THE UNIVERSITY OF CHICAGO

MEASURING PERCEPTIONS AND MITIGATING BIAS IN TEXT AND VOICE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
JENNIFER ROSE CRYAN

CHICAGO, ILLINOIS
AUGUST 2023

"Out there things can happen

and frequently do

to people as brainy

and footsy as you."

— Dr. Seuss

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

To Ben + Heather - it's been quite a long journey, but absolutely would not have been possible without the care + effort you both put towards my potential for success. Thank you. To Marge + Ralph - your endless endurance and vivacity continues to inspire me. I will remain forever grateful for your love + support. To Shannon - thank you for always leading the way with power + grace. To Georgia - love you bunches. To Alex - thank you for always reminding us every moment is better with a little laughter. To Steve + Susan - thanks, for everything. You know better than me all of the things you've done that allowed me to get to where I am. It's all very much appreciated. To Marie + Vivon - thanks for being the best gypsy adventure partners out there. You inspire me daily to continue pushing the limits of what's possible in life and never settle for anything less than. To Bruno, Yuli, Pranav, Jean, Andrew, Suhail, Nik, Areej + others in my cohort - thank you for your friendships and good times through all the ups and downs. To Zhujun, Emily, Shawn, Bolun, Yanzi, Shiliang, Kevin, Xinyi, Zhuolin, Huiying, Yuxin, Wenxin + others - thank you for all of the knowledge and experience passed along, and the lovely times we've shared in Chicago, Santa Barbara, and elsewhere around the world.

ABSTRACT

Our ability as humans to effectively communicate depends heavily on the language we use and the way we speak to one another. The values of our society are both reflected in and reinforced by our use of language. Detecting how language could reflect biases needs to remain effective as these values evolve over time. This dissertation evaluates methods to measure how people perceive written text and spoken voice, how these perceptions may perpetuate societal stereotypes, and how to prevent biased perceptions.

Specifically, gendered language in text often affirms gender stereotypes and perpetuates bias and discrimination. As readers absorb written content, gendered language used settings such as biographies, recommendation letters, and job advertisements can negatively impact the subjects. Gender stereotypes have been studied extensively, however, the current methods used today still rely on word banks from nearly 50 years ago. Since then, societal views have continued to evolve and it's important to be able to reflect these changes. Additionally, significant advances have been made in developing new methods for analyzing how words are used in larger bodies of text. To address this, I first examine how descriptive language reflects societal perceptions of gender roles. Then, I demonstrate a crowd-sourced method for updating gender lexicons to reflect modern language and train deep learning models to detect gendered language more efficiently.

In addition to written text, efficient and unbiased communication depends upon not only the content, but the manner in which it is presented. The tone of voice of a speaker can heavily influence how they are perceived (e.g., perceived trustworthiness, competence). Further, changes in emotion tone of voice can reduce biases and more effective communication. This work explores ways to improve methods for measuring perceptions of gendered language in text and emotion tone in voice, and ways to mitigate resulting biases.

CHAPTER 1

INTRODUCTION

Biases influence how we interact with each other and society at large. Everyday decisions often rely upon simple heuristics (mental frameworks) to help quickly guide us towards likely scenarios. Mental heuristics develop based on situations we experience and observe throughout our lives. Our individual heuristics, heavily informed by societal stereotypes, represent our expectations for people and scenarios. These judgements may often be harmless or even positive (halo effect). However, potentially harmful misrepresentations of others occur when judgements reaffirming negative stereotypes (conformity bias).

The ease of misreading intentions brings attention a need for better understanding initial presumptions. While biases have long been studied in psychology and sociolinguistics, methods vary greatly and are not universally agreed upon. More recently, widespread availability of large language datasets and machine learning methods provide new ways to study human interactions. Without continuous analysis of the data and methods used, they both risk becoming outdated and no longer viable. By looking closely at frequently used dataset sources and analyzing the methods that use them, we can improve understanding of modern societal perceptions in the real world. Due to the vastness of language use, I focus my work on methods to measure and manipulating gendered language in written text and emotional tone of voice.

