12 *not shared groups* (groups that exist uniquely in each country; three groups for each country). For shared groups, we manually select groups from the list of social groups from Table 5.3. To collect not shared groups, we conduct a survey among native speakers. For each language, we ask 6 native speakers to list $5 - 10$ social groups that they believe are unique to their culture. We then choose 3 social groups per language based on the outcome of the majority vote.

In our human study, we further verify that each group matches the property of its category. To illustrate, stereotypes of groups in the first category exhibit an average correlation score of 0.60 across languages. In contrast, groups in the second and third categories demonstrate progressively lower correlation scores of 0.50 and 0.26, respectively.

Category	Groups
Shared/ Shared	man, woman, gay, lesbian, single mother, housewife, software engineer, wealthy person, poor person, disabled person
Shared/ Non-shared	Asian person, Black person, Muslim person, immigrant, government official, civil servant, feminist, veteran
Non-shared/ Non-shared	*USA*: Texan, Mormon, Puerto Rican *Russia*: VDV soldier, Muscovite, Chechenets *China*: migrant worker, Hui person, Shanghainese person *India*: Brahmin person, Gujarati person, Shudra person

Table 5.7: Categories and corresponding social groups were used for the model and human experiments. "Shared/Shared" represents shared groups and shared stereotypes. "Shared/Non-shared" represents shared groups and non-shared stereotypes. "Non-shared/Non-shared" represents non-shared groups and non-shared stereotypes.

5.2.2.2 Human stereotypes

To collect human stereotypes, we conduct a human study on Prolific[12] for each of the four languages with native speakers of the respective languages who lived or still live in the United States, Russia, China, and India[13]. In the survey, participants are first asked to mark at least 4 social groups that they feel they are familiar with. Then they are asked to rate the group-trait associations of 4 social groups from their list of familiar groups. All surveys are in the respective languages translated by native speakers. For shared/shared and shared/non-shared groups, we collect at least 5 participants' annotations per group per language. For non-shared groups with non-shared stereotypes, we collect at least 5 annotations for the language they originate from, with no minimum limit of annotations for other languages.

Human Study We follow the same approach as in subsection 5.1.5 to collect human stereotypes. Participants first read the consent form, and if they agree to participate in the study, they see the survey's instructions. For each social group, participants read in their respective language,

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

[13] Approved by our institutional IRB, #1724519-3.

"As viewed by American/Russian/Chinese/Indian society, (while my own opinions may differ), how `[e.g., powerless, dominant, poor]` versus `[e.g., powerful, dominated, wealthy]` are `<group>`?".

Participant Demographics We collect participants' demographic information including gender, age, education level, and (for non-English speakers) information about how frequently they read American social media. Participants can refrain from providing answers to any of these questions.

After averaging the gender distribution across all languages: men 0.49, women 0.45, non-binary/transgender/gender fluid 0.05, and the rest of the participants preferred not to answer. Participants in the Russian survey were the ones who read American media most frequently: 0.44 read it regularly compared to 0.35 and 0.28 percent for Hindi and Chinese, respectively. On average, 0.39 respondents read American media from time to time. Around 0.05 never read the media.

All approved participants stated that they were fluent in the survey's languages. Each participant was paid $\$2.00$ to rate five social groups on 16 pairs of traits, and on average, participants spent about ten minutes on the survey. This resulted in a pay of $\$12.00$ per hour.[14]

5.2.2.3 Model stereotypical associations

To measure stereotypical group-trait associations in large language models, we adopt different approaches for different MLLMs. For mT5, we use the increased log probability score (ILPS) [Kurita et al., 2019], which computes the likelihood for the model to generate a trait given

[14]Maryland's current minimum wage is $\$12.20$ https://www.minimum-wage.org/maryland.

a templated sentence about a group. For example, $p(\texttt{competent} | \text{“}\textit{Asian people are}\ \_\_\text{.”})$ indicates the correlation between `competent` and “Asian people”.

For mBERT, we use the sensitivity test (SeT) [Cao et al., 2022b], which is shown to have better alignment with human stereotypes [Cao et al., 2022b]. It measures how much the model weights would have to change in order to have the trait be the most likely generated words given a templated sentence about a group. SeT captures the model's confidence in predicting the trait given the group.

In evaluating GPT-3.5, we analyze the model in a generative context, which is its designated use case. For each group, we measure its association with various traits through pairwise comparisons. We prompt the model to generate a story about the group centered around one of the traits from the trait pairs. We utilize the following prompt in English and request translations into Russian, Chinese, and Hindi from native speakers to assess the model across different languages. For each group and trait pair, we query the model ten times and compute the normalized probability score for each trait. Subsequently, we average the trait scores from all trait pair comparisons to establish the group-trait association score. Our experimentation employs the gpt-3.5-turbo-0125 model. Notably, we observe a tendency in the GPT model to favor positive traits (e.g., `powerful, trustworthy`) over negative traits (e.g., `powerless, untrustworthy`). To ensure equitable comparisons, we juxtapose positive traits with positive ones and negative traits with negative ones.

prompt sys = “You are a writer.”

prompt = “You are writing about [GROUP]. Before writing, think about what theme you want to pick. You can choose either ‘[TRAIT1]’ or ‘[TRAIT2]’ as your theme.

You can also choose 'neither' if you think neither of these themes fits. Note that you can choose only one theme. Output the exact name of the theme only, without any punctuation."

When processing the outputs of GPT-3.5, we employ an exact match criterion to assign scores to traits. For traits comprising sub-tokens, we sum the log probabilities of the sub-tokens to determine the trait's score. In instances where outputs do not precisely match the traits in the prompt, we have them manually processed by native speakers of the respective language. Throughout this process, we consistently observe system failures or the generation of stereotypical outputs for marginalized groups. Notably, for certain groups like "feminist" and "Muslim person" in Chinese, the model often disregards the prompt and simply outputs the group name. Moreover, in some cases, the model alters the trait specified in the prompt. For example, it changes `dominating` to `dominated` for "disabled person" in English or `poor` to `wealthy` for "migrant worker" in Russian. Additionally, the model may overlook the traits provided in the prompt and generate stereotypical traits instead. For instance, in Russian, it generates `rape` and `patriot` for "Puerto Rican" or `cowboy` for "Texan". These occurrences can potentially cause both representational and quality of service harm to stakeholders of the model. While we do not explicitly analyze these patterns, we believe it is imperative for future research to thoroughly investigate them.

5.2.3 Stereotype Leakage and Its Effects

In this section, we present our quantitative and qualitative results of the assessment of stereotype leakage across languages in MLLMs. We study the extent to which human stereotypes

from the four languages are represented in the respective languages in MLLMs' stereotypical associations.

5.2.3.1 Quantitative Results

We compute the stereotype leakage across languages within three MLLMs based on Equation 5.3. The findings are presented in Figure 5.3, illustrating the extent to which stereotypical associations in the target language model are influenced by human stereotypes present in the culture associated with the source language. For example, in Figure 5.3, we observe that within GPT-3.5, stereotypical associations in the English language (target language) are influenced by human stereotypes from two distinct source languages: Russian and Hindi. This observation suggests the presence of stereotype leakages within the GPT-3.5 model.

In our analysis of mBERT, we observe significant leakages of stereotypes from Hindi to English and Chinese with coefficients of 0.02 ($p = 0.009$) and 0.06 ($p = 0.00$), respectively. We also observe English human stereotypes manifesting in mBERT Hindi with a coefficient of 0.02 ($p = 0.048$). Within the mT5 model, we find two significant stereotype leakages, both of which are leakages targeting Hindi. Russian and Chinese human stereotypes manifest in mT5 Hindi with coefficients of 0.02 ($p = 0.047$) and 0.06 ($p = 0.00$), respectively. For GPT-3.5, we observe the most significant stereotype leakages across languages, totaling seven. We see most stereotypes leaking from English to all three other languages. The largest flows are from English to Chinese and Hindi, with coefficients of 0.02 ($p = 0.00$). Overall, we observe that GPT-3.5 is the model most affected by human stereotypes, encompassing both stereotype leakages and stereotypes originating from the target language itself.

Moreover, among all languages, Hindi experiences the highest degree of stereotype leakage — it has four cases of significant stereotype leakage from other languages across three MLLMs. Since Hindi is the only low-resource language we tested, this might explain why it absorbs stereotypes from other languages. Both Chinese and English languages have three leakages across the models. The Chinese language has leakages from all three other languages, while the English language has the most leakage from Hindi. The Russian language has two significant leakages from English and Hindi.

Finally, we report the coefficients of effects from monolingual language models (LM_{tgt}) in Table 5.8. All the effects are statistically significant and are stronger than the effects from human stereotypes. This is not surprising because monolingual language models and multilingual language models share similar training data and model structures.

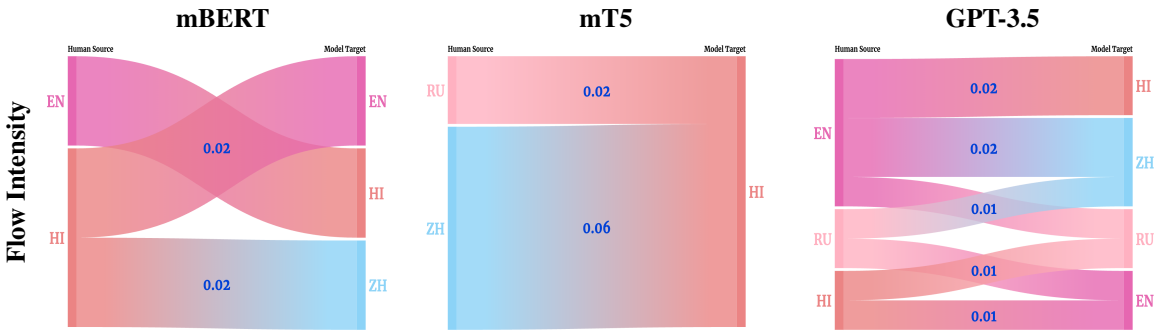

Figure 5.3: The figures show the stereotype leakages for three models: mBERT, mT5, and GPT-3.5 respectively. Each figure illustrates the flow from the human source language (the left column) to the target language in a particular model (the right column). If no flow for a particular language is presented, this means that no significant leakage is happening.

5.2.3.2 Qualitative Results

Moving forward, we delve into the specific stereotypical associations that leak from one language to another, considering the potential implications of such strengthened associations.

Monolingual BERT	**EN**	**RU**	**ZH**	**HI**
mBERT	0.33	0.29	0.17	0.08
mT5	0.10	0.45	0.14	0.14
GPT-3.5	0.07	0.05	0.05	0.06

Table 5.8: Mixed-effect coefficients of monolingual BERTs in the respective languages contributing to the same languages in multilingual language models. All of the effects are statistically significant. Note that the coefficients are not comparable across multilingual language models as the score ranges are different.

We focus on the GPT-3.5 model, in which we observe the most leakage from human stereotypes. Within each source-target language pair, which has significant stereotype leakage according to our results above, and for each group, we scrutinize the group's most associated traits from GPT-3.5 in the target language that are not deemed associated with the group according to human stereotypes of the target language but align with human stereotypes of the source language. Our analysis reveals two primary types of leakages: the amplification of both positive and negative representations for certain languages. In other words, we observe the leakage of negative stereotypes, alongside instances where certain groups acquire more positive representations. Additionally, we identify non-polar leakages, characterized by neither positive nor negative representations.

Positive Leakage According to human annotation, "Asian people" are more positively perceived in English language than in Russian. We observe the strengthening of such traits in GPT-3.5 Russian language as `wealthy, likable`, and `high status`. Moreover, "housewives" become more `warm` in English following leakages from Russian and Hindi. "Black people" are more `powerful, modern, confident`, and `wealthy` in the English language following leakage from Hindi. Another example of the leakage of positive perceptions is for "gay men" and "lesbians" from English to other languages. Traits such as `likable, confident, warm,`

dominant, sincere, and powerful become stronger in Russian, Chinese, and Hindi.

Negative Leakage On the other hand, there are negative stereotypes that leak across languages. From "feminists", we observe a leakage from English to Chinese and Hindi and from Russian to Chinese of such stereotypical associations as egoistic, threatening, repellent, and cold, while, for instance, in human survey in Hindi this group is perceived as warm. Another example is "immigrants". From Russian and English languages, traits such as threatening, repellent, dishonest, egoistic, and unconfident leak to Chinese and Hindi. Based on human data, we found that people surveyed in Chinese view this group quite favorably since the majority of immigrants to China are highly qualified professionals [Pieke, 2012]. In Russia, immigrants are mostly coming from poorer neighboring countries and are negatively stereotyped in society, while in the U.S., immigrants are diverse and could be both marginalized and more privileged. Moreover, there is a notable leakage from English to Chinese and Hindi for "Black people" for traits dominated and poor. This aligns with known stereotypes about African Americans and Africans in U.S. society [Miller-Cribbs and Farber, 2008, Galster, 1992, BERESFORD, 1996].

Non-polar Leakage There are also non-polar leakages, which are neither positive nor negative. From Hindi to English and Russian, we see the strengthening of religious for various groups such as "women", "disabled people", "Black people", and "Asian people". It has been shown that there are more than 70.00% believers of the total population in India as of 2011[Evans and Sahgal, 2021].

Non-shared Groups Leakage In the case of non-shared groups, we expected uni-directional transferring of the groups' perceptions from the language of origin to other languages. Our findings confirm this hypothesis. For example, the group "VDV soldiers" is a widely known military unit in Russia. There are strong stereotypes in Russian society about this group, but the group is mostly unknown to Americans. Out of the 34 survey English survey respondents who passed the quality tests, no one chose this group as a familiar one. This group's representation leaks from Russian to English, strengthening traits such as `confident, traditional, competitive`, and `threatening`. Another example is "Hui people", a group widely unknown to Russian and Hindi society: out of 76 respondents for both surveys, no one chose this group as the familiar one. This social group is a minority in China and is composed of Chinese-speaking followers of Islam. Originally, "Hui people" are marginalized in China and viewed as more traditional, religious, and conservative [Hillman, 2004, Hong, 2005]. Accordingly, we observed the leakage of such traits as `irrational, traditional, threatening, repellent, religious, and egoistic`. All groups specific to the Hindi language — "Gujarati, Brahmin", and "Shudra people" — have certain traits leaking to the English and Russian languages. For example, high caste groups ("Gujarati" and "Brahmin people") strengthen such positive traits as `wealthy, likable, sincere, powerful, high status, competitive`, and `confident`. In addition, "Brahmin people" become more associated in GPT-3.5 with traits `poor, low status, powerless, traditional, religious`, and `dominated`. This leakage corresponds to the perception of these groups in Indian society and by our survey respondents [Witzel, 1993, Milner, 1993].

5.2.4 Discussion and Limitations

Multilingual large language models have the potential to spread stereotypes beyond the societal context they emerge from, whether by generating new stereotypes, amplifying existing ones, or reinforcing prevailing social perceptions from dominant cultures. In this study, we demonstrate that this concern is indeed valid. To do so, we establish a framework for measuring the leakage of stereotypical associations in multilingual large language models across languages. We are limited in our ability to run a causal analysis, because none of the studied languages can be easily removed from the training data to see their genuine impact on stereotypical associations in other languages. Retraining GPT-3.5, for instance, is not a feasible option. Nonetheless, if our hypothesis holds, indicating that stereotype leakage is occurring, we would anticipate observing associations of stereotypes cross-lingually, and indeed, we do identify this association. As a proxy for re-training the models without a particular language, we run an association on a monolingual version of the model. We perform this experiment on both monolingual BERTs and multilingual BERT to measure how well the monolingual BERT in a specific language can predict the behavior of mBERT in the same language. We find stronger associations between stereotypes in monolingual English and Russian models with mBERT in the same languages than for the case of Chinese and Hindi languages. In addition, we find that there is interaction and exchange between languages in multilingual large language models. The stereotype leakage occurs bidirectionally. On the example of GPT-3.5, as the best-aligned with human judgment model, we observe the strengthening of positive, negative, and non-polar associations in the model. In addition, our study underscores the role of "native" languages in framing social groups unknown to other linguistic communities. Such leakage of stereotypes amplifies the complexity of societal

perceptions by introducing a complex interconnected bias from different languages and cultures. In the context of shared groups, stereotype leakage may manifest as the manifestation of stereotypes that were not previously present within the cultural setting of a particular group. In the case of non-shared groups, stereotype leakage can extend the reach of existing stereotypes from the source culture to other cultural contexts.

To our knowledge, we are the first to introduce the concept of stereotype leakage across languages in multilingual LLMs. We propose a framework for quantifying this leakage in multilingual models, which can be easily applied to unstudied social groups. We show that multilingual large language models could facilitate the transmission of biases across different cultures and languages. We demonstrate the existence of stereotype leakage within MLLMs, which are trained on diverse linguistic datasets. As multilingual models begin to play an increasingly influential role in AI applications and across societies, understanding their potential vulnerabilities and the level of bias propagation across linguistic boundaries becomes important. As a result, we lay the groundwork for advancing both the theoretical comprehension of multilingual models and the practical implementation for bias mitigation in AI systems.

Limitations First, stereotypes were selected based on the ABC model, which was developed and tested using U.S. and German stereotypes. We translated our surveys into the other languages but this might result in patterns that better reflect Anglocentric stereotypes [Talat et al., 2022] than other stereotypes. In our study, we try to control the influence of the U.S. culture by asking crowd workers how frequently they read U.S. social media. We see that on average 39% of respondents in Russia, China, and India read the media. This American cultural dominance might affect the collected data as these data may not fully capture the range of stereotypes typical for these

cultures. In addition, we have language-culture limitations as English language survey results only apply to the U.S., Russian to Russia, Chinese to China, and Hindi to India. Lastly, while we indirectly consider culture through survey results on associations, we do not measure or account for culture in a comprehensive manner.

5.3 Conclusion

In this chapter, we have introduced a systematic approach for analyzing stereotypes in large language models, which is both theory-grounded and easy to generalize to as yet unexamined social groups. This methodology enables us to explore stereotypes about a wide range of identities. Expanding upon this framework, our study goes beyond the examination of stereotypes in the English language and those specific to the U.S. culture. We look into cross-cultural stereotypes within multilingual language models. Specifically, we proposed a framework to examine stereotype leakages across language boundaries in these models. Our findings, particularly within the GPT-3.5 model, indicate the presence of such stereotype leakages. Given the widespread adoption of GPT models, the leakage of stereotypes poses a potential risk of perpetuating harm across language and cultural boundaries. Looking ahead, we hope the framework we have established will lay a foundation for future research aimed at a deeper understanding of stereotypes within language models as well as reducing any negative effects these stereotypes may have on all stakeholders involved with each AI system.

Chapter 6: Conclusion and Future Directions

In this thesis, I have investigated the crucial issue of aligning artificial intelligence systems with the needs of users and the values of diverse stakeholders. Through a series of studies, I have identified challenges and proposed approaches to bridge the gap between current AI technologies and their meaningful application in real-world contexts.

First, I examined potential causes of misalignment in AI system development, focusing on gender biases in natural language processing systems. By assessing each stage of the machine learning lifecycle, I demonstrated how biases can enter the system when developers fail to carefully consider the system's real-life stakeholders.

Next, I explored the alignment of AI systems with specific user groups in two application contexts: content moderation assistance for volunteer moderators and visual question answering for blind and visually impaired users. In both cases, I identified significant gaps between the AI systems' capabilities and the users' needs. These findings provide valuable direction for future work to improve the alignment of AI technologies with specialized user requirements.

Finally, I addressed the critical issue of aligning AI systems with human values, concentrating on the problem of stereotyping in large language models. I proposed a theory-grounded method for systematically evaluating stereotypical associations, considering diverse user identities, intersectional stereotypes, and cross-cultural stereotype leakage. This approach contributes

to the development of safe and inclusive AI systems that align with the values of all stakeholders.

Overall, the findings have demonstrated that AI systems with a "one-size-fits-all" approach are inadequate in real-world scenarios. Coreference resolution systems, for instance, fail to adequately represent a spectrum of gender identities due to their development based on oversimplified gender concepts. Similarly, visual question answering models, designed around challenge datasets, fall short in addressing practical needs of blind and visually impaired users. Their standard evaluation metrics do not accurately capture the diverse preferences of different users either. In the realm of content moderation, current NLP models are not equipped to handle the varied moderation rules required by moderators, nor can they cater to the specific performance preferences of different moderators for various rules. Finally, large language models have been found to perpetuate stereotypes across a range of identities and cultures. My study also revealed that individuals from different backgrounds perceive stereotypes differently, posing a significant challenge in developing systems that align with the values of all stakeholders. It's clear that AI systems must prioritize safety for all stakeholders. However, the definition of safety and the values an AI system aligns with should be tailored to the specific applications of each system and the needs of its users.

These findings underscore the importance of adopting a human-centered approach in the development of AI systems. In my thesis, I have demonstrated how the AI systems development process can be improved by incorporating strategies from the field of Human-Computer Interaction (HCI). Firstly, involve more users in the development process and anchor it in specific applications. In developing content moderation assistance systems, my study revealed that focusing on specific applications can lead to innovative NLP task definitions, previously unconsidered. Interaction with volunteer content moderators has also provided valuable insights into the practi-

cal use of these systems, guiding improvements in current models and highlighting areas where research efforts may not be necessary. In the development of VQA systems, my study highlighted the benefits of basing models on real user data. This approach helps in identifying model issues that genuinely impact user experience and can be barriers to the effective use of these systems. Finally, as pointed out in the VQA example, the involvement of users is crucial in the evaluation phase of AI systems. If the evaluation metrics are not well-aligned with user needs and preferences, they may not accurately represent the model's performance. Such misalignment can also lead to misguided directions in system development.

Moreover, my studies have highlighted the advantages of collaborating with experts from fields outside of machine learning. For example, in the study of assessing stereotypes within large language models, we integrated a stereotype model from the domain of social psychology. This field has a long history of studying stereotypes, with extensive research into the development of stereotype models. Building based on their models provided a critical dimension to our research, allowing for a more nuanced and comprehensive exploration of stereotypes within large language models. In addition to this, my other studies also benefitted greatly from collaboration with domain-specific experts. Engaging with professionals who specialize in these particular areas brought a fusion of diverse perspectives and expertise. Such collaborations led to innovations that might have been unattainable within the silos of a single discipline.

Overall, the methodologies and insights presented in this work offer a foundation for future research and development efforts aimed at aligning AI systems with the complexities of human needs and values in diverse contexts.

6.1 Future Directions

6.1.1 User-Centered Evaluation Metrics for Blind and Low Vision VQA Users

With the overarching aim of creating AI systems that resonate with the requirements of real-life users, I intend to bridge the gaps identified in our research. As highlighted in section 4.2, our findings indicate that current evaluation metrics for Visual Question Answering (VQA) systems don't accurately capture user preferences or the model's actual performance. Therefore, my goal is to develop a more effective evaluation metric grounded in user preferences. The key research question is to determine if Blind and Visually Impaired (BVI) users have distinct needs for model-generated responses in different contexts.

For instance, BVI users might need concise, factual answers during social events but prefer more comprehensive responses when at home. To tackle this research challenge, my plan includes conducting a diary study followed by a user study involving BVI participants. This approach will help us gather in-depth insights into user preferences and frustrations across various scenarios. The outcomes of this research are anticipated to lead to more accurate evaluation metrics and contribute significantly to the development of VQA models. These models will be better equipped to produce answers that cater specifically to the diverse needs of BVI users.

6.1.2 Challenges Faced by Users with Intersectional Identities in Large Pre-trained Models

Existing works on fairness mostly focus on social groups with single identities. Similarly, accessibility works often focus on users with disabilities without considering other demographic

identities [Harrington et al., 2023]. However, as discussed in subsubsection 5.1.6.2, intersectional experience is non-trivial and can be greater than the sum of their component identities [Combahee River Collective, 1977]. Hence, I aim to explore and improve on emerging challenges faced by individuals with demographic identities that intersect with disability when interacting with large pre-trained models.

To mitigate biases and stereotyping, developers of publicly released large pre-trained models often implement restrictions, leading the models to refuse to speak about specific topics. While well-intentioned, refusing to engage with sensitive subjects about marginalized groups creates additional erasure harm — excluding these identities and perspectives. For example, even when explicitly prompted, models like DALL-E sometimes refuse to generate figures of neurodivergent people. This demonstrates the downside of quick-fix attempts to avoid stereotypes; they can erase identities, running counter to goals of equitable treatment. As we consider fairness in AI systems, we must balance mitigating stereotypes with avoiding further marginalization of minority groups. This can be increasingly complicated with intersectional identities, with overlapping and sometimes contradictory stereotypes and assumptions made about them. While large pre-trained models may be able to manage single identities, it remains unclear how well they can hold through multiple overlapping identities. Hence, with the vision to mitigate stereotyping and harm to stakeholders with intersectional identities with disability, my goal is first to understand these intersections.

I plan to continue collaborating with social psychologists to gain better insights into intersectional identities with disability. Implicitly, I will develop frameworks to assess the stereotypes present in these models concerning intersectional identities, going beyond the union of their com-

ponent identities. Explicitly, I plan to conduct user studies to understand how users with diverse intersectional identities with disability interact with these large pre-trained models. From that, I will curate evaluation datasets and develop evaluation paradigms to assess the models' levels of over-sensitivity or under-sensitivity regarding content about or from intersectional identities with disability identities.

In addition, towards bias mitigation, I plan to explore the generalizability of the existing mitigation approaches – extending from individual groups to intersectional groups. Based on my work on stereotype measurement, I intend to explore the extent to which bias mitigation applied to individual identities generalizes to intersectional groups and the scope of this generalization.

Appendix A:

A.1 Annotation of ACL Anthology Papers

Below we list the complete set of annotations we did of the papers described in **??**. For each of the papers considered, we annotate the following items:

- Coref: Does the paper discuss coreference resolution?
- L.G: Does the paper deal with linguistic gender (grammatical gender or gendered pronouns)?
- S.G: Does the paper deal with social gender?
- Eng: Does the paper study English?
- L≠G: (If yes to L.G and S.G:) Does the paper distinguish linguistic from social gender?
- 0/1: (If yes to S.G:) Does the paper explicitly or implicitly assume that social gender is binary?
- Imm: (If yes to S.G:) Does the paper explicitly or implicitly assume social gender is immutable?
- Neo: (If yes to S.G and to English:) Does the paper explicitly consider uses of definite singular "they" or neopronouns?

For each of these, we mark with [Y] if the answer is yes, [N] if the answer is no, and [-] if this question is not applicable (ie it doesn't pass the conditional checks).

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Sidner [1981]	Y	Y	Y	Y	N	-	-	-
Bainbridge [1985]	Y	Y	N	Y	-	-	-	-
Kameyama [1986]	Y	Y	Y	Y	N	Y	Y	N
Mellish [1988]	N	Y	N	Y	-	-	-	-
Danlos and Namer [1988]	N	Y	N	N	-	-	-	-
Yoshimoto [1988]	N	Y	N	N	-	-	-	-
Zock et al. [1988]	N	Y	N	N	-	-	-	-
Popowich [1989]	N	Y	N	Y	-	-	-	-
Mani et al. [1993]	Y	N	Y	Y	-	Y	-	-
Narayanan and Hashem [1993]	N	Y	N	N	-	-	-	-
Soloman and Wood [1994]	N	Y	N	Y	-	-	-	-
Quantz [1994]	N	Y	N	Y	-	-	-	-
Baker et al. [1994]	-	-	-	-	-	-	-	-
Genthial et al. [1994]	N	Y	N	N	-	-	-	-
Levinger et al. [1995]	N	Y	N	N	-	-	-	-
Holan et al. [1997]	N	Y	N	N	-	-	-	-
Dorna et al. [1998]	N	N	N	Y	-	-	-	-
Harabagiu and Maiorano [1999]	Y	Y	Y	Y	N	Y	Y	N
Avgustinova and Uszkoreit [2000]	N	Y	N	N	-	-	-	-
Channarukul et al. [2000]	N	Y	N	Y	-	-	-	-
Abuleil et al. [2002]	N	Y	N	N	-	-	-	-

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Cucerzan and Yarowsky [2003]	N	Y	N	N	-	-	-	-
Pakhomov et al. [2003]	N	N	Y	Y	-	-	-	-
Tadić and Fulgosi [2003]	N	Y	N	N	-	-	-	-
Debowski [2003]	N	Y	N	N	-	-	-	-
Navarretta [2004]	Y	Y	Y	N	N	Y	Y	-
Carl et al. [2004]	Y	Y	Y	N	N	Y	Y	-
Mota et al. [2004]	N	Y	N	Y	-	-	-	-
Eisner and Karakos [2005]	N	Y	N	Y	-	-	-	-
Boulis and Ostendorf [2005]	N	N	Y	Y	-	Y	Y	N
Smith et al. [2005]	N	Y	N	N	-	-	-	-
Bergsma and Lin [2006]	Y	Y	Y	Y	N	Y	Y	N
Vogt and André [2006]	N	N	Y	N	-	Y	Y	-
Quirk and Corston-Oliver [2006]	N	Y	N	Y	-	-	-	-
Dada [2007]	N	Y	N	N	-	-	-	-
Streiter et al. [2007]	N	N	Y	N	-	-	-	-
Jing et al. [2007]	Y	Y	Y	Y	N	Y	-	N
Badr et al. [2008]	N	Y	N	N	-	-	-	-
Marchal et al. [2008]	N	Y	N	N	-	-	-	-
van Peursen [2009]	N	Y	N	N	-	-	-	-
Badr et al. [2009]	N	Y	N	N	-	-	-	-
Garera and Yarowsky [2009]	N	Y	Y	Y	N	Y	Y	N

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Bergsma et al. [2009]	Y	Y	Y	Y	N	Y	Y	N
Nastase and Popescu [2009]	N	Y	N	N	-	-	-	-
Nanba et al. [2009]	N	N	N	Y	-	-	-	-
Robaldo and Di Carlo [2009]	N	N	N	Y	-	-	-	-
Mukherjee and Liu [2010]	N	N	Y	Y	-	Y	Y	-
Ng [2010]	Y	Y	Y	Y	N	Y	Y	N
Burkhardt et al. [2010]	N	N	Y	N	-	Y	Y	-
Marton et al. [2010]	N	Y	N	N	-	-	-	-
Le Nagard and Koehn [2010]	Y	Y	Y	Y	N	Y	Y	N
Rojas-Barahona et al. [2011]	N	Y	N	N	-	-	-	-
Mukund et al. [2011]	N	Y	N	N	-	-	-	-
Sarawgi et al. [2011]	N	N	Y	Y	-	Y	Y	N
Li et al. [2011]	Y	Y	Y	Y	N	Y	Y	N
Burger et al. [2011]	N	N	Y	Y	-	Y	Y	N
Mohammad and Yang [2011]	N	N	Y	Y	-	Y	Y	N
Sapena et al. [2011]	Y	Y	Y	Y	N	Y	Y	N
Charton and Gagnon [2011]	Y	Y	Y	Y	N	Y	Y	N
Alkuhlani and Habash [2011]	N	Y	N	N	-	-	-	-
Mareček et al. [2011]	N	Y	N	N	-	-	-	-
López-Ludeña et al. [2011]	N	Y	N	N	-	-	-	-
Declerck et al. [2012]	Y	Y	N	Y	-	-	-	-

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Bergsma et al. [2012]	N	N	Y	Y	-	Y	Y	N
Alkuhlani and Habash [2012]	N	Y	N	N	-	-	-	-
Filippova [2012]	N	N	Y	Y	-	Y	-	-
Dinu et al. [2012]	N	Y	N	N	-	-	-	-
El Kholy and Habash [2012]	N	Y	N	N	-	-	-	-
Yu [2012]	N	N	N	N	-	-	-	-
Guillou [2012]	Y	Y	Y	Y	Y	Y	-	-
Vogel and Jurafsky [2012]	N	N	Y	Y	-	Y	Y	N
Goldberg and Elhadad [2013]	N	Y	N	N	-	-	-	-
Marton et al. [2013]	N	Y	N	N	-	-	-	-
Weller et al. [2013]	N	Y	N	Y	-	-	-	-
Ciot et al. [2013]	N	N	Y	N	-	Y	Y	-
Volkova et al. [2013]	N	N	Y	Y	-	Y	Y	N
Levitan [2013]	N	N	Y	Y	-	N	N	N
Bojar et al. [2013]	N	Y	N	N	-	-	-	-
Glavaš et al. [2013]	N	Y	N	N	-	-	-	-
Liu et al. [2013]	N	N	N	N	-	-	-	-
Kestemont [2014]	N	N	N	Y	-	-	-	-
Novák and Žabokrtský [2014]	Y	Y	N	Y	-	-	-	-
Babych et al. [2014]	N	Y	N	N	-	-	-	-
Soler-Company and Wanner [2014]	N	N	Y	Y	-	Y	Y	N

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Chen and Ng [2014]	Y	Y	Y	Y	N	Y	Y	N
Sap et al. [2014]	N	N	Y	Y	-	Y	Y	-
Nguyen et al. [2014a]	N	N	Y	Y	-	Y	Y	N
Prabhakaran et al. [2014]	N	N	Y	Y	-	Y	Y	N
Sidorov et al. [2014]	N	N	Y	Y	-	Y	Y	N
Darwish et al. [2014]	N	Y	N	N	-	-	-	-
Ahmed Khan [2014]	N	Y	N	N	-	-	-	-
Nguyen et al. [2014b]	N	N	Y	N	-	Y	Y	-
Stewart [2014]	N	N	Y	Y	-	Y	Y	-
Matthews et al. [2014]	N	Y	N	N	-	-	-	-
Vaidya et al. [2014]	N	Y	N	N	-	-	-	-
Kokkinakis et al. [2015]	N	Y	Y	N	N	Y	-	-
Johannsen et al. [2015]	N	N	Y	Y	-	Y	Y	-
Schwartz et al. [2015]	N	N	N	Y	-	-	-	-
Hovy [2015]	N	N	Y	Y	-	Y	Y	N
Agarwal et al. [2015]	N	Y	Y	Y	N	Y	Y	N
Preoţiuc-Pietro et al. [2015]	N	N	Y	Y	N	Y	Y	-
Ramakrishna et al. [2015]	N	Y	Y	Y	N	Y	Y	N
Taniguchi et al. [2015]	N	N	Y	Y	-	N	Y	N
Schofield and Mehr [2016]	N	N	Y	Y	-	Y	Y	N
Levitan et al. [2016]	N	N	Y	Y	-	Y	Y	N

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Flekova et al. [2016]	N	N	Y	Y	-	Y	Y	N
Tran and Ostendorf [2016]	N	N	N	Y	-	-	-	-
Qian et al. [2016]	N	Y	N	Y	-	-	-	-
Li et al. [2016]	N	N	Y	Y	-	Y	Y	N
Zhang et al. [2016]	N	N	Y	Y	-	Y	Y	N
Garimella and Mihalcea [2016]	N	N	Y	Y	-	Y	Y	N
Reddy and Knight [2016]	N	N	Y	Y	-	Y	Y	N
Li and Dickinson [2017]	N	N	Y	N	-	Y	Y	-
Pérez Estruch et al. [2017]	N	N	Y	Y	-	Y	Y	N
Pérez-Rosas et al. [2017]	N	N	Y	Y	-	Y	Y	N
Rabinovich et al. [2017]	N	N	Y	N	-	Y	Y	-
Costa-jussà [2017]	N	Y	N	N	-	-	-	-
Sap et al. [2017]	N	N	Y	Y	-	Y	-	-
Zhao et al. [2017]	N	N	Y	Y	-	Y	Y	N
Mandravickaitė and Krilavičius [2017]	N	N	Y	Y	-	Y	Y	N
Verhoeven et al. [2017]	N	N	Y	Y	-	Y	Y	N
Larson [2017b]	N	Y	Y	Y	Y	N	N	Y
Koolen and van Cranenburgh [2017]	N	N	Y	N	-	N	Y	-
Tatman [2017]	N	N	Y	Y	-	Y	Y	N
Soler-Company and Wanner [2017]	N	N	Y	Y	-	Y	Y	N
Ljubešić et al. [2017]	N	N	Y	N	-	Y	Y	-

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Litvinova et al. [2017]	N	N	Y	N	-	Y	Y	-
Mohammad et al. [2018]	N	N	Y	Y	-	Y	-	-
Wang and Jurgens [2018]	N	Y	Y	Y	Y	N	N	N
Kraus et al. [2018]	N	N	Y	Y	-	Y	-	-
Martinc and Pollak [2018]	N	N	Y	Y	-	Y	Y	N
Chan and Fyshe [2018]	N	N	Y	Y	-	Y	Y	N
Durmus and Cardie [2018]	N	N	N	Y	-	-	-	-
Zaghouani and Charfi [2018]	N	Y	Y	N	N	Y	Y	-
Plank [2018]	N	N	Y	Y	-	Y	Y	N
Wood-Doughty et al. [2018]	N	N	Y	Y	-	Y	Y	N
Moorthy et al. [2018]	N	N	Y	Y	-	Y	-	-
Levitan et al. [2018]	N	N	Y	Y	-	Y	Y	N
Webster et al. [2018]	Y	Y	Y	Y	N	Y	Y	N
Park et al. [2018]	N	Y	Y	Y	N	Y	Y	N
Vanmassenhove et al. [2018]	N	Y	Y	N	N	Y	Y	-
Kleinberg et al. [2018]	N	N	Y	Y	-	Y	Y	N
Zhao et al. [2018b]	N	N	Y	Y	-	Y	Y	N
Balusu et al. [2018]	N	N	N	Y	-	-	-	-
Rudinger et al. [2018]	Y	Y	Y	Y	N	N	-	Y
Zhao et al. [2018a]	Y	Y	Y	Y	N	Y	Y	N
Kiritchenko and Mohammad [2018]	-	-	-	-	-	-	-	-

Citation	Coref	L.G	S.G	Eng	L≠S	0/1	Imm	Neo
Barbieri and Camacho-Collados [2018]	N	N	Y	Y	-	Y	N	-
van der Goot et al. [2018]	N	N	Y	N	-	Y	Y	-
Karlekar et al. [2018]	N	N	Y	Y	-	Y	Y	N
de Gibert et al. [2018]	N	N	N	Y	-	-	-	-
Mickus et al. [2019]	N	Y	N	N	-	-	-	-

A.2 Example GICoref Document from Wikipedia: Dana Zzyym

[[Source: https://en.wikipedia.org/wiki/Dana_Zzyym]]

Dana Alix Zzyym$_A$ is an Intersex activist and former sailor who was the first military veteran in the United States to seek a non - binary gender U.S. passport , in a lawsuit Zzyym$_A$ v. Pompeo$_C$.