1.1 Bias in Text

In the domain of written text, language affirming gender stereotypes are often observed in various contexts today, from recommendation letters to Wikipedia entries and fiction novels and movie dialogue. Yet to date, there is little agreement on the methodology to quantify gender stereotypes in natural language (specifically the English language). Common

methodology (including those adopted by companies tasked with detecting gender bias) relies on a gender word inventory approach largely based on psychology studies from the 1970s. Further, the methods used to evaluate gendered wording most often simply counted up feminine or masculine words, and giving an overall rating of how they balance in number. In more recent years, deep learning language models trained on very large corpora of text data show great promise for more complex analysis of large bodies of text (e.g., producing a meaningful summary of a news article). However, they have yet to be applied to the task of measuring gendered language in text.

To evaluate how well this lexicon represents modern perspectives, I first analyze how gendered language appears in society today, and validate associations between gender and words via crowd-sourced ratings. The updated lexicon includes thousands of words with gender score ratings, and can then be applied to machine learning classification models applied to bodies of text (e.g., news articles). However, this method remains simplistic and does not account for the context the words are used in, because it still relies on single word tokens. To fully reexamine gender stereotype detection in the context of modern tools, this work then comparatively analyzes the efficacy of lexicon-based approaches and end-to-end, ML-based approaches prevalent in state-of-the-art natural language processing systems. In an attempt to account for words in context (as opposed to single words), crowd-sourcing of paragraphs and articles provided additional ratings of entire bodies of text. Overall, the use of a large language dataset shows that even compared to an updated lexicon-based approach, end-to-end classification approaches are significantly more robust and accurate, even when trained by moderately sized corpora. This work demonstrates the need and significance of updating data and methodology, and how it can drastically improve the way we measure how gendered language is used.

1.2 Bias in Voice

The way humans express themselves through language includes not only the words used, but also how they say it. Human speech also carries with it factors such as tone, emotion, and intention. Listeners, on the other hand, may perceive some or all of these factors, but also properties about the speaker, including external characteristics (e.g., race/ethnicity, gender appearance, socioeconomic status, regionality) and personal characteristics (e.g., emotion, confidence, trustworthiness). This work seeks a better understanding of how listeners perceive these voice biases in speech, and the feasibility of altering such perceptions using machine-learning based tools for speech processing. Traditional models for voice conversion are typically developed without consideration of real-world applications, and how the alterations in voice affect perceptions. In addition to voice conversion models, the datasets they are trained on vary widely in quality, source, and methods of establishing and/or validating the "ground-truth" data labels.

Through a combination of user studies and experimental voice processing, I attempt to understand: can ML-based tools correctly detect human emotion and manipulate human emotion in speech; how do these alterations impact the perceptions of the speaker by human listeners; and how willing are users accept ML-based voice alteration software as tools for reducing voice bias. User studies show that people do not universally interpret the same emotion from the same audio, and therefore emotion speech datasets should not be used as universal ground-truths. Though, reducing emotion tones (i.e., anger, surprise) does show increases in perceptions of competency and trustworthiness, and are preferred in several real-world scenarios. However, people show strong skepticism of the concept of alter one's voice due to fears of deception and misuse. Overall, while emotion voice conversion models show the ability to improve conversation experiences, researchers should maintain awareness of people's perspectives, and use caution when implementing such tools in the real world.

Methods and datasets require continuous reflection and updating to represent evolving real world perspectives. Leveraging deep learning methods improves understanding of perceptions of ourselves and others, and thereby how we can reduce the perpetuation of biases.