Early life

Zzyym$_A$ has expressed that their$_A$ childhood as a military brat made it out of the question for them$_A$ to be associated with the queer community as a youth due to the prevalence of homophobia in the armed forces . Their$_A$ parents$_B$ hid Zzyym$_A$'s status as intersex from them$_A$ and Zzyym$_A$ discovered their$_A$ identity and the surgeries their$_A$ parents$_B$ had approved for them$_A$ by themselves$_B$ after their$_A$ Navy service . In 1978 , Zzyym$_A$ joined the Navy as a machinist 's mate .

Activism

Zzyym$_A$ has been an avid supporter of the Intersex Campaign for Equality .

Legal case

Zzyym$_A$ is the first veteran to seek a non - binary gender U.S. passport . In light of the State Department 's continuing refusal to recognize an appropriate gender marker , on June 27 , 2017 a federal court granted Lambda Legal 's motion to reopen the case . On September 19 , 2018 , the United States District Court for the District of Colorado enjoined the U.S. Department of State from relying upon its binary - only gender marker policy to withhold the requested passport .

A.3 Example GICoref Document from AO3: Scar Tissue

[[Source: https://archiveofourown.org/works/14476524]]

[[Author: cornheck]]

Despite dreading their$_A$ first true series of final exams , Crona$_A$'s relieved to have a particularly absorbative memory , lucky to recall all the material they$_A$ 'd been required to catch up on . Half a semester of attendance , a whole year of course content .

The only true moment of discomfort came when they$_A$ 'd arrived at the essay portion . Thankful it was easy enough to answer , however , their$_A$ subtle eye - roll stemmed entirely from just how much writing it asked of them$_A$, hands already beginning to ache at the thought of scrawling out two pages on the origins , history , and importance of partnered and grouped soul resonance .

By the end of it all , their$_A$ neck , wrist , back , and ribs ached from the strain of their$_A$ typical , hunched posture – a habit they$_A$ defaulted to , and Miss Marie$_B$ silently wished they$_A$ 'd be more mindful of . It was a relief , at least to them$_A$, not to be the last one out of the lecture hall . Booklet turned in , they$_A$ left the room as quietly as possible and lingered just outside , an air of hesitance settling upon them$_A$ as they$_A$ considered what to do now that , it seemed , everything was over with . No more class , no more lessons , just ... students on break from their studies for the season .

“ Kind of a breeze , was n't it ? ” Evans$_C$ ' voice echoes in the arched hall and Crona$_A$'s shoulders jump , their$_A$ frame still a tense and anxious mess .

“ Oh , ” they$_A$ sigh , “ I$_A$ ... I$_A$ suppose so . It was n't ... necessarily hard . ” Crona$_A$ answers , putting forth a vaguely forced smile .

Smiling with the assumed purpose of making Soul$_C$ comfortable with the interaction . A defense mechanism .

“ I$_A$ - I$_A$ guess , for a final , it was easier than I$_A$ expected ... everyone ... made it sound like it 'd be difficult . ”

“ If by everyone , you mean Black Star , then yeah , ” Soul chuckles , “ he does n’t really do well on ‘ em ... bad test - taker . ”

“ Ah , ” their facade falls just in time to be replaced by a much more genuine grin .

Of the little they ’d spent talking to Black Star , he certainly had confidence and skill enough to make up for the lost exam points given his performance in every other grading category .

“ That ... makes sense . ”

“ Maka ’s always the first one done when it comes to this stuff , she practically studies in her sleep . I ’m convinced she must be practicing clairvoyance the way she burns through essay questions , ” Soul laughs , turning to the meek teen who gives him a simple nod in response .

Determined not to let an impending awkward silence fall between them , Soul pipes up again , “ So , are you staying here for break ? ”

“ Ye - well , I ... I think so , ” they begin , stuttering , but encouraged to continue by a cock of Soul ’s head ; a social cue even they could read , “ The professor ... and Miss Marie asked if I ’d like to come and stay with them for the time being . ”

“ Oh , huh , Stein and Marie ? Nice , ” his brows lift , clearly some varying degree of happy for the other .

The optimism is short - lived , observing as Crona ’s expression falls back to its characteristic expressionless gaze .

“ It seems like you ’ve got a good thing going with those two . ”

“ I have n’t decided , yet , if I should accept the invitation , ” they shift a bit where they stand .

Never having been the best at reassuring others , even his own meister , Soul kept his mouth shut to avoid stuttering while he searched for the right words a web of thoughts .

“ Y ’ know , I think it ’s less of an invitation and more of an extended welcome . ”

The other raises their head , taken aback , “ Oh , ” Crona mutters , in a poignant tone , “ I ... never considered something like that . ”

Soul does n’t leave much wiggle room for their mood to fall any further (nothing past a flat - lipped frown) , “ They ’d probably love to have you , I bet they drive each other nuts sometimes all

by themselves . ”

Though Evans wo n’t admit it , he knows it ’s all too likely Stein might actually put some more effort into taking care of himself if he had someone else besides Marie to look after .

“ I - I see , ” they exhale with a nod , giving Soul a hint of affirmation that he ’d done something to boost the kid ’s confidence .

“ I mean , it ’s got ta be lonely not to mention boring hanging here all summer ... and the weather , ” Soul nearly gasps , dramatizing it for added effect , “ Oh , man , I do n’t know how you can stay cooped up in that room of yours when it ’s so nice out , ” he grins .

“ But ... meh . Different strokes . I ca n’t judge . ”

His comments comfort them , an for a moment they forget how this came to be . The cathedral in Italy , Lady Medusa ’s wrath , and the black blood that infected him . Every moment they spent in the presence of Soul Evans builds always up to this ; fixation on the memories of their first encounters and all the pain they ’ve caused him , the pain they ’ve caused he and Maka both . As quickly as Soul had lifted the swordsman ’s spirits , they ’d weighed themselves down once more . It seemed so normal , though . Soul could n’t bring himself to feel any sense of accomplishment in the coaxing - out of Crona ’s smile when the return of their self doubt was as certain as the sun in the sky . His own stubbornness could n’t let his diminished self worth lie . With another encouraging smile , rows of sharpened incisors appearing oddly charismatic , he opens his mouth to speak – but finds himself cut off before he can even squeeze a word in .

“ Soul , I ’m sorry , ” the meister blurts .

Having been pent - up for months , the apology comes forth without inhibition , rolling effortlessly off their tongue .

“ Sorry ... ? For what ? ” Evans quirks a brow , chuckling .

He adjusts his stance to face Crona with the whole of his body , maintaining his positive demeanor .

“ F - for what ... ? ”

They stammer , shaking their head . For all their remorse , they thought this would have been

obvious .

“ For everything , it ’s ... the first time we$_F$ dueled , I$_A$ was the enemy ! I$_A$ - I$_A$ almost killed you$_C$, I$_A$ - I$_A$ ... I$_A$ really , really hurt you$_C$, ” they$_A$ answer , still so sick with guild that even their$_A$ confession of responsibility is tainted with frustration .

Soul$_C$ seems stunned for a moment before harnessing his$_C$ quick wit .

“ Hey , now , you$_A$ ca n’t take all the credit like that , Ragnarok$_L$ did most of the damage , ” he$_C$ . . .

Appendix B:

B.1 LLMs Performances under Zero-shot and Few-shot Settings

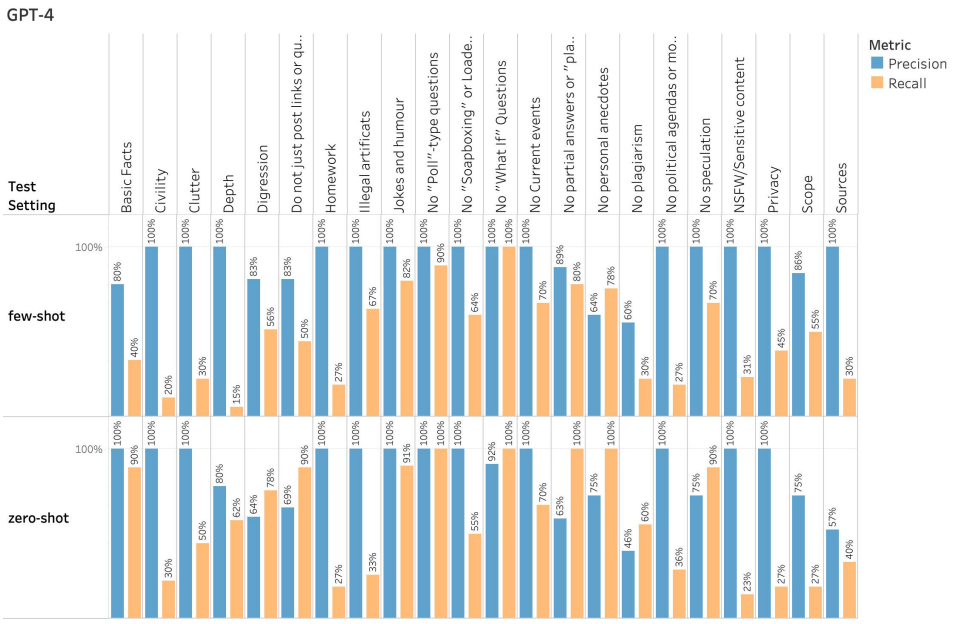

Figure B.1: Model performance on detecting violations of moderation rules with GPT-4 model under few-shot setting (top) and zero-shot setting (bottom). The x-axis is the moderation rules. Each bar pair is the precision (left) and recall (right) scores on the specific rule.

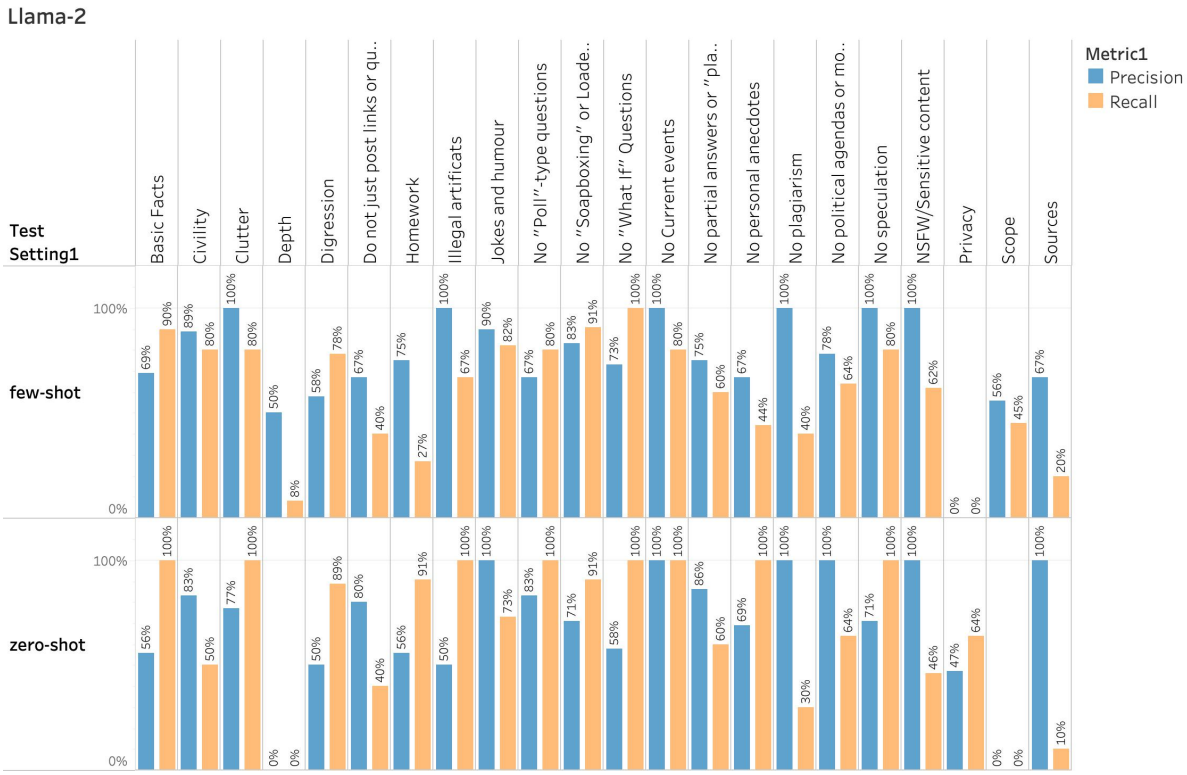

Figure B.2: Model performance on detecting violations of moderation rules with Llama-2 model under few-shot setting (top) and zero-shot setting (bottom). The x-axis is the moderation rules. Each bar pair is the precision (left) and recall (right) scores on the specific rule.

B.2 User Survey

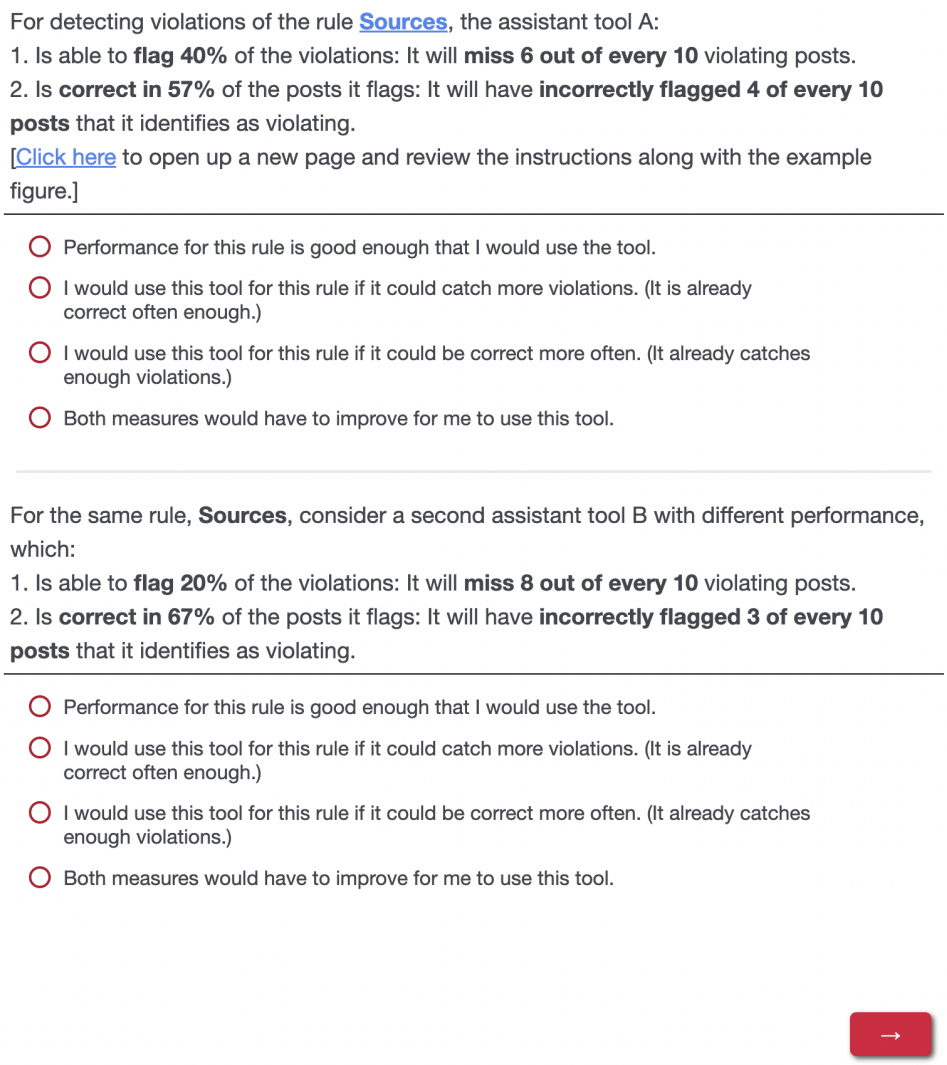

Figure B.3: The evaluation survey question for the rule `Sources`. The question content is similar for all the rules.

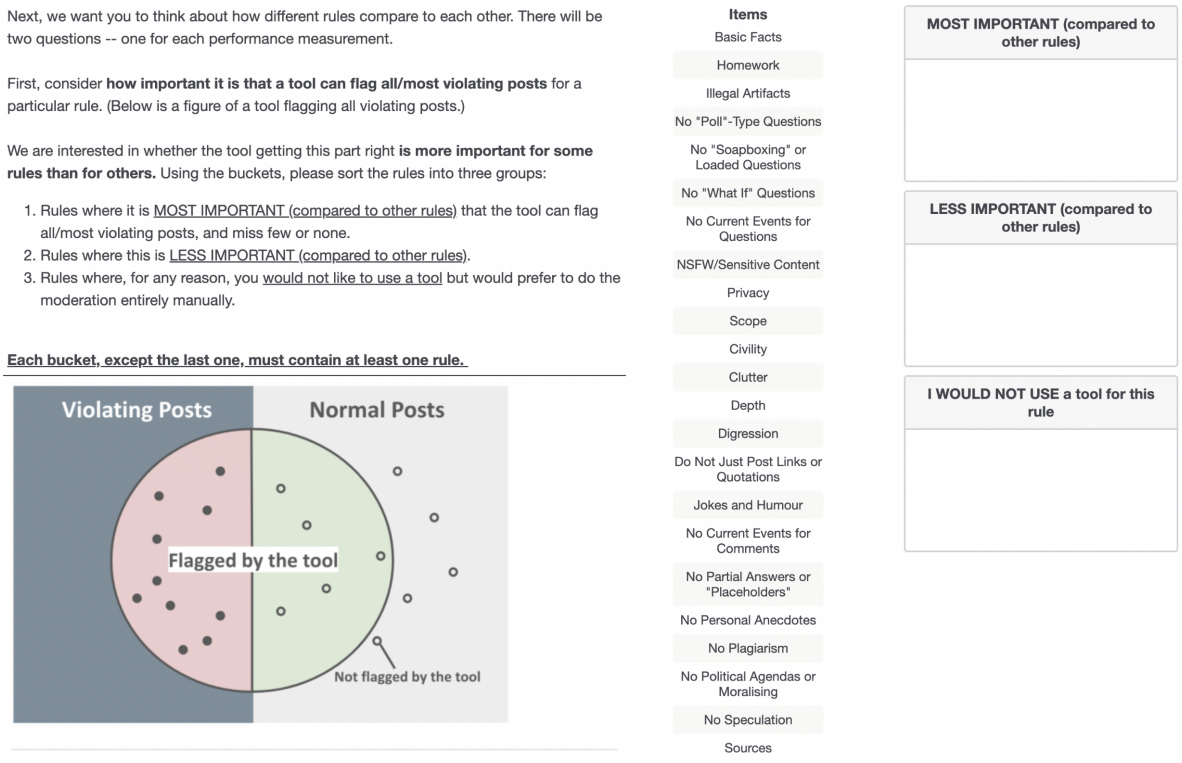

Figure B.4: The clustering survey question for recall importance. The same question is asked for precision importance.

Categories of Rules	AskHistorian	Atheism	Movies
Advertising & Commercialization			No Spam & Self-promotion - Same Source Posts
Consequences/ Moderation/ Enforcement			
Content/Behavior	Homework; No "what if" questions Basic facts Depth; No speculation; Jokes and humor	Proselytizing	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs; No Extraneous Comic Book Movie Submission"
Copyright	No plagiarism		
Doxxing/Personal Info	Privacy; No personal anecdotes	Personal Attacks or Flaming	
Format	No partial answers or "placeholders"; Do not just post links or quotations		No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title
Harassment	Civility	Harassment or Bigotry	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
Hate Speech	Civility		No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
Images			
Links & Outside Content	Sources		Prohibited Popular Websites
Low-Quality Content	Scope; Clutter	Low-Effort Posts	
NSFW	NSFW/Sensitive content		
Off-topic	Digression	Off Topic	No Extraneous Comic Book Movie Submission
Personal Army		Brigading	No Subreddit Brigading
Personality	Civility	Personal Attacks or Flaming	No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dogwhisteling/ Homophobic Slurs
Politics	No political agendas or moralising; No "Soapboxing" or loaded questions		
Prescriptive			
Reddiquette			
Reposting			No Repost or Discussion Threads of New Releases
Restrictive			
Spam		Spam	No Spam & Self-promotion - Same Source Posts
Spoilers			No Repost or Discussion Threads of New Releases; No Spoilers
Trolling		No Trolling	No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title
Voting	"Poll-type" question		
Others	Current event; Illegal artifacts	Don't complain about the use of AAVE or slang	No Encouraging Piracy

Table B.1: Subreddit rules from AskHistorian, Atheism, and Movies subreddits and their cross-

Rule	Rule Description
Civility	All users are expected to behave with courtesy and politeness at all times. We will not tolerate racism, sexism, or any other forms of bigotry. This includes Holocaust denialism. Nor will we accept personal insults of any kind, and do not allow minor nitpicking of grammar or spelling.
Scope	Submissions must be about a question about the human past, a META post about the state of the subreddit, or an AMA ("Ask Me Anything") with a historical expert or panel of experts.
No Current events	To discourage off-topic discussions of current events, questions, answers, and all other comments must be confined to events that happened 20 years ago or more, inclusively (e.g., 2003 and older).
Homework	Our users aren't here to do your homework for you, but they might be willing to help. Don't just give us your essay/assignment topic and ask us for ideas.
NSFW/Sensitive content	Questions with NSFW titles will be deleted, and we will ask you to repost it with a different title.
No "Poll"-type questions	"Poll"-type questions aren't appropriate here: "Who was the most influential person in history?" or "Who was the worst general in your period?" or "Who are your Top 10 favorite people in history?" If your question includes the words "most" or "least", or "best" or "worst" (or can be reworded to include these words), it's probably a "poll"-type question.
No "Soapboxing" or Loaded Questions	All questions must allow a back-and-forth dialogue based on the desire to gain further information and not be predicated on a false and loaded premise in order to push an agenda. Example: Good Question: "People say that Nixon is the worst President of all time. Why is this so?" Bad Question: "Nixon was the worst President of all time. Why isn't Obama considered the worst?" The bad question is a fishing expedition to try to start a debate about Obama's presidency. Most of these questions will break our 20-year rule, or try to set up a debate about an issue using a long wall of text in the main post.
No "What If" Questions	Questions should be about what did happen, not what could have happened.
Privacy	We do not allow questions that pose possible privacy issues for living or recently deceased persons who are not in the public eye. The cut-off for "recent" is 100 years.
Illegal artifacts	It is our policy to disallow posts asking for further information on artifacts where there is a likelihood that the acquisition or possession of the item might be illegal, unethical, and/or run contrary to sound, historical practices.
Basic Facts	Questions looking for specific, basic facts - for the purpose of this rule, seeking a name, a date or time, a number, a location, the origin of a word, the first or last known instance/example of an object/phenomenon/etc., or a simple list of examples or facts - are not allowed as standalone threads.
Depth	An in-depth answer provides the necessary context and complexity that the given topic calls for, going beyond a simple cursory overview. Your answer should be giving context to the events being discussed, not simply listing some related facts.
Sources	We do not require sources to be preemptively listed in an answer here, but do expect that respondents be familiar with relevant and reliable literature on the topic, and that answers reflect current academic understanding or debates on the subject at hand. But sole reliance on tertiary sources for context and analysis is not allowed, and will result in the removal of a response.
No personal anecdotes	Personal anecdotes are not acceptable answers in this subreddit.
No speculation	Suppositions and personal opinions are not a suitable basis for an answer here. Warning phrases for speculation include: "I guess..." or "My guess is...", "I believe...", "I think...", "... to my understanding", "It makes sense to me that...", "It's only common sense."
No partial answers or "placeholders"	An answer should be full and complete in and of itself.
No political agendas or moralizing	This subreddit is a place for learning and open-minded discussion. As such, answers should not be written in the interests of advancing a personal agenda, but should represent a sincere effort to make an argument from the historical record. They should be constructed in keeping with the principles of the historical method - that is to say, your evidence should not be chosen selectively to support an argument that you feel is right; your argument should instead demonstrably flow from

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No Trolling	trolling; provocative	TRAC2020_IBEN_A_bert-base-multilingual-uncased	No	Yes	Yes		
Personal Attacks or Flaming	doxxing; cyberbullying	TRAC2020_IBEN_A_bert-base-multilingual-uncased	No	Yes	Yes		
Off Topic	topic classifier	tweet-topic-21-multi	No	No	Yes, but with changes	Twitter topic classification	Need to be finetuned on this topic
Image not submitted correctly							
Spam	spam	bert-tiny-finetuned-sms-spam-detection	No	No	Yes, but with changes	Detect message spam	1. Need to be adapted to reddit texts rather than message texts. 2. Need to be adapted to detect self-promotion of some youtube account/website etc.
Low-Effort Posts	low effort; low quality	NA	No	No	No		
Proselytizing	proselytizing; Soapbox	NA	No	No	No		
Harassment or Bigotry	harassment; bigotry	Text-Moderation	No	Yes	Yes, but with changes		any and all curse words are permitted
Brigading	brigading	NA - Toxicity	No	Yes	Yes, but with changes		These models may be able to detect some of the brigading comments, but there is no specific model for brigading and some brigading encouraging comments may not be caught by the toxicity models

Table B.3: Model review for the Atheism subreddit. The second and third columns are the keywords used for model searching and the best-matching model from hugging face. The rest columns are annotation results. Crossed-out rules are rules not related to text input.

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
Do Familiarize Yourself With Our Rules							
Incivility - No Antagonism/ Flamewars/ Attention Whoring - No Racial/ Sexist/ Dog-whistling/ Homophobic Slurs	offensive; hatespeech	Text-Moderation	No	Yes	Yes, but with changes	Yes	Quoting something offensive from a movie is okay but needs to be put in quotation marks.
No Spam & Self-promotion - Same Source Posts	spam	bert-tiny-finetuned-sms-spam-detection	No	No	Yes, but with changes	No	Detect message spam - 1. Need to be adapted to Reddit texts rather than message texts. 2. Need to be adapted to detect self-promotion of some YouTube account/website etc.
No Image Posts & Memes							
No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title - misinformation	misinformation	TwHIN-BERT-Misinformation-Classifier	No	No	Yes, but with changes	No	Detect misinformation for Twitter posts - 1. Need to be adapted to title-like texts. 2. Need to be adapted to a movie-related context
No Ambiguous/ Misleading/ Inaccurate Information or Clickbait in The Submission Title - Clickbait	clickbait	distilroberta-clickbait	No	No	Yes	Yes	
No Extraneous Comic Book Movie Submission	misinformation	TwHIN-BERT-Misinformation-Classifier	No	No	Yes, but with changes	No	Detect misinformation for Twitter posts - Need to be adapted to a movie-related context
No Repost	rule-based		Yes	No			
No Discussion Threads of New Releases	rule-based		Yes	No			
No Spoilers	spoiler	roberta-base-finetuned-imdb-spoilers	No	No	Yes	Yes	

Rules	Keywords	Annotated model	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
Civility	toxicity; hate-speech	Text-Moderation	No	Yes	Yes		The model may not be able to catch Holocaust denialism, which is vital in this case.
Current event	time/year prediction for event; history event; event extractor	GoLLIE-7B	No	No	Yes, but with changes	Information extraction	1. The model needs to be adapted to question-contexts and extract event names from the question. 2. Then by matching the event name with Wikipedia page, we may get the year or time of the event.
Homework	homework	NA	No	No	No		
Scope	topic classifier	tweet-topic-21-multi	No	No	Yes, but with changes	Twitter topic classification	Need to be fine-tuned to this topic.
NSFW/Sensitive content	NSFW; sexual	NSFW_text_classifier	No	No	Yes		

Table B.5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
"Poll-type" question	poll	NA	No	No	No		
"Soapboxing" or loaded questions	soapboxing; agenda pushing	NA	No	No	No		
No "what if" questions	hypothetical; counterfactual; alternative history	NA	No	No	No		

Table B.5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
Privacy	PII; personal private	lgpd_pii_identifier	No	No	Yes, but with changes	Identify sensitive data in the scope of LGPD. The goal is to have a tool to identify document numbers like CNPJ, CPF, people's names, and other kinds of sensitive data, allowing companies to find and anonymize data according to their business needs, and governance rules.	Need to be adapted to the sensitive data here

Table B.5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
Illegal artifacts	illegal; safety	beaver-dam-7b	No	No	Yes		
Basic facts	basic facts; factoid	nf-cats	No	No	Yes		
Depth	depth	NA	No	No	No		
Sources	reliable source	NA	No	No	No		
No personal anecdotes	personal anecdote	muppet-roberta-base-joke_detector	No	No	Yes, but with changes	Detect jokes, stories, and anecdotes	The model is trained with 2000 jokes. Here the detection target is not about jokes. In addition, here the need is focusing on personal anecdotes.

Table B.5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No speculation	speculation; factual; fact-check	verdict-classifier-en	No	No	Yes, but with changes	Fact-checking model for detecting misinformation	The model does not detect speculation or personal opinion. Here the detection target is the speculation of history events, which can be especially hard for such a fact-checking model.
No partial answers or "placeholders"	partial answer	NA	No	No	No		
No political agendas or moralizing	political agenda; moralizing	NA	No	No	No		
Do not just post links or quotations	rule-based	NA	Yes	No			

Table B.5 – *Continued from previous page*

Rules	**Keywords**	**Annotated model**	RULE-BASED?	TOXIC?	MATCH?	PURPOSE?	GAP?
No plagiarism	plagiarism	longformer-base-plagiarism-detection	No	No	Yes, but with changes	Detect machine-paraphrased plagiarisms	The real-world source of plagiarism is way larger than the training and testing dataset size of this model.
Jokes and humor	joke; humor; flippant	muppet-roberta-base-joke_detector	No	No	Yes		
Digression	relevance	response-quality-classifier-large	No	No	Yes, but with changes	Rate the relevancy and specificity of a response within a dialogue	Need to be adapted to the Reddit (question-comment) texts.
Clutter	clutter	NA	No	No	No		

Table B.5: Model review for the AskHistorian subreddit. The second and third columns are the keywords used for model searching and the best-matching model from Hugging Face. The rest of the columns are annotation results. Crossed-out rules are rules not related to text input.

Appendix C:

C.1 Traits

The full list of traits and respective adjectives is in the Table C.1

C.2 Experiment Results with Single Groups

Table C.2 presents the Kendall's τ correlation scores between the model and human at the group level, while Table C.3 and Table C.4 shows the alignment with the precision at 3 scores (former computed with the top 3 traits and latter with the bottom 3 traits).

C.3 Experiment Results of Intersectional Groups

Table C.5 presents the dominating relationship between domains, while Table C.6 lists the average correlation scores of the paired group with each of its identities' domain for each domain pair. Table C.7 shows the top 50 emergent group-trait associations.

C.3.1 Human study setup

The survey for the collection of associated traits is presented in Figure C.1.

Trait	Adjectives
Agency scale	
Powerful	forceful, formidable, capable
Powerless	weak, helpless, incapable
High status	privileged, elite, advantaged
Low status	unskilled, lowly, inferior
Dominant	commanding, authoritative
Dominated	subservient, submissive, deferential
Wealthy	affluent, rich, prosperous
Poor	impoverished, destitute, needy
Confident	self-assured, assured, self-possessed
Unconfident	bashful, meek, timid
Competitive	ambitious, driven, zealous
Unassertive	submissive, diffident, passive
Beliefs scale	
Modern	radical, forward-looking
Traditional	old-fashioned
Science-oriented	analytical, logical, atheistic
Religious	devout, pious, reverent
Alternative	unorthodox, avant-garde, eccentric
Conventional	mainstream
Liberal	left-wing, Democrat, progressive
Conservative	right-wing, Republican
Communion scale	
Trustworthy	reliable, dependable, truthful
Untrustworthy	unreliable, undependable
Sincere	genuine, forthright, honest
Dishonest	insincere, deceitful
Warm	friendly, kind, loving
Cold	unfriendly, unkind, aloof
Benevolent	considerate, generous
Threatening	intimidating, menacing, frightening
Likable	pleasant, amiable, lovable
Repellent	vile, loathsome, nasty
Altruistic	helpful, charitable, selfless
Egotistic	selfish, self-centered, insensitive

Table C.1: Full list of traits and corresponding adjectives

	CEAT		ILPS		ILPS*		SeT	
	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT
White people	0.150	-0.033	-0.117	-0.383	0.117	-0.350	-0.033	-0.217
Hispanic people			0.533	0.200	0.133	0.300	0.483	0.283
Asian people			0.092	0.126	0.159	0.126	0.243	0.326
Black people	-0.209	-0.075	0.209	0.142	0.176	0.042	0.393	0.209
Immigrants	-0.117	-0.267	0.233	0.350	0.217	0.383	0.283	0.400
Men	0.183	-0.033	0.083	0.433	0.233	0.183	0.200	0.383
Women	-0.433	0.083	0.217	0.017	-0.100	0.050	0.083	0.067
Wealthy people	0.100	-0.133	0.067	0.017	0.150	0.167	0.067	0.083
Jewish people	0.250	0.083	0.017	-0.067	0.150	-0.217	0.033	-0.100
Muslim people	0.233	-0.050	0.000	-0.167	0.183	-0.017	0.250	-0.233
Christians	0.343	0.393	0.209	0.075	0.410	-0.176	0.243	0.142
Cis people	0.167	-0.067	-0.167	-0.033	0.217	-0.400	0.050	0.033
Trans people	-0.283	-0.050	0.067	-0.067	0.033	0.083	0.133	0.050
Working class people	0.050	0.300	0.183	-0.117	-0.300	0.017	0.250	-0.033
Nonbinary people			0.050	-0.183	0.117	-0.050	0.067	-0.250
Native Americans	-0.217	-0.017	0.117	0.350	0.000	-0.183	0.200	0.283
Buddhists	0.000	0.300	0.417	0.517	0.483	0.217	0.383	0.533
Mormons	0.167	0.367	-0.033	0.100	0.283	-0.333	-0.083	0.283
Veterans	0.100	0.417	0.250	-0.083	0.267	-0.083	0.217	-0.033
Unemployed people	-0.233	0.083	0.067	0.500	0.067	0.400	0.050	0.500
Teenagers	-0.150	-0.133	0.200	-0.267	0.367	-0.033	0.217	-0.250
Elderly people	0.017	0.417	0.650	0.333	0.533	0.117	0.700	0.400
Blind people	0.017	0.367	0.217	0.267	0.100	0.150	0.200	0.267
Autistic people			0.350	-0.117	0.317	0.250	0.267	-0.050
Neurodivergent people	-0.167	0.000	0.083	-0.017	-0.100	0.050	0.017	-0.117

Table C.2: Overall alignment scores with human annotations for Kendall's τ. There are some missing scores for CEAT because there are no occurrences of these groups in the Reddit 2014 dataset

	CEAT		ILPS		ILPS*		SeT	
	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT
White people	1.00	1.00	0.33	0.33	0.67	0.67	0.67	0.67
Hispanic people			1.00	0.67	0.67	0.67	0.67	0.67
Asian people			1.00	1.00	1.00	1.00	1.00	1.00
Black people	0.00	0.33	0.33	0.33	0.33	0.00	0.67	0.33
Immigrants	0.33	0.00	0.67	0.00	0.33	0.00	0.33	0.33
Men	0.67	0.00	0.67	1.00	0.67	0.33	0.67	1.00
Women	0.33	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wealthy people	1.00	0.67	0.33	0.33	0.67	0.67	0.67	0.67
Jewish people	0.67	0.67	0.00	0.33	0.33	0.33	0.33	0.33
Muslim people	0.00	0.00	0.00	0.00	0.33	0.33	0.33	0.00
Christians	1.00	1.00	1.00	1.00	1.00	0.67	1.00	1.00
Cis people	1.00	1.00	1.00	0.67	1.00	0.67	1.00	1.00
Trans people	0.33	0.33	1.00	0.00	0.67	0.67	1.00	0.33
Working class people	0.67	0.67	0.67	0.33	0.33	1.00	0.67	0.67
Non-binary people			1.00	0.67	1.00	0.67	1.00	0.67
Native Americans	0.33	0.67	0.67	1.00	0.33	0.67	0.67	0.67
Buddhists	0.33	0.67	1.00	1.00	1.00	1.00	0.677	1.00
Mormons	0.67	1.00	1.00	1.00	1.00	0.67	1.00	1.00
Veterans	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Unemployed people	0.33	0.00	0.00	0.67	0.00	0.00	0.00	0.67
Teenagers	0.00	0.33	0.67	0.33	0.67	0.33	0.67	0.67
Elderly people	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Blind people	0.67	0.67	1.00	1.00	0.67	1.00	1.00	1.00
Autistic people			1.00	0.67	1.00	1.00	1.00	0.67
Neurodivergent people	0.33	0.00	0.00	0.33	0.00	0.33	0.00	0.33