CHAPTER 2

VERIFYING TRADITIONAL METHODS OF MEASURING GENDERED LANGUAGE IN WRITTEN TEXT

2.1 Introduction

The values of our society are both reflected in and reinforced by our use of language. In that context, sexism and gender discrimination is often perpetrated and reproduced through lexical choices in everyday communication. Recent studies have identified descriptions that reflect gender stereotypes in different types of articles, such as biographical pages of notable people [181], recommendation letters [104], fictional stories [55] and movie dialogue [140]. These issues are further exacerbated today by the ubiquitous usage of machine learning tools in language processing. We know that machine learning algorithms often translate and incorporate gender biases from training data [168], and such biases have been proven in popular techniques including word embeddings [23, 30], coreference resolution [197] and sentence encoders [109].

While the case is obvious for accurately identifying gender stereotypes in language, today's tools for this process are woefully lacking. Gender stereotypes are traditionally captured by a gender word inventory: a pre-compiled word lexicon which contains items describing social traits and behaviors that supposedly differentiate male or female genders. The original lexicon was a bag of words hand-picked from a survey of psychology students in 1974 [17]. These word banks have been further extended in later studies to detect gender bias in job postings [62, 170]. Though widely utilized, items in traditional gender word inventories are less endorsed by women in recent years [177, 48], and their efficacy in detecting gender stereotypes in language remains unclear.

Meanwhile, today's natural language processing tasks are dominated by an end-to-end approach using deep neural network models. Instead of using a pre-compiled word lexicon,

the end-to-end approach trains a neural network model that produces the desired output using labeled raw text as input. This approach has shown great success in most NLP tasks such as sentiment analysis [81] and hate speech detection [14], which were traditionally solved by a lexicon approach [169, 45]. Unfortunately two risks remain with this approach: a) these models often require tens of thousands of samples to train an accurate model, and b) bias can seep into training data and affect detection results.

The goal of this work is to empirically analyze different approaches to the problem of detecting gender stereotypes in natural language, in order to understand the best methodology for ongoing and future studies. More specifically, we are interested in three general questions. *First*, can we update the traditional lexicon-based approach to reflect gender stereotypes in modern society, with modern machine learning tools at our disposal? *Second*, is it feasible to build a gender stereotype detection model using the end-to-end approach? Given the dependence of deep learning models on large training sets, how accurate can this approach be given moderately sized datasets? *Third*, how do these two approaches (lexicon-based and end-to-end deep learning) compare in practice? What accounts for the differences in their accuracy results? The ultimate goal is to develop methodology guidelines for identifying gender stereotypes moving forward.

This study consists of two key components: building a gender stereotype lexicon (an update to the lexicon approach), and a careful empirical comparison between the lexicon approach with the end-to-end deep learning model. First, to build the gender stereotype lexicon, we begin by extracting verbs and adjectives that are used to describe humans from English Wikipedia. After selecting a set of frequently used words, users are asked to evaluate the masculinity and femininity of each word. Then a supervised learning approach is applied, using user-labeled words to generate scores for all remaining words, resulting in a gender stereotype lexicon that contains stereotype scores of over 10,000 words. Second, we build an end-to-end deep learning model by training an NLP BERT model with a dataset

of online articles marked by crowdworkers as consistent with or contradictory to common gender stereotypes. By comparing the end-to-end and lexicon approaches, findings show the end-to-end models significantly outperform. Finally, the results are manually analyzed to understand underlying causes of misclassifications for the lexicon-based approach.

This work makes three key contributions:

- We develop a robust gender stereotype lexicon reflecting modern language and interpretations, using a combination of data-mining, crowd-sourcing, and supervised learning using linear classifiers. We also empirically show that existing unsupervised methods fall short in comparison on accuracy measures.
- We collect the first human labeled text corpus (4,333 articles) for gender stereotypes, and use it to train an end-to-end, deep learning classification model based on the BERT language representation tool.
- We evaluate both approaches using a secondary user study, and find that our end-to-end approach significantly outperforms our lexicon approach in its ability to recognize gender stereotypes. We manually study errors made by the lexicon classifier, and identify key underlying reasons for those errors.