Table C.3: Overall alignment scores with human annotations for Precision at the top 3 traits

	CEAT		ILPS		ILPS*		SeT	
	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT
White people	0.67	0.33	0.00	0.00	0.33	0.67	0.67	0.67
Hispanic people			1.00	0.33	1.00	0.67	0.67	0.67
Asian people			0.33	0.00	0.67	1.00	1.00	1.00
Black people	0.33	0.33	1.00	0.67	1.00	0.00	0.67	0.33
Immigrants	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Men	0.33	0.67	0.33	1.00	0.67	1.00	0.67	1.00
Women	0.00	0.33	0.00	0.00	0.00	0.33	0.00	0.00
Wealthy people	0.33	0.00	0.33	0.00	0.33	0.67	0.33	0.00
Jewish people	0.67	0.33	1.00	0.67	1.00	0.00	1.00	0.67
Muslim people	0.67	0.67	0.67	0.33	1.00	1.00	1.00	0.67
Christians	0.67	1.00	0.33	0.33	0.33	0.00	0.33	0.67
Cis people	0.33	0.33	0.00	0.33	0.33	0.00	0.33	0.33
Trans people	0.00	0.67	0.33	0.33	0.33	0.33	0.33	0.33
Working class people	0.67	0.67	0.33	0.33	0.67	0.33	0.33	0.67
Non-binary people			0.00	0.00	0.33	0.67	0.00	0.00
Native Americans	0.33	0.33	0.33	0.67	0.67	0.33	0.67	0.67
Buddhists	0.33	0.67	1.00	1.00	0.33	0.67	1.00	0.67
Mormons	0.67	1.00	0.33	0.33	0.33	0.00	0.33	0.67
Veterans	0.33	0.67	0.67	0.00	0.33	0.33	0.67	0.00
Unemployed people	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Teenagers	0.33	0.33	1.00	0.33	1.00	1.00	0.67	0.00
Elderly people	0.33	1.00	1.00	0.67	1.00	0.33	1.00	1.00
Blind people	1.00	0.67	0.33	0.33	0.67	0.33	0.33	0.33
Autistic people			0.67	0.33	1.00	0.67	0.33	0.33
Neurodivergent people	0.67	0.67	0.67	1.00	0.67	0.67	0.67	0.67

Table C.4: Overall alignment scores with human annotations for Precision at the bottom 3 traits

	Dominates	Dominated by
age	gender/sexuality, race/ethnicity, nationality, politics, religion, socio-economic	-
politics	nationality, socio-economic, disability	age, religion
gender/ sexuality	race/ethnicity, nationality	age
disability	race/ethnicity, nationality	politics
social-economic	race/ethnicity, nationality	age, politics
religion	politics	-
race/ ethnicity	-	age, gender/sexuality, socio-economic, disability
nationality	-	age, gender/sexuality, politics, socio-economic, disability

Table C.5: Domination relations between social domains

Domain A	Domain B	Correlation A	Correlation B
age	disability	0.532	0.475
gender	disability	0.418	0.356
age	gender	0.552	0.320
age	nationality	0.583	0.337
disability	nationality	0.543	0.309
gender	nationality	0.481	0.225
political stance	nationality	0.287	0.179
race	nationality	0.594	0.525
religion	nationality	0.490	0.525
socio	nationality	0.540	0.338
age	political stance	0.319	0.177
disability	political stance	0.019	0.397
gender	political stance	0.315	0.375
race	political stance	0.376	0.348
religion	political stance	0.380	0.271
age	race	0.520	0.395
disability	race	0.538	0.392
gender	race	0.478	0.371
age	religion	0.502	0.449
disability	religion	0.465	0.463
gender	religion	0.439	0.360
race	religion	0.522	0.460
age	socio	0.562	0.406
disability	socio	0.420	0.419
gender	socio	0.374	0.397
political stance	socio	0.433	0.290
race	socio	0.387	0.488
religion	socio	0.404	0.439

Table C.6: Full list of correlations for paired social groups. The table shows two domains, which comprise group AB, correlations between group AB and group A, group AB and group B.

Group AB	Emerged Trait	Increased Score	Max Score
Jamaican mechanic	trustworthy	0.1055	-0.0449
gay with a disability	conventional	0.0931	0.0017
gay with a disability	threatening	0.0922	-0.0316
Hispanic unemployed person	egotistic	0.0919	-0.1546
gay with a disability	liberal	0.0882	0.0401
female Native American	dominant	0.0860	0.0682
Democrat teenager	altruistic	0.0858	-0.0986
Deaf mechanic	likable	0.0854	0.0046
Black mechanic	likable	0.0821	-0.0118
Democrat mechanic	trustworthy	0.0819	-0.0449
male doctor	benevolent	0.0819	-0.0230
female Indian person	dominant	0.0808	0.0471
Latina	dominant	0.0808	0.0720
Filipino mechanic	trustworthy	0.0802	-0.0137
Native American mechanic	trustworthy	0.0796	-0.0449
teenage Democrat	altruistic	0.0794	-0.0986
trans mechanic	likable	0.0792	-0.0118
Democrat mechanic	sincere	0.0792	-0.0205
Democrat teenager	sincere	0.0790	-0.0205
female Black person	dominant	0.0785	0.0471
unemployed Italian person	poor	0.0784	0.0384
female doctor	alternative	0.0779	0.0052
Irish autistic person	egotistic	0.0775	-0.0708
Russian mechanic	likable	0.0773	-0.0118
unemployed Hispanic person	egotistic	0.0772	-0.1546
Russian unemployed person	egotistic	0.0762	-0.1788
female doctor	traditional	0.0750	0.0107
Amish mechanic	trustworthy	0.0748	-0.0170
Republican mechanic	sincere	0.0745	-0.0164
male teenager	conventional	0.0738	-0.0589
Hispanic French person	egotistic	0.0733	-0.1210
Cuban person with a disability	poor	0.0731	0.0486
atheist mechanic	trustworthy	0.0727	-0.0381
Hispanic Irish person	egotistic	0.0725	-0.1322
female Indian person	dominated	0.0721	0.0421
gay with a disability	traditional	0.0717	0.0229
unemployed German person	poor	0.0715	0.0384
female American person	dominated	0.0709	0.0328
Irish mechanic	trustworthy	0.0709	-0.0300
Muslim autistic person	egotistic	0.0708	-0.0708
male teenager	traditional	0.0705	-0.0490
Russian autistic person	egotistic	0.0704	-0.0708
Japanese autistic person	egotistic	0.0700	-0.0708
trans Republican	sincere	0.0698	-0.0164
German White person	egotistic	0.0696	-0.0833
male Buddhist	benevolent	0.0696	-0.0148
Irish Deaf person	egotistic	0.0693	-0.0589
Native American mechanic	sincere	0.0690	-0.0249
German Republican	egotistic	0.0688	-0.0517

Table C.7: Top 50 emergent group-trait associations

	Social Group			
Trait pair	**Women**	**Men**	**White**	**Black**
powerless-powerful	46.8	81.4	80.7	37.1
low status-high status	44.9	76.3	78.6	25.5
dominated-dominant	34.3	84.8	72.6	26.3
poor-wealthy	55.2	67.7	76.6	28.8
unconfident-confident	57.3	78.3	77.4	54.7
unassertive-competitive	53.8	75.5	79.3	49.9
traditional-modern	61.8	53.3	60.8	31.7
religious-science oriented	59.9	56.1	52.8	27.0
conventional-alternative	55.3	46.7	47.1	44.2
conservative-liberal	61.7	40.8	43.0	56.8
untrustworthy-trustworthy	52.2	50.9	58.2	29.9
dishonest-sincere	52.4	45.3	56.6	37.4
cold-warm	53.8	42.3	56.8	53.0
threatening-benevolent	64.3	39.7	54.2	31.4
repellent-likable	65.5	59.7	59.1	40.3
egoistic-altruistic	50.1	42.8	50.6	47.5

Table C.8: Group-trait associations from White annotators for a subset of social groups. Scores that are closer to 0 indicate closer to the trait on the left (powerless, low status, etc.) and scores closer to 100 indicate closer to the trait on the right (powerful, high status, etc.).

C.3.2 Comparison of Results Across Race and Gender Demographics

Four tables C.8, C.9, C.10, C.11 show how perceptions of "White people", "Black people", "White men", and "White women" differ from each other across annotator demographics. In Table C.12, we show correlation scores for all social groups and overall score between the model and Black, White, White female, and White male annotators. We may see that for certain social groups, the model has better alignment with White people rather than Black people. However, we may see that overall the model doesn't correlate well with human annotators.

	Social Group			
Trait pair	**Women**	**Men**	**White**	**Black**
powerless-powerful	61.0	93.0	73.8	56.6
low status-high status	67.8	86.0	74.3	49.3
dominated-dominant	56.0	94.0	72.5	55.3
poor-wealthy	59.0	91.0	76.8	40.6
unconfident-confident	82.3	85.0	69.7	75.9
unassertive-competitive	54.0	57.0	80.5	76.3
traditional-modern	64.8	67.0	80.3	53.7
religious-science oriented	35.5	65.0	81.8	21.7
conventional-alternative	66.0	62.0	52.5	57.9
conservative-liberal	71.3	82.0	71.5	67.7
untrustworthy-trustworthy	78.5	57.0	62.8	46.9
dishonest-sincere	78.5	61.0	62.3	42.7
cold-warm	87.5	66.0	50.7	58.3
threatening-benevolent	78.3	38.0	35.5	49.7
repellent-likable	85.0	59.0	49.3	62.1
egoistic-altruistic	80.8	77.0	59.8	39.6

Table C.9: Group-trait associations from Black annotators for a subset of social groups. Scores which are closer to 0 indicate closer to the trait on the left (powerless, low status, etc.) and scores closer to 100 indicate closer to the trait on the right (powerful, high status, etc.).

	Social Group			
Trait pair	**Women**	**Men**	**White**	**Black**
powerless-powerful	37.5	80.0	81.9	29.8
low status-high status	44.0	77.0	83.4	18.3
dominated-dominant	42.0	83.3	69.8	18.0
poor-wealthy	47.0	70.5	83.0	12.5
unconfident-confident	55.5	75.5	81.6	51.0
unassertive-competitive	61.0	83.3	82.3	39.0
traditional-modern	59.5	59.3	76.8	26.3
religious-science oriented	46.0	62.5	61.3	21.5
conventional-alternative	51.0	55.0	64.6	42.3
conservative-liberal	54.0	36.7	55.1	53.0
untrustworthy-trustworthy	49.5	45.7	47.5	32.5
dishonest-sincere	48.0	42.5	52.5	34.0
cold-warm	50.0	43.0	55.6	48.0
threatening-benevolent	56.5	34.0	48.3	24.0
repellent-likable	50.5	57.3	57.0	40.5
egoistic-altruistic	51.5	44.8	47.6	53.8

Table C.10: Group-trait associations from White male annotators for a subset of social groups. Scores which are closer to 0 indicate closer to the trait on the left (powerless, low status, etc.) and scores closer to 100 indicate closer to the trait on the right (powerful, high status, etc.).

	Social Group			
Trait pair	**Women**	**Men**	**White**	**Black**
powerless-powerful	48.1	82.8	81.8	41.3
low status-high status	45.1	75.5	76.8	29.6
dominated-dominant	33.2	86.2	78.1	31.0
poor-wealthy	56.4	64.8	73.5	38.1
unconfident-confident	57.5	81.7	76.2	56.9
unassertive-competitive	52.8	67.7	78.9	56.9
traditional-modern	62.1	47.2	51.0	34.9
religious-science oriented	58.5	49.7	50.6	30.2
conventional-alternative	55.9	38.3	37.4	45.3
conservative-liberal	62.8	45.0	38.6	59.0
untrustworthy-trustworthy	52.6	56.2	61.0	28.4
dishonest-sincere	53.1	48.2	53.9	39.1
cold-warm	54.3	41.7	51.4	55.9
threatening-benevolent	65.4	45.3	53.4	35.6
repellent-likable	67.7	62.0	53.3	40.1
egoistic-altruistic	49.9	40.7	47.7	44.0

Table C.11: Group-trait associations from White female annotators for a subset of social groups. Scores which are closer to 0 indicate closer to the trait on the left (powerless, low status, etc.) and scores closer to 100 indicate closer to the trait on the right (powerful, high status, etc.).

Figure C.1: Example of the survey for one group

Trait pair	Social Group			
	Black	**White**	**White Men**	**White Women**
White person	-0.130	0.080	-0.180	0.220
Hispanic person	0.360	**0.470**	0.200	**0.570**
Asian person	**0.560**	0.100	0.190	0.050
Black person	**0.470**	0.370	0.250	0.370
immigrant	0.010	**0.420**	0.300	**0.420**
man	-0.130	0.220	0.180	0.320
woman	-0.060	-0.030	0.080	-0.080
wealthy person	-0.600	0.050	0.050	0.080
Jewish person	0.020	-0.020	-0.120	0.070
Muslim person	——	0.230	0.140	0.280
Christian	0.270	**0.390**	0.280	0.010
cis person	-0.840	0.090	-0.020	0.170
trans person	0.190	0.150	0.180	0.120
working class person	0.010	0.290	0.290	0.220
non-binary	-0.040	0.050	-0.030	0.120
Native American	0.140	0.070	0.080	0.130
Buddhist	0.230	0.320	0.250	0.320
Mormon	-0.030	0.030	0.100	-0.180
veteran	0.220	0.200	0.180	0.190
unemployed person	0.030	0.020	-0.040	0.000
teenager	0.200	0.200	0.220	0.130
elderly person	**0.540**	**0.650**	**0.710**	**0.620**
blind person	0.226	0.217	0.217	0.217
autistic person	0.267	0.217	0.267	0.167
neurodivergent person	0.092	0.050	0.092	0.033
overall	**0.151**	**0.187**	**0.177**	**0.164**

Table C.12: Correlation scores between the model and White, Black, White male, and White female annotators. Scores with p-values less than 0.05 are marked bold.

	CEAT		**ILPS**		**ILPS***		**SeT**	
	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT
Kendall's τ	0.028	0.123†	0.142†	0.071	0.173†	-0.007	**0.174**†	0.093

Table C.13: Overall alignment scores with human annotations with only test groups. The highest scores are bold for each row. For correlation scores, we mark scores where the p-value is < 0.05 with †.

Bibliography

Han-Wei Liu, Ching-Fu Lin, and Yu-Jie Chen. Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology*, 27(2):122–141, 02 2019a. ISSN 0967-0769. doi: 10.1093/ijlit/eaz001. URL https://doi.org/10.1093/ijlit/eaz001.

Charnpreet Kaur and Urvashi Garg. Artificial intelligence techniques for cancer detection in medical image processing: A review. *Materials Today: Proceedings*, 81:806–809, 2023. ISSN 2214-7853. doi: https://doi.org/10.1016/j.matpr.2021.04.241. URL https://www.sciencedirect.com/science/article/pii/S2214785321031618. International Virtual Conference on Sustainable Materials (IVCSM-2k20).

Ravi Kokku, Sharad Sundararajan, Prasenjit Dey, Renuka Sindhgatta, Satya Nitta, and Bikram Sengupta. Augmenting classrooms with ai for personalized education. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6976–6980, 2018. doi: 10.1109/ICASSP.2018.8461812.

Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider,

Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3):391–437, 2023. doi: 10.1080/10447318.2022.2153320. URL https://doi.org/10.1080/10447318.2022.2153320.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. 2023.

Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *AI SOCIETY*, 35, 12 2020. doi: 10.1007/s00146-020-00950-y.

Patrick Shafto, Noah D. Goodman, and Michael C. Frank. Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, 7(4):341–351, 2012. ISSN 17456916, 17456924. URL http://www.jstor.org/stable/41613573.

Wenlong Sun, O. Nasraoui, and Patrick Shafto. Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE*, 15, 2020. URL https://api.semanticscholar.org/CorpusID:221125180.

United Nations Department of Economic and Social Affairs. The 17 goals, 2018. URL https://sdgs.un.org/goals.

Ge Wang. Humans in the loop: The design of interactive ai systems. *Stanford HAI*, Oct 2019. URL https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

Ben Shneiderman. Design lessons from ai's two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1(2):73–82, 2020. doi: 10.1109/TTS.2020.2992669.

ISO. Iso 9241-11:1998(en), last visited 14/05/2015. 1998. URL: https://www.iso.org/obp/ui/iso:std:iso:9241:-11:ed-1:v1:en.

Cynthia Rudin and Joanna Radin. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, 1 (2), nov 22 2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

Sushil K. Oswal. Breaking the exclusionary boundary between user experience and access: steps toward making ux inclusive of users with disabilities. In *Proceedings of the 37th ACM International Conference on the Design of Communication*, SIGDOC '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367905. doi: 10.1145/3328020.3353957. URL https://doi.org/10.1145/3328020.3353957.

Long Cheng, Fang Liu, and Danfeng Daphne Yao. Enterprise data breach: causes, challenges, prevention, and future directions: Enterprise data breach. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7:e1211, 06 2017. doi: 10.1002/widm.1211.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3): 330–347, jul 1996a. ISSN 1046-8188. doi: 10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

Sigal Samuel. Ai's islamophobia problem. *Vox*, Sep 2021. URL https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim.

Timothy J. Hazen, Alexandra Olteanu, Gabriella Kazai, Fernando Diaz, and Michael Golebiewski. On the social and technical challenges of web search autosuggestion moderation. *arXiv:2007.05039 [cs]*, Jul 2020. URL http://arxiv.org/abs/2007.05039. arXiv: 2007.05039.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Beyond the surface: A global-scale analysis of visual stereotypes in text-to-image generation. *ArXiv*, abs/2401.06310, 2024a. URL https://api.semanticscholar.org/CorpusID:266977586.

Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *CoRR*, abs/2011.08512, 2020. URL https://arxiv.org/abs/2011.08512.

Incident Database. Ai incident database, 2024. URL https://incidentdatabase.ai/.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.415. URL https://aclanthology.org/2023.emnlp-main.415.

Sarah A Gilbert. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a public scholarship site on Reddit: A case study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020.

Donghee Yvette Wohn. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

Bryan Dosono and Bryan Semaan. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

Conversational AI Jigsaw. Jigsaw unintended bias in toxicity classification. *Kaggle*, Nov 2019a. URL https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data.

John Pavlopoulos, Jeffrey Scott Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:219176615.

Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2117–2126, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481291. URL https://doi.org/10.1145/2470654.2481291.

Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–31, oct 2020. doi: 10.1145/3415220. URL https://doi.org/10.1145%2F3415220.

Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish The Poor*. St. Martin's Press, 2017. ISBN 978-1-250-07431-7. URL https://www.amazon.com/Automating-Inequality-High-Tech-Profile-Police/dp/1250074312.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in

language model decisions. (arXiv:2312.03689), December 2023. doi: 10.48550/arXiv.2312.03689. URL http://arxiv.org/abs/2312.03689. arXiv:2312.03689 [cs].

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. The abc of stereotypes about groups: Agency/socioeconomic success, conservative-progressive beliefs, and communion. *Journal of personality and social psychology*, 110:675–709, 05 2016. doi: 10.1037/pspa0000046.

Benjamin Müller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *CoRR*, abs/2010.12858, 2020. URL https://arxiv.org/abs/2010.12858.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020. URL https://arxiv.org/abs/2010.11934.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and Dario Amodei. Language models are few-shot learners, 05 2020.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Ronald M. Baecker, Jonathan Grudin, William A. S. Buxton, and Saul Greenberg. *A historical and intellectual perspective*, page 35–47. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995. ISBN 1558602461.

Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, 39:494 – 518, 2021. URL https://api.semanticscholar.org/CorpusID:236428680.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423. URL http://www.jstor.org/stable/2251299.

Terry Winograd. Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artif. Intell.*, 170(18):1256–1258, dec 2006. ISSN 0004-3702. doi: 10.1016/j.artint.2006.10.011. URL https://doi.org/10.1016/j.artint.2006.10.011.

John Haugeland. *Artificial Intelligence: The Very Idea*. The MIT Press, 01 1989. ISBN 9780262291149. doi: 10.7551/mitpress/1170.001.0001. URL https://doi.org/10.7551/mitpress/1170.001.0001.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha,

Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Lee Rainie, Cary Funk, Monica Anderson, and Alec Tyson. Ai and human enhancement: Americans' openness is tempered by a range of concerns. *Pew Research Center: Internet Technology*, Mar 2022. URL https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns/.

Ben Shneiderman. *Human-Centered AI*. Oxford University Press, 2022a. ISBN 0192845292, 9780192845290. URL https://books.google.com/books/about/Human_centered_AI.html?id=YS9VEAAAQBAJ.

Jan Auernhammer. Human-centered ai: The role of human-centered design research in the development of ai. *DRS2020: Synergy*, 2020. URL https://api.semanticscholar.org/CorpusID:225225693.

Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, pages 277–284, 2003.

Eliezer Yudkowsky. Coherent extrapolated volition. *Technical report, Singularity Institute for Artificial Intelligence*, 2004.

Daniel Dewey. Learning what to value. In *Proceedings of the 4th International Conference on Artificial General Intelligence*, AGI'11, page 309–314, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642228865.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *ArXiv*, abs/1811.07871, 2018. URL https://api.semanticscholar.org/CorpusID:53745764.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://dx.doi.org/10.1007/s11023-020-09539-2.

Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. Ai alignment in the design of interactive ai: Specification alignment, process alignment, and evaluation support, 2023.

Charles Kiene and Benjamin Mako Hill. Who uses bots? A statistical analysis of bot usage in moderation teams. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–8, 2020.

Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–29, 2018.

Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an "eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156, 2016.

Catherine Han, Joseph Seering, Deepak Kumar, Jeffrey T Hancock, and Zakir Durumeric. Hate raids on twitch: Echoes of the past, new modalities, and implications for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–28, 2023.

Zhihua Cui, Xuechun Jing, Peng Zhao, Wensheng Zhang, and Jinjun Chen. A new subspace clustering strategy for ai-based data analysis in iot system. *IEEE Internet of Things Journal*, 8 (16):12540–12549, 2021. doi: 10.1109/JIOT.2021.3056578.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349451. doi: 10.1145/3172944.3172983. URL https://doi.org/10.1145/3172944.3172983.

Xi Victoria Lin. Program synthesis from natural language using recurrent neural networks. 2017. URL https://api.semanticscholar.org/CorpusID:3809743.

Kasra Ferdowsifard, Allen Ordookhanians, Hila Peleg, Sorin Lerner, and Nadia Polikarpova. Small-step live programming by example. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 614–626, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375146. doi: 10.1145/3379337.3415869. URL https://doi.org/10.1145/3379337.3415869.

Ben Shneiderman. Human-centered ai: ensuring human control while increasing automation. In *Proceedings of the 5th Workshop on Human Factors in Hypertext*, HUMAN '22, New York,

NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450394017. doi: 10.1145/3538882.3542790. URL https://doi.org/10.1145/3538882.3542790.

Jenny T. Liang, Chenyang Yang, and Brad A. Myers. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, ICSE '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400702174. doi: 10.1145/3597503.3608128. URL https://doi.org/10.1145/3597503.3608128.

Hyunjin Kang and Chen Lou. AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement. *Journal of Computer-Mediated Communication*, 27(5):zmac014, 08 2022. ISSN 1083-6101. doi: 10.1093/jcmc/zmac014. URL https://doi.org/10.1093/jcmc/zmac014.

Douglas Schuler and Aki Namioka. *Participatory Design: Principles and Practices*. Lawrence Erlbaum Associates, 1993. URL https://books.google.com/books?hl=en&lr=&id=pWOEk6Sk4YkC&oi=fnd&pg=PR7&dq=Participatory+Design+Schuler+Namioka&ots=p-zwtrs8Qf&sig=bxhby78WvlsffKftVGz26uQkiOk.

Sari Kujala. User involvement: A review of the benefits and challenges. *Behaviour & Information Technology*, 22(1):1–16, 2003. doi: 10.1080/01449290301782. URL https://doi.org/10.1080/01449290301782.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference*

on Human Factors in Computing Systems, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300233. URL https://doi.org/10.1145/3290605.3300233.

Dennis Wixon and Chauncey Wilson. Chapter 27 - the usability engineering framework for product design and evaluation. In Marting G. Helander, Thomas K. Landauer, and Prasad V. Prabhu, editors, *Handbook of Human-Computer Interaction (Second Edition)*, pages 653–688. North-Holland, Amsterdam, second edition edition, 1997. ISBN 978-0-444-81862-1. doi: https://doi.org/10.1016/B978-044481862-1.50093-5. URL https://www.sciencedirect.com/science/article/pii/B9780444818621500935.

John D. Gould and Clayton Lewis. Designing for usability: key principles and what designers think. *Commun. ACM*, 28(3):300–311, mar 1985. ISSN 0001-0782. doi: 10.1145/3166.3170. URL https://doi.org/10.1145/3166.3170.

Christiane Floyd, Wolf-Michael Mehl, Fanny-Michaela Reisin, Gerhard Schmidt, and Gregor Wolf. Out of scandinavia: alternative approaches to software design and system development. *Hum.-Comput. Interact.*, 4(4):253–350, dec 1989. ISSN 0737-0024. doi: 10.1207/s15327051hci0404_1. URL https://doi.org/10.1207/s15327051hci0404_1.

Linda Neuhauser and Gary L. Kreps. Participatory design and artificial intelligence: Strategies to improve health communication for diverse audiences. In *AAAI Spring Symposium: AI and Health Communication*, 2011. URL https://api.semanticscholar.org/CorpusID:8484616.

Rachel Smith Christian Dindler and Ole Sejer Iversen. Computational empowerment: participatory design in education. *CoDesign*, 16(1):66–80, 2020. doi: 10.1080/15710882.2020.

1722173. URL https://doi.org/10.1080/15710882.2020.1722173.

Batya Friedman, David G. Hendry, and Alan Borning. 2017. doi: 10.1561/1100000015.

Nigel Bevan, James Carter, and Susan Harker. Iso 9241-11 revised: What have we learnt about usability since 1998? In Masaaki Kurosu, editor, *Human-Computer Interaction: Design and Evaluation*, page 143–151, Cham, 2015. Springer International Publishing. ISBN 978-3-319-20901-2.

Shari Trewin. Ai fairness for people with disabilities: Point of view. *ArXiv*, abs/1811.10670, 2018. URL https://api.semanticscholar.org/CorpusID:53757043.

Brianna Blaser and Richard E. Ladner. Why is data on disability so hard to collect and understand? In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, volume 1, pages 1–8, 2020. doi: 10.1109/RESPECT49803.2020.9272466.

Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. An autoethnographic case study of generative artificial intelligence's utility for accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702204. doi: 10.1145/3597638.3614548. URL https://doi.org/10.1145/3597638.3614548.

Rie Kamikubo, Kyungjun Lee, and Hernisa Kacorri. Contributing to accessibility datasets: Reflections on sharing study data by blind people. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association

for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581337. URL https://doi.org/10.1145/3544548.3581337.

Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. Designing an online infrastructure for collecting ai data from people with disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 52–63, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445870. URL https://doi.org/10.1145/3442188.3445870.

Hernisa Kacorri, Utkarsh Dwivedi, Sravya Amancherla, Mayanka Jha, and Riya Chanduka. Incluset: A data surfacing repository for accessibility datasets. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371032. doi: 10.1145/3373625.3418026. URL https://doi.org/10.1145/3373625.3418026.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76893–76907. Cur-

ran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f29cf8f8b4996a4a453ef366cf496354-Paper-Datasets_and_Benchmarks.pdf.

A.M. Martinez, R.B. Wilbur, R. Shay, and A.C. Kak. Purdue rvl-slll asl database for automatic recognition of american sign language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172, 2002. doi: 10.1109/ICMI.2002.1166987.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://www.aclweb.org/anthology/N18-2002.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 2018. doi: 10.1162/tacl_a_00240. URL https://www.aclweb.org/anthology/Q18-1240.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *ACL Workshop on Ethics in NLP*, pages 74–79, 2017.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *arXiv preprint arXiv:1901.09451*, 2019a.

Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL https://www.aclweb.org/anthology/S18-2005.

Joel Escudé Font and Marta R. Costa-jussà. Equalizing gender biases in neural machine translation with word embeddings techniques. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

Marcelo Prates, Pedro Avelar, and Luis C. Lamb. Assessing gender bias in machine translation – a case study with google translate. *Neural Computing and Applications*, 2019.

Matthew S. Dryer. Expression of pronominal subjects. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

Anke Frank, Chr Hoffmann, Maria Strobel, et al. Gender issues in machine translation. *Univ. Bremen*, 2004.

Mario Wandruszka. *Sprachen: vergleichbar und unvergleichlich*. R. Piper & Company, 1969.

Uwe Kjær Nissen. Aspects of translating gender. *Linguistik online*, 11(2):02, 2002.

Ursula Doleschal and Sonja Schmid. Doing gender in Russian. *Gender Across Languages. The linguistic representation of women and men*, 1:253–282, 2001.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of NeurIPS*, 2016a.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT*, 2019.

Brian N. Larson. Gender as a variable in natural-language processing: Ethical considerations. In *ACL Workshop on Ethics in NLP*, 2017a.

Chandler May. Neurips2019, 2019. URL https://slideslive.com/38923450/deconstructing-gender-prediction-in-nlp.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. A survey of race, racism, and anti-racism in NLP. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.149. URL https://aclanthology.org/2021.acl-long.149.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017a. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://aclanthology.org/N19-1062.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. Twitter Universal Dependency parsing for African-American and mainstream American English. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1131. URL https://aclanthology.org/P18-1131.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. Investigating African-American Vernacular English in transformer-based text generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.473. URL https://aclanthology.org/2020.emnlp-main.473.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.239. URL https://aclanthology.org/2023.emnlp-main.239.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL https://aclanthology.org/W19-3504.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. Toward fairness in ai for people with disabilities sbg@a research roadmap. *SIGACCESS Access. Comput.*, (125), mar 2020. ISSN 1558-2337. doi: 10.1145/3386296.3386298. URL https://doi.org/10.1145/3386296.3386298.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Unintended machine learning biases as social barriers for persons with disabilitiess. *SIGACCESS Access. Comput.*, (125), mar 2020. ISSN 1558-2337. doi: 10.1145/3386296.3386305. URL https://doi.org/10.1145/3386296.3386305.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.113.

Daniela Massiceti, Camilla Longden, Agnieszka Słowik, Samuel Wills, Martin Grayson, and Cecily Morrison. Explaining clip's performance disparities on data from blind/low vision users. 2023.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *CoRR*, abs/2004.09456, 2020. URL https://arxiv.org/abs/2004.09456.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154. URL https://www.aclweb.org/anthology/2020.emnlp-main.154.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. UNQOVERing stereotyping biases via underspecified questions. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.311. URL https://aclanthology.org/2020.findings-emnlp.311.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Analyzing stereotypes in generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.355. URL https://aclanthology.org/2021.findings-acl.355.

Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing biases and the impact of multilingual training across multiple languages, 2023.

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in english, spanish, and arabic, 2022.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2896–2907, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.260. URL https://aclanthology.org/2020.acl-main.260.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.209. URL https://aclanthology.org/2020.emnlp-main.209.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4365–4381. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258823008.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in

text-to-image generation, 2024b.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462536. URL https://doi.org/10.1145/3461702.3462536.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337, 2019. URL http://arxiv.org/abs/1906.07337.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 862–872, Mar 2021. doi: 10.1145/3442188.3445924. arXiv: 2101.11718.

Conversational AI Jigsaw. Jigsaw unintended bias in toxicity classification. *Kaggle*, Nov 2019b. URL https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 120–128, Jan 2019b. doi: 10.1145/3287560.3287572. arXiv: 1901.09451.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*, pages 4349–4357, 2016b.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017b. ISSN 0036-8075. doi: 10.1126/science.aal4230.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *CoRR*, abs/1911.03891, 2019. URL http://arxiv.org/abs/1911.03891.

I.J. Wod. Weight of evidence: A brief survey. *Bayesian statistics*, 2:249–270, 1985.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. 2015.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Richard A. Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50:3 – 44, 2017. URL https://api.semanticscholar.org/CorpusID:12924416.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FvevdI0aA_h.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. Analyzing stereotypes in generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online, August 2021b. Association for Compu-

tational Linguistics. doi: 10.18653/v1/2021.findings-acl.355. URL https://aclanthology.org/2021.findings-acl.355.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996b.

Maria Bustillos. Our desperate, 250-year-long search for a gender-neutral pronoun, 2011. permalink.

Merriam-Webster. Words we're watching: Singular 'they', 2016. permalink.

Evan Bradley, Julia Salkind, Ally Moore, and Sofi Teitsort. Singular 'they' and novel pronouns: gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America*, 4(1):36–1–7, 2019. ISSN 2473-8689. doi: 10.3765/plsa.v4i1.4542. URL https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4542.

Levi C. R. Hord. Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german. 2016.

Michael Spivak. *The Joy of TEX: A Gourmet Guide to Typesetting with the AMS-TEX Macro Package*. American Mathematical Society, USA, 1st edition, 1997. ISBN 0821829971.

Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. 2018.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. page 8. ACM, 2018.

GLAAD. Media reference guide–transgender. permalink, 2007.

Max Lambert and Melina Packer. How gendered language leads scientists astray. New York Times, June 2019.

Emily M. Bender. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. The Future of Artificial Intelligence: Language, Ethics, Technology, March 25 2019.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. In *Proceedings of SIGCIS*, 2017.

KC Clements. Misgendering: What is it and why is it harmful?, Sep 2018. URL https://www.healthline.com/health/transgender/misgendering. Blog post, permalink.

Jennifer Wortman Vaughan and Hanna Wallach. Microsoft research webinar: Machine learning and fairness, 2019. permalink.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *arXiv:1803.09010*, 2018a.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal representation and gender stereotypes in image search results for occupations. 2015.

Latanya Sweeney. Discrimination in online ad delivery. *ACM Queue*, 2013.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. 2018.

Lauren Ackerman. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4, 10 2019. doi: 10.5334/gjgl.721.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. What's in a name? reducing bias in bios without access to protected attributes. 2019.

Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. Language (technology) is power: The need to be explicit about NLP harms. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2020.

Brian Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017b. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://www.aclweb.org/anthology/W17-1601.

Alan Garnham, Jane Oakhill, Marie-France Ehrlich, and Manuel Carreiras. Representations and processes in the interpretation of pronouns: New evidence from spanish and french. *Journal of Memory and Language*, 34(1), 1995.

Manuel Carreiras, Alan Garnham, Jane Oakhill, and Kate Cain. The use of stereotypical gender information in constructing a mental model: Evidence from english and spanish. *The Quarterly Journal of Experimental Psychology Section A*, 49(3), 1996.

Yulia Esaulova, Chiara Reali, and Lisa von Stockhausen. Influences of grammatical and stereotypical gender during reading: eye movements in pronominal and noun phrase anaphor resolution. *Language, Cognition and Neuroscience*, 29(7), 2014.

Chiara Reali, Yulia Esaulova, and Lisa Von Stockhausen. Isolating stereotypical gender in a grammatical gender language: evidence from eye movements. *Applied Psycholinguistics*, 36 (4), 2015.

Lee Osterhout and Linda A Mobley. Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, 34(6), 1995.

Peter Hagoort and Colin M Brown. Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of psycholinguistic research*, 28(6), 1999.

Lee Osterhout, Michael Bersick, and Judith McLaughlin. Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, 25(3), 1997.

Mary Bucholtz. Gender. *Journal of Linguistic Anthropology*, 1999. Special issue: Lexicon for the New Millennium, ed. Alessandro Duranti.

Susan Stryker. *Transgender history*. Seal Press, 2008.

Cheris Kramarae and Paula A. Treichler. *A feminist dictionary*. Pandora Press, 1985.

Candace West and Don H. Zimmerman. Doing gender. *Gender & society*, 1(2):125–151, 1987.

Judith Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, 1989.

Barbara J. Risman. From doing to undoing: Gender as we know it. *Gender & Society*, 23(1), 2009.

Julia Serano. *Whipping Girl: A Transsexual Woman on Sexism and the Scapegoating of Femininity*. Seal Press, 2007.

Suzanne J. Kessler and Wendy McKenna. *Gender: An ethnomethodological approach*. University of Chicago Press, 1978.

Kristen Schilt and Laurel Westbrook. Doing gender, doing heteronormativity. *Gender & Society*, 23(4), 2009.

Helana Darwin. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334, 2017.

Christina Richards, Walter Pierre Bouman, and Meg-John Barker. *Genderqueer and Non-Binary Genders*. Springer, 2017.

Judah HaNasi. Mishnah bikkurim, 189. In Mishnah, Chapter 4.

Richard Burton. *Kama Sutra, Translation*. 1883.

Karen Olsen Bruhns. Gender archaeology in native north america. In Sarah Nelson, editor, *Handbook of Gender in Archaeology*. 2006.

James Neill. *The Origins and Role of Same-Sex Relations in Human Societies*. 2008.

Greville G. Corbett. *Gender*. Cambridge University Press, 1991.

Elinor Ochs. Indexing gender. *Rethinking context: Language as an interactive phenomenon*, 11: 335, 1992.

Colette G. Craig. Classifier languages. *The encyclopedia of language and linguistics*, 2:565–569, 1994.

Greville G. Corbett. Number of genders. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

Marlis Hellinger and Heiko Motschenbacher. *Gender across languages*, volume 4. John Benjamins Publishing Company, 2015.

Robin Lakoff. Language and woman's place. *New York ao: Harper and Row*, 1975.

Michael Silverstein. Language structure and linguistic ideology. *The elements: A parasession on linguistic units and levels*, pages 193–247, 1979.

Pedro A. Fuertes-Olivera. A corpus-based view of lexical gender in written business english. *English for Specific Purposes*, 26(2):219–234, 2007.

Naomi S. Baron. A reanalysis of english grammatical gender. *Lingua*, 27:113–140, 1971. doi: 10.1016/0024-3841(71)90082-9. URL https://doi.org/10.1016/0024-3841(71)90082-9.

Bronwyn M. Bjorkman. Singular they and the syntactic representation of gender in english. *Glossa: A Journal of General Linguistics*, 2(1):80, 2017. doi: 10.1016/0024-3841(71)90082-9. URL http://doi.org/10.5334/gjgl.374.

Grusha Prasad and Joanna Morris. The p600 for singular "they": How the brain reacts when john decides to treat themselves to sushi. PsyArXiv, Jan 2020. doi: 10.31234/osf.io/hwzke. URL psyarxiv.com/hwzke.

Kirby Conrod. Changes in singular they. In *Cascadia Workshop in Sociolinguistics*, 2018a.

Anonymous. Understanding drag, Apr 2017. URL https://transequality.org/issues/resources/understanding-drag. Blog post, permalink.

Osten Dahl. Animacy and the notion of semantic gender. *Trends in linguistics studies and monographs*, 124:99–116, 2000.

John Lyons. *Semantics*. Cambridge University Press, 1977.

Kirby Conrod. What does it mean to agree? coreference with singular they. In *Pronouns in Competition workshop*, 2018b.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. 2015.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. 2018.

Bill Walsh. The post drops the 'mike'–and the hyphen in 'e-mail', December 2015. Washington Post; permalink.

Travis M. Andrews. The singular, gender-neutral 'they' added to the Associated Press Stylebook, March 2017. Washington Post; permalink.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. *OntoNotes: A Large Training Corpus for Enhanced Processing*. 01 2011.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*, 2019.

Aravind K. Joshi and Scott Weinstein. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'81, pages 385–387, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 44–50, Cambridge, Massachusetts, USA, June 1983. Association for Computational Linguistics. doi: 10.3115/981311.981320. URL https://www.aclweb.org/anthology/P83-1007.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363, 2004. doi: 10.1162/0891201041850911. URL https://www.aclweb.org/anthology/J04-3003.

Jennifer E Arnold. Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, 23(4):495–527, 2008.

Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé, III. Why discourse affects speakers' choices of referring expressions. 2015a.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

Hal Daumé, III and Daniel Marcu. A large-scale exploration of effective global features for a joint entity detection and tracking model. pages 97–104, 2005.

Naho Orita, Eliana Vornov, Naomi H. Feldman, and Hal Daumé, III. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2015b.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv:1803.07640*, 2017.

Thomas Wolf. State-of-the-art neural coreference resolution for chatbots, Oct 2017. URL https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30. Blog post, permalink.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. 2010.

Kevin Clark and Christopher D. Manning. Entity-centric coreference resolution with model stacking. 2015.

Kevin Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. 2016.

KC Clements. What is deadnaming?, 2017. URL https://www.healthline.com/health/transgender/deadnaming. Blog post, permalink.

Nafise Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. pages 632–642, 01 2016. doi: 10.18653/v1/P16-1060.

Nafise Sadat Moosavi. *Robustness in Coreference Resolution*. PhD dissertation, University of Heidelberg, 2020.

Elizabeth Du and Ross Liddy. Anaphora in natural language processing and information retrieval. *Inf. Process. Manage.*, 26, 1990.

Ari Pirkola and Kalervo Järvelin. The effect of anaphor and ellipsis resolution on proximity searching in a text database. *Inf. Process. Manage.*, 32, 1996.

Richard J. Edens, Helen L. Gaylard, Gareth J. F. Jones, and Adenike M. Lam-Adesina. An investigation of broad coverage automatic pronoun resolution for information retrieval. In *SIGIR*, 2003.

Ruslan Mitkov. Introduction: special issue on anaphora resolution in machine translation and multilingual NLP. *Machine translation*, 14(3), 1999.

Christian Hardmeier and Marcello Federico. Modelling pronominal anaphora in statistical machine translation. In *IWSLT*, 2010.

Liane Guillou. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-3001.

Christian Hardmeier and Liane Guillou. Pronoun translation in english-french machine translation: An analysis of error types. *arXiv preprint arXiv:1808.10196*, 2018.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt cheap web-scale parallel text from the common crawl. 2013.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. 2012.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics, 1995.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1, 2005.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada, 1998.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. 2009.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics, 2005.

Jie Cai and Michael Strube. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36, 2010.

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, 2019.

Nancy Fraser. Abnormal Justice. *Critical Inquiry*, 34(3):393–422, 2008.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Matt Post. A call for clarity in reporting Bleu scores. In *WMT*, 2018.

Vincent Ng and Claire Cardie. Bootstrapping coreference classifiers with multiple machine learning algorithms. 2003.

Florian Laws, Florian Heimerl, and Hinrich Schütze. Active learning for coreference resolution. 2012.

Timothy A Miller, Dmitriy Dligach, and Guergana K Savova. Active learning for coreference resolution. In *Workshop on Biomedical NLP*, 2012.

Mrinmaya Sachan, Eduard Hovy, and Eric P Xing. An active learning approach to coreference resolution. 2015.

Altaf Rahman and Vincent Ng. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824. Association for Computational Linguistics, 2011.

Mohit Bansal and Dan Klein. Coreference semantics from web features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 389–398. Association for Computational Linguistics, 2012.

Shane Bergsma and Dekang Lin. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220180. URL https://www.aclweb.org/anthology/P06-1005.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. Annotating twitter data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. 01 2019.

Meg Young, Lassana Magassa, and Batya Friedman. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, 2019.

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and A. Nenkova. Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models. *ArXiv*, abs/2004.04123, 2020.

Ellen Bard and Matthew Aylett. Referential form word duration and modeling the listener in spoken dialogue. 01 2004.

Emiel Krahmer and Kees van Deemter. Computational generation of referring expressions: A survey. *Comput. Linguist.*, 38(1):173–218, March 2012. ISSN 0891-2017. doi: 10.1162/COLI_a_00088. URL https://doi.org/10.1162/COLI_a_00088.

Teresa De Lauretis. Feminism and its differences. *Pacific Coast Philology*, pages 24–30, 1990.

Annamarie Jagose. *Queer theory: An introduction*. NYU Press, 1996.

Ann Light. HCI as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with Computers*, 23(5), 2011.

Shaowen Bardzell and Elizabeth F Churchill. Iwc special issue "feminism and HCI: new perspectives" special issue editors' introduction. *Interacting with Computers*, 23(5), 2011.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 15009–15018, June 2023. doi: 10.1609/aaai.v37i12.26752.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit rules! Characterizing an ecosystem of governance. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(11), June 2018. ISSN 2334-0770. doi: 10.1609/icwsm.v12i1.15033. URL https://ojs.aaai.org/index.php/ICWSM/article/view/15033.

Joseph Seering. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4:107, 2020.

Amy Binns. Don't feed the trolls! Managing troublemakers in magazines' online communities. *Journalism practice*, 6(4):547–562, 2012.

Julian Dibbell. A rape in cyberspace or how an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.*, page 471, 1994.

Hanlin Li, Brent Hecht, and Stevie Chancellor. Measuring the monetary value of online volunteer work. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 596–606, 2022.

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019.

Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on Twitter. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3006. URL https://aclanthology.org/W17-3006.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144. URL https://aclanthology.org/N19-1144.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL https://doi.org/10.1145/3534678.3539147.

Ilan Price, Jordan Gifford-Moore, Jory Fleming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. Six attributes of unhealthy conversation. *arXiv:2010.07410 [cs]*, October 2020. URL http://arxiv.org/abs/2010.07410. arXiv: 2010.07410.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. Watch your language: Large language models and content moderation. (arXiv:2309.14517), September 2023. URL http://arxiv.org/abs/2309.14517. arXiv:2309.14517 [cs].

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, November 2018. ISSN 2573-0142. doi: 10.1145/3274301.

Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. Identifying machine-paraphrased plagiarism. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future*, pages 393–413, Cham, 2022. Springer International Publishing. ISBN 978-3-030-96957-8.

Jessica Shieh. Best practices for prompt engineering with openai api, 2023. URL https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api. [Online; accessed Nov 2023].

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European*

conference on computer vision, pages 740–755. Springer, 2014.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.7. URL https://aclanthology.org/2020.acl-tutorials.7.

L Stephen Coles. An on-line question-answering systems with natural language and pictorial input. In *Proceedings of the 1968 23rd ACM national conference*, pages 157–167, 1968.

Mariet Theune, Boris van Schooten, Rieks op den Akker, WAUTER Bosma, DENNIS Hofs, Anton Nijholt, EMIEL Krahmer, Charlotte van Hooijdonk, and Erwin Marsi. Questions, pictures, answers: Introducing pictures in question-answering systems. *ACTAS-1 of X Symposio Internacional de Comunicacion Social, Santiago de Cuba*, pages 450–463, 2007.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690, 2014.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1): 4–31, may 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0966-6. URL https://doi.org/10.1007/s11263-016-0966-6.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4271–4280, 2019.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. URL http://arxiv.org/abs/1505.00468.

Chun-Ju Yang, Kristen Grauman, and Danna Gurari. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

Walter Lippmann. *Public Opinion*. New York :Free Press, 1965.

J.S. Bruner, Brunswik E, L. Festinger, F. Heider, K.F. Muenzinger, C.E. Osgood, and D. Rapaport. Going beyond the information given. *Contemporary approaches to cognition*, pages 41–67, 1957.

S. Wheeler and Richard Petty. The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological bulletin*, 127:797–826, 12 2001. doi: 10.1037/0033-2909.127.6.797.

Dr. Charles Stangor. Principles of social psychology – 1st international edition. BCcampus, 2014.

Lynne M. Jackson. The psychology of prejudice: From attitudes to social action. American Psychological Association, 2011. ISBN 978-1-4338-0920-0. URL https://books.google.com/books?id=Q8MkAQAAMAAJ.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82 6:878–902, 2002.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Debora Nozza, Federico Bianchi, and Dirk Hovy. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.191. URL https://aclanthology.org/2021.naacl-main.191.

Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.49. URL https://aclanthology.org/2021.naacl-main.49.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Negin Ghavami and Letitia Anne Peplau. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1):113–127, 2013. doi: 10.1177/0361684312464203. URL https://doi.org/10.1177/0361684312464203.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

Alex Koch, Vincent Yzerbyt, Andrea Abele, Naomi Ellemers, and Susan T. Fiske. *Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts*, volume 63, page 1–68. Elsevier, 2021. ISBN 978-0-12-824578-1. doi: 10.1016/bs.aesp.2020.11.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0065260120300265.

Patricia Hill Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2002.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets, 2018b. URL https://arxiv.org/abs/1803.09010.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. *CoRR*, abs/0812.4952, 2008. URL http://arxiv.org/abs/0812.4952.

Victor Bittorf, Benjamin Recht, Christopher Ré, and Joel Tropp. Factoring nonnegative matrices with linear programs. *Advances in Neural Information Processing Systems*, 2, 06 2012.

Hal Daumé, III and Abhishek Kumar. Column squishing for multiclass updates (blog post), 2017. URL https://nlpers.blogspot.com/2017/08/column-squishing-for-multiclass-updates.html.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing Bayesian Models of Annotation. *Transactions of the Association for Compu-*

tational Linguistics, 6:571–585, 12 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00040. URL https://doi.org/10.1162/tacl_a_00040.

Robert E. Botsch. *Significance and Measures of Association*. Aug 2011. URL http://polisci.usca.edu/apls301/Text/Chapter12.SignificanceandMeasuresofAssociation.htm.

Combahee River Collective. *A Black Feminist Statement*. na, 1977.

Combahee River Collective. The combahee river collective statement. *Home girls: A Black feminist anthology*, 1:264–274, 1983.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

Irene Browne and Joya Misra. The intersection of gender and race in the labor market. *Annual review of sociology*, 29(1):487–513, 2003.

Amy C Steinbugler, Julie E Press, and Janice Johnson Dias. Gender, race, and affirmative action: Operationalizing intersectionality in survey research. *Gender & Society*, 20(6):805–825, 2006.

Maxine Baca Zinn and Bonnie Thornton Dill. Theorizing difference from multiracial feminism. *Feminist studies*, 22(2):321–331, 1996.

Deborah K King. Multiple jeopardy, multiple consciousness: The context of a black feminist ideology. *Signs: Journal of women in culture and society*, 14(1):42–72, 1988.

Eduardo Bonilla-Silva. Rethinking racism: Toward a structural interpretation. *American sociological review*, pages 465–480, 1997.

bell hooks. Yearning: Race, gender, and cultural politics. *Hypatia*, 7(2), 1992.

Sheldon Stryker. *Symbolic interactionism: a social structural version*. Benjamin/Cummings Pub. Co, 1980.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. 2022a. doi: 10.48550/ARXIV.2203.13928. URL https://arxiv.org/abs/2203.13928.

Carl A. Latkin, Catie Edwards, Melissa A. Davey-Rothwell, and Karin E. Tobin. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in baltimore, maryland. *Addictive Behaviors*, 73:133–136, Oct 2017. ISSN 1873-6327. doi: 10.1016/j.addbeh.2017.05.005.

Joel E. Martinez, Lauren A. Feldman, Mallory J. Feldman, and Mina Cikara. Narratives shape cognitive representations of immigrants and immigration-policy preferences. *Psychological Science*, 32(2):135–152, Feb 2021. ISSN 0956-7976. doi: 10.1177/0956797620963610.

Sarah Ariel Lamer, Paige Dvorak, Ashley M. Biddle, Kristin Pauker, and Max Weisbuch. The transmission of gender stereotypes through televised patterns of nonverbal bias. *Journal of Personality and Social Psychology*, 123(6):1315–1335, 2022. ISSN 1939-1315. doi: 10.1037/pspi0000390.

Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34):13526–13531, Aug 2012. doi: 10.1073/pnas.1208951109.

CHINUA THELWELL. *Front Matter*, pages i–iv. University of Massachusetts Press, 2020. ISBN 9781625345165. URL http://www.jstor.org/stable/j.ctv160btb3.1.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.616.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021. URL https://arxiv.org/abs/2112.04359.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019. URL http://arxiv.org/abs/1901.07291.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.92. URL https://aclanthology.org/2022.naacl-main.92.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.197. URL https://aclanthology.org/2022.naacl-main.197.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, 2022.

Alex Koch, Angela Dorrough, Andreas Glöckner, and Roland Imhoff. The abc of society: Perceived similarity in agency/socioeconomic success and conservative-progressive beliefs increases intergroup cooperation. *Journal of Experimental Social Psychology*, 90:103996, 2020. ISSN 0022-1031. doi: https://doi.org/10.1016/j.jesp.2020.103996.

Frank N. Pieke. Immigrant china. *Modern China*, 38(1):40–77, 2012. doi: 10.1177/0097700411424564.

Julie E. Miller-Cribbs and Naomi B. Farber. Kin Networks and Poverty among African Americans: Past and Present. *Social Work*, 53(1):43–51, 01 2008. ISSN 0037-8046. doi: 10.1093/sw/53.1.43. URL https://doi.org/10.1093/sw/53.1.43.

George Galster. Housing discrimination and urban poverty of african-americans. *Journal of Housing Research*, 2, 01 1992.

PETER BERESFORD. Poverty and disabled people: Challenging dominant debates and policies. *Disability & Society*, 11(4):553–568, 1996. doi: 10.1080/09687599627598. URL https://doi.org/10.1080/09687599627598.

Jonathan Evans and Neha Sahgal. Key findings about religion in india. *Pew Research Centre*, 2021.

Ben Hillman. The rise of the community in rural china: Village politics, cultural identity and religious revival in a hui hamlet. *The China Journal*, (51):53–73, 2004.

Ding Hong. A comparative study on the cultures of the dungan and the hui peoples. *Asian Ethnicity*, 6(2):135–140, 2005. doi: 10.1080/14631360500135765. URL https://doi.org/10.1080/14631360500135765.

Michael Witzel. Toward a history of the brahmins. *Journal of the American Oriental Society*, 113:264, 1993. URL https://api.semanticscholar.org/CorpusID:163531550.

Murray Milner. Hindu eschatology and the indian caste system: An example of structural reversal. *The Journal of Asian Studies*, 52:298–319, 1993. URL https://doi.org/10.2307/2059649.

Christina N. Harrington, Aashaka Desai, Aaleyah Lewis, Sanika Moharana, Anne Spencer Ross, and Jennifer Mankoff. Working at the intersection of race, disability and accessibility. In *ASSETS*. ACM, 2023. ISBN 9798400702204. doi: 10.1145/3597638.3608389. URL https://doi.org/10.1145/3597638.3608389.

Candace L. Sidner. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, 7(4):217–231, 1981. URL https://www.aclweb.org/anthology/J81-4001.

R. I. Bainbridge. Montagovian definite clause grammar. In *Second Conference of the European Chapter of the Association for Computational Linguistics*, Geneva, Switzerland, March 1985. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E85-1004.

Megumi Kameyama. A property-sharing constraint in centering. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, New York, New York, USA, July 1986. Association for Computational Linguistics. doi: 10.3115/981131.981159. URL https://www.aclweb.org/anthology/P86-1031.

C. S. Mellish. Implementing systemic classification by unification. *Computational Linguistics*, 14(1), 1988. URL https://www.aclweb.org/anthology/J88-1004.

Laurence Danlos and Fiametta Namer. Morphology and cross dependencies in the synthesis of personal pronouns in romance languages. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988. URL https://www.aclweb.org/anthology/C88-1029.

Kei Yoshimoto. Identifying zero pronouns in Japanese dialogue. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*, 1988. URL https://www.aclweb.org/anthology/C88-2159.

Michael Zock, Gil Francopoulo, and Abdellatif Laroui. Language learning as problem solving. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*, 1988. URL https://www.aclweb.org/anthology/C88-2164.

Fred Popowich. Tree unification grammar. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 228–236, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics. doi: 10.3115/981623.981651. URL https://www.aclweb.org/anthology/P89-1028.

Inderjeet Mani, T. Richard Macmillan, Susann Luperfoy, Elaine Lusher, and Sharon Laskowski. Identifying unknown proper names in newswire text. In *Acquisition of Lexical Knowledge from Text*, 1993. URL https://www.aclweb.org/anthology/W93-0105.

Ajit Narayanan and Lama Hashem. On abstract finite-state morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands, April 1993. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E93-1035.

Danny Soloman and Mary McGee Wood. Learning a radically lexical grammar. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 1994. URL https://www.aclweb.org/anthology/W94-0115.

J. Joachim Quantz. An HPSG parser based on description logics. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, 1994. URL https://www.aclweb.org/anthology/C94-1067.

Janet Baker, Larry Gillick, and Robert Roth. Research in large vocabulary continuous speech recognition. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://www.aclweb.org/anthology/H94-1097.

Damien Genthial, Jacques Courtin, and Jacques Menezo. Towards a more user-friendly correction. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*, 1994. URL https://www.aclweb.org/anthology/C94-2176.

Moshe Levinger, Uzzi Ornan, and Alon Itai. Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew. *Computational Linguistics*, 21(3):383–404, 1995. URL https://www.aclweb.org/anthology/J95-3004.

Tomáš Holan, Vladislav Kuboň, and Martin Plátek. A prototype of a grammar checker for Czech. In *Fifth Conference on Applied Natural Language Processing*, pages 147–154, Washington, DC, USA, March 1997. Association for Computational Linguistics. doi: 10.3115/974557.974579. URL https://www.aclweb.org/anthology/A97-1022.

Michael Dorna, Anette Frank, Josef van Genabith, and Martin C. Emele. Syntactic and semantic transfer with f-structures. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 341–347, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. doi: 10.3115/980845.980903. URL https://www.aclweb.org/anthology/P98-1056.

Sanda M. Harabagiu and Steven J. Maiorano. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *The Relation of Discourse/Dialogue Structure and Reference*, 1999. URL https://www.aclweb.org/anthology/W99-0104.

Tania Avgustinova and Hans Uszkoreit. An ontology of systematic relations for a shared grammar of Slavic. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000. URL https://www.aclweb.org/anthology/C00-1005.

Songsak Channarukul, Susan W. McRoy, and Syed S. Ali. Enriching partially-specified representations for text realization using an attribute grammar. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 163–170, Mitzpe Ramon, Israel, June 2000. Association for Computational Linguistics. doi: 10.3115/1118253.1118276. URL https://www.aclweb.org/anthology/W00-1422.

Saleem Abuleil, Khalid Alsamara, and Martha Evens. Acquisition system for Arabic noun morphology. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1118637.1118640. URL https://www.aclweb.org/anthology/W02-0503.

Silviu Cucerzan and David Yarowsky. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47, 2003. URL https://www.aclweb.org/anthology/N03-1006.

Serguei V. Pakhomov, James Buntrock, and Christopher G. Chute. Identification of patients with congestive heart failure using a binary classifier: A case study. In *Proceedings of the*

ACL 2003 Workshop on Natural Language Processing in Biomedicine, pages 89–96, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1118958.1118970. URL https://www.aclweb.org/anthology/W03-1312.

Marko Tadić and Sanja Fulgosi. Building the Croatian morphological lexicon. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 41–45, Budapest, Hungary, April 2003. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W03-2906.

Łukasz Debowski. A reconfigurable stochastic tagger for languages with complex tag structure. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 63–70, Budapest, Hungary, April 2003. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W03-2909.

Costanza Navarretta. An algorithm for resolving individual and abstract anaphora in Danish texts and dialogues. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 95–102, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-0713.

Michael Carl, Sandrine Garnier, Johann Haller, Anne Altmayer, and Bärbel Miemietz. Controlling gender equality with shallow NLP techniques. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 820–826, Geneva, Switzerland, aug 23–aug 27 2004. COLING. URL https://www.aclweb.org/anthology/C04-1118.

Cristina Mota, Paula Carvalho, and Elisabete Ranchhod. Multiword lexical acquisition and dictionary formalization. In *Proceedings of the Workshop on Enhancing and Using Elec-*

tronic Dictionaries, pages 73–76, Geneva, Switzerland, August 29th 2004. COLING. URL https://www.aclweb.org/anthology/W04-2115.

Jason Eisner and Damianos Karakos. Bootstrapping without the boot. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 395–402, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1050.

Constantinos Boulis and Mari Ostendorf. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 435–442, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219894. URL https://www.aclweb.org/anthology/P05-1054.

Noah A. Smith, David A. Smith, and Roy W. Tromble. Context-based morphological disambiguation with random fields. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 475–482, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1060.

Thurid Vogt and Elisabeth André. Improving automatic emotion recognition from speech via gender differentiaion. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/392_pdf.pdf.

Chris Quirk and Simon Corston-Oliver. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 62–69, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W06-1608.

Ali Dada. Implementation of the Arabic numerals and their syntax in GF. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 9–16, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W07-0802.

Oliver Streiter, Leonhard Voltmer, and Yoann Goudin. From tombstones to corpora: TSML for research on language, culture, identity and gender differences. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*, pages 450–458, Seoul National University, Seoul, Korea, November 2007. The Korean Society for Language and Information (KSLI). doi: http://hdl.handle.net/2065/29135. URL https://www.aclweb.org/anthology/Y07-1047.

Hongyan Jing, Nanda Kambhatla, and Salim Roukos. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1040–1047, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-1131.

Ibrahim Badr, Rabih Zbib, and James Glass. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 153–156, Columbus,

Ohio, June 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P08-2039.

Harmony Marchal, Benoît Lemaire, Maryse Bianco, and Philippe Dessus. A MDL-based model of gender knowledge acquisition. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 73–80, Manchester, England, August 2008. Coling 2008 Organizing Committee. URL https://www.aclweb.org/anthology/W08-2110.

Wido van Peursen. How to establish a verbal paradigm on the basis of ancient Syriac manuscripts. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 1–9, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-0801.

Ibrahim Badr, Rabih Zbib, and James Glass. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 86–93, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E09-1011.

Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P09-1080.

Shane Bergsma, Dekang Lin, and Randy Goebel. Glen, glenda or glendale: Unsupervised and semi-supervised learning of English noun gender. In *Proceedings of the Thirteenth Conference*

on Computational Natural Language Learning (CoNLL-2009), pages 120–128, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-1116.

Vivi Nastase and Marius Popescu. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D09-1142.

Hidetsugu Nanba, Haruka Taguma, Takahiro Ozaki, Daisuke Kobayashi, Aya Ishino, and Toshiyuki Takezawa. Automatic compilation of travel information from automatically identified travel blogs. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 205–208, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P09-2052.

Livio Robaldo and Jurij Di Carlo. Disambiguating quantifier scope in DTS. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 195–209, Tilburg, The Netherlands, January 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-3718.

Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA, October 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D10-1021.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-1142.

Felix Burkhardt, Martin Eckert, Wiebke Johannsen, and Joachim Stegmann. A database of age and gender annotated telephone speech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/262_Paper.pdf.

Yuval Marton, Nizar Habash, and Owen Rambow. Improving Arabic dependency parsing with lexical and inflectional morphological features. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 13–21, Los Angeles, CA, USA, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-1402.

Ronan Le Nagard and Philipp Koehn. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-1737.

Lina Maria Rojas-Barahona, Thierry Bazillon, Matthieu Quignard, and Fabrice Lefevre. Using MMIL for the high level semantic annotation of the French MEDIA dialogue corpus. In *Pro-*

ceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), 2011. URL https://www.aclweb.org/anthology/W11-0146.

Smruthi Mukund, Debanjan Ghosh, and Rohini Srihari. Using sequence kernels to identify opinion entities in Urdu. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 58–67, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-0308.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-0310.

Dingcheng Li, Tim Miller, and William Schuler. A pronoun anaphora resolution system based on factorial hidden Markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1169–1178, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1117.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1120.

Saif Mohammad and Tony Yang. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and*

Sentiment Analysis (WASSA 2.011), pages 70–79, Portland, Oregon, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-1709.

Emili Sapena, Lluís Padró, and Jordi Turmo. RelaxCor participation in CoNLL shared task on coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 35–39, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-1903.

Eric Charton and Michel Gagnon. Poly-co: a multilayer perceptron approach for coreference detection. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 97–101, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-1915.

Sarah Alkuhlani and Nizar Habash. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-2062.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. Two-step translation with grammatical post-processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2152.

Verónica López-Ludeña, Rubén San-Segundo, Syaheerah Lufti, Juan Manuel Lucas-Cuesta, Julián David Echevarry, and Beatriz Martínez-González. Source language categorization

for improving a speech into sign language translation system. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 84–93, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2309.

Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-1006.

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N12-1033.

Sarah Alkuhlani and Nizar Habash. Identifying broken plurals, irregular gender, and rationality in Arabic text. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–685, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-1069.

Katja Filippova. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning, pages 1478–1488, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D12-1135.

Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Şulea. The Romanian neuter examined through a two-gender n-gram classification system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 907–910, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/651_Paper.pdf.

Ahmed El Kholy and Nizar Habash. Rich morphology generation using statistical machine translation. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 90–94, Utica, IL, May 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-1514.

Bei Yu. Function words for Chinese authorship attribution. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 45–53, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-2506.

Adam Vogel and Dan Jurafsky. He said, she said: Gender in the ACL anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-3204.

Yoav Goldberg and Michael Elhadad. Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics*, 39(1):121–160,

2013. doi: 10.1162/COLI_a_00137. URL https://www.aclweb.org/anthology/J13-1007.

Yuval Marton, Nizar Habash, and Owen Rambow. Dependency parsing of modern standard Arabic with lexical and inflectional features. *Computational Linguistics*, 39(1):161–194, 2013. doi: 10.1162/COLI_a_00138. URL https://www.aclweb.org/anthology/J13-1008.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. Using subcategorization knowledge to improve case prediction for translation to German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–603, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1058.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1114.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1187.

Rivka Levitan. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages

84–90, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-2012.

Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-2208.

Goran Glavaš, Damir Korenčić, and Jan Šnajder. Aspect-oriented opinion mining from user reviews in Croatian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 18–23, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-2404.

Yuanchao Liu, Ming Liu, Xiaolong Wang, Limin Wang, and Jingjing Li. PAL: A chatterbot system for answering domain-specific questions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 67–72, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-4012.

Mike Kestemont. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0908. URL https://www.aclweb.org/anthology/W14-0908.

Michal Novák and Zdeněk Žabokrtský. Cross-lingual coreference resolution of pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics:*

Technical Papers, pages 14–24, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1003.

Bogdan Babych, Jonathan Geiger, Mireia Ginestí Rosell, and Kurt Eberle. Deriving de/het gender classification for Dutch nouns for rule-based MT generation tasks. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 75–81, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1014. URL https://www.aclweb.org/anthology/W14-1014.

Juan Soler-Company and Leo Wanner. How to use less features and reach better performance in author gender identification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1315–1319, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/104_Paper.pdf.

Chen Chen and Vincent Ng. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–774, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1084. URL https://www.aclweb.org/anthology/D14-1084.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar, Oc-

tober 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1121. URL https://www.aclweb.org/anthology/D14-1121.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland, August 2014a. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1184.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1211. URL https://www.aclweb.org/anthology/D14-1211.

Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. Comparison of gender- and speaker-adaptive emotion recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3476–3480, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/322_Paper.pdf.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. Using stem-templates to improve Arabic POS and gender/number tagging. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2926–2931, Reykjavik, Iceland,

May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/335_Paper.pdf.

Tafseer Ahmed Khan. Automatic acquisition of Urdu nouns (along with gender and irregular plurals). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2846–2850, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/844_Paper.pdf.

Dong Nguyen, Dolf Trieschnigg, and Theo Meder. TweetGenie: Development, evaluation, and lessons learned. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 62–66, Dublin, Ireland, August 2014b. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-2014.

Ian Stewart. Now we stronger than ever: African-American English syntax in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-3004. URL https://www.aclweb.org/anthology/E14-3004.

Austin Matthews, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. The CMU machine translation systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 142–149, Baltimore, Maryland, USA, June 2014. Association

for Computational Linguistics. doi: 10.3115/v1/W14-3315. URL https://www.aclweb.org/anthology/W14-3315.

Ashwini Vaidya, Owen Rambow, and Martha Palmer. Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 127–136, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. doi: 10.3115/v1/W14-5816. URL https://www.aclweb.org/anthology/W14-5816.

Dimitrios Kokkinakis, Ann Ighe, and Mats Malm. Gender-based vocation identification in Swedish 19th century prose fiction using linguistic patterns, NER and CRF learning. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 89–97, Denver, Colorado, USA, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0710. URL https://www.aclweb.org/anthology/W15-0710.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1011. URL https://www.aclweb.org/anthology/K15-1011.

H. Andrew Schwartz, Gregory Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman, and Lyle Ungar. Extracting human temporal orientation from Facebook language. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 409–419, Denver, Colorado, May–

June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1044. URL https://www.aclweb.org/anthology/N15-1044.

Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1073. URL https://www.aclweb.org/anthology/P15-1073.

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. Key female characters in film have more to talk about besides men: Automating the Bechdel test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1084. URL https://www.aclweb.org/anthology/N15-1084.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, Colorado, June 5 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1203. URL https://www.aclweb.org/anthology/W15-1203.

Anil Ramakrishna, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *Proceed-*

ings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1996–2001, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1234. URL https://www.aclweb.org/anthology/D15-1234.

Tomoki Taniguchi, Shigeyuki Sakaki, Ryosuke Shigenaka, Yukihiro Tsuboshita, and Tomoko Ohkuma. A weighted combination of text and image classifiers for user gender inference. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 87–93, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-2814. URL https://www.aclweb.org/anthology/W15-2814.

Alexandra Schofield and Leo Mehr. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, California, USA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0204. URL https://www.aclweb.org/anthology/W16-0204.

Sarah Ita Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 40–44, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0806. URL https://www.aclweb.org/anthology/W16-0806.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 843–854, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1080. URL https://www.aclweb.org/anthology/P16-1080.

Trang Tran and Mari Ostendorf. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1108. URL https://www.aclweb.org/anthology/D16-1108.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. Investigating language universal and specific properties in word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1140. URL https://www.aclweb.org/anthology/P16-1140.

Shoushan Li, Bin Dai, Zhengxian Gong, and Guodong Zhou. Semi-supervised gender classification with joint textual and social modeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2092–2100, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1197.

Dong Zhang, Shoushan Li, Hongling Wang, and Guodong Zhou. User classification with multiple textual perspectives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2112–2121, Osaka, Japan, December

2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1199.

Aparna Garimella and Rada Mihalcea. Zooming in on gender differences in social media. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 1–10, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-4301.

Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5603. URL https://www.aclweb.org/anthology/W16-5603.

Wen Li and Markus Dickinson. Gender prediction for Chinese social media data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 438–445, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_058. URL https://doi.org/10.26615/978-954-452-049-6_058.

Carlos Pérez Estruch, Roberto Paredes Palacios, and Paolo Rosso. Learning multimodal gender profile using neural networks. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 577–582, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_075. URL https://doi.org/10.26615/978-954-452-049-6_075.

Verónica Pérez-Rosas, Quincy Davenport, Anna Mengdan Dai, Mohamed Abouelenien, and Rada Mihalcea. Identity deception detection. In *Proceedings of the Eighth International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 885–894, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1089.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1101.

Marta R. Costa-jussà. Why Catalan-Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1207. URL https://www.aclweb.org/anthology/W17-1207.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1247. URL https://www.aclweb.org/anthology/D17-1247.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–

2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Justina Mandravickaitė and Tomas Krilavičius. Stylometric analysis of parliamentary speeches: Gender dimension. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 102–107, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1416. URL https://www.aclweb.org/anthology/W17-1416.

Ben Verhoeven, Iza Škrjanec, and Senja Pollak. Gender profiling for Slovene twitter communication: the influence of gender marking, content and style. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 119–125, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1418. URL https://www.aclweb.org/anthology/W17-1418.

Corina Koolen and Andreas van Cranenburgh. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1602. URL https://www.aclweb.org/anthology/W17-1602.

Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1606. URL https://www.aclweb.org/anthology/W17-1606.

Juan Soler-Company and Leo Wanner. On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 681–687, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2108.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. Language-independent gender prediction on twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 1–6, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2901. URL https://www.aclweb.org/anthology/W17-2901.

Olga Litvinova, Pavel Seredin, Tatiana Litvinova, and John Lyell. Deception detection in Russian texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 43–52, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-4005.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001. URL https://www.aclweb.org/anthology/S18-1001.

Zijian Wang and David Jurgens. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 33–45, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1004. URL https://www.aclweb.org/anthology/D18-1004.

Matthias Kraus, Johannes Kraus, Martin Baumann, and Wolfgang Minker. Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1018.

Matej Martinc and Senja Pollak. Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1082.

Sophia Chan and Alona Fyshe. Social and emotional correlates of capitalization on twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 10–15, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1102. URL https://www.aclweb.org/anthology/W18-1102.

Esin Durmus and Claire Cardie. Understanding the effect of gender and stance in opinion expression in debates on "abortion". In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 69–75,

New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1110. URL https://www.aclweb.org/anthology/W18-1110.

Wajdi Zaghouani and Anis Charfi. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1111.

Barbara Plank. Predicting authorship and author traits from keystroke dynamics. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 98–104, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1113. URL https://www.aclweb.org/anthology/W18-1113.

Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, and Mark Dredze. Predicting twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 105–111, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1114. URL https://www.aclweb.org/anthology/W18-1114.

Sridhar Moorthy, Ruth Pogacar, Samin Khan, and Yang Xu. Is Nike female? exploring the role of sound symbolism in predicting brand name gender. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1128–1132, Brus-

sels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1142. URL https://www.aclweb.org/anthology/D18-1142.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1176. URL https://www.aclweb.org/anthology/N18-1176.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://www.aclweb.org/anthology/D18-1302.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL https://www.aclweb.org/anthology/D18-1334.

Bennett Kleinberg, Maximilian Mozes, and Isabelle van der Vegt. Identifying the sentiment styles of YouTube's vloggers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3581–3590, Brussels, Belgium, October-November

2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1394. URL https://www.aclweb.org/anthology/D18-1394.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://www.aclweb.org/anthology/D18-1521.

Murali Raghu Babu Balusu, Taha Merghani, and Jacob Eisenstein. Stylistic variation in social media part-of-speech tagging. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 11–19, New Orleans, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1602. URL https://www.aclweb.org/anthology/W18-1602.

Francesco Barbieri and Jose Camacho-Collados. How gender and skin tone modifiers affect emoji semantics in twitter. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 101–106, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2011. URL https://www.aclweb.org/anthology/S18-2011.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2061. URL https://www.aclweb.org/anthology/P18-2061.

Sweta Karlekar, Tong Niu, and Mohit Bansal. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2110. URL https://www.aclweb.org/anthology/N18-2110.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL https://www.aclweb.org/anthology/W18-5102.

Timothee Mickus, Olivier Bonami, and Denis Paperno. Distributional effects of gender contrasts across categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 174–184, 2019. doi: 10.7275/g11b-3s25. URL https://www.aclweb.org/anthology/W19-0118.