Hopefully these results will inform best practices moving forward for detecting gender stereotypes in text.

2.2 Background

Gender Stereotypes in Language. Gender stereotypes are common beliefs about what men and women's physical and personality traits are and should be like. According to traditional gender stereotypes, women should display *communal* traits (e.g., nice, caring, warm) and men should display *agentic* traits (e.g., assertive, competent, effective) [51, 43].

Gender stereotypes emerge in language choices used in written and verbal communication

[111]. It has been found that a category label used to refer to a group triggers mental connections with characteristics stereotypically associated with the group, even in supposedly unprejudiced people who do not explicitly endorse the stereotype [103, 111]. This also applies when the category label is one's gender. For example, after primed by words consistent with gender stereotypes (e.g., "nurse"), people are faster to associate gender pronouns (e.g., "she") with the corresponding gender (e.g., "female") [15].

As a result, gender stereotypes are common in contemporary languages, both in written and spoken communication. For example, in fiction writing, traditional gender stereotypes such as dominant men and submissive women are common throughout nearly every genre, regardless of the gender of the author [55]. On Wikipedia, articles about notable women emphasize more on romantic relationships or family-related issues compared to articles about notable men [181]. When writing recommendation letters for faculty positions, women are often described as more communal and less agentic than men [104]. Additionally, in movie dialogue, male characters use more words related to achievement than female characters [140].

Gender Word Inventory. Stereotypes can be captured by *gender word inventories* – pre-compiled lists of items describing social traits and behaviors that differentiate males and females [17, 147]. Gender word inventories are historically extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [164]) and Bem Sex Role Inventory (BSRI [17]) are two of the most representative questionnaires in early studies. The items extracted for the BSRI and PAQ typically associate females with more communal attributes (i.e., gentle, warm) and men with more agentic attributes (i.e., aggressive, competitive), which are highly consistent with traditional perceptions regarding gender stereotypes. Other studies generalized these words into *expressive* and *instrumental* traits [159]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered

wording in job advertisements [62].

These gendered word inventories are traditionally used as a way to measure gender role self-concepts, i.e., whether people see themselves as masculine or feminine. Among them, BSRI is considered as a golden standard in gender role evaluation, and has been used in thousands of studies in the more than 40 years since it was developed [46]. However, perceptions captured by BSRI are less endorsed by women in recent years [177, 48]. These works reviewed a large collection of studies that apply BSRI, and tracked how user responses change over a long period of time. Women's femininity scores have decreased significantly over the years, indicating that societal gender norms may require an update of masculine and feminine stereotyped characteristics.

Given previous results showing that existing gender word inventories may not properly reflect these concepts in the modern world, we seek to develop a lexicon that captures people's perceptions of gender stereotypes in contemporary society.

Gendered Stereotype Studies in NLP. There are few tools or algorithms to determine if any piece of text perpetuates modern gender stereotypes. One class of tools shares some similarity to ours are tools used to detect gender biased language in job advertisements [170]. These tools leverage precompiled lists of gender biased words aggregated from previous psychology studies [62] that may affect decisions of job applicants, and calculate gender bias of a job advertisement based on the number of occurrence of these words.

Although detecting gender stereotypes in natural language is still an under-explored area, the NLP community has increasingly focused on issues of fairness and bias in NLP models. Many projects focus on identifying and removing biases in algorithms, e.g. in word embeddings [23]. Prior studies also observed performance discrepancies across genders in systems including coreference resolution [197], image caption generation [74], and sentiment analysis [131]. Such biases can be mitigated by creating an augmented dataset that counters gender bias in the original training dataset [131], or adding constraints during model training

to enforce gender neutral prediction [198].

2.3 Motivation for Updating Data and Methods

Understanding the evolution of characteristic gender differences in language helps to mitigate the effects of masculine and feminine stereotypes, thus reducing gender bias. One domain where gender biased language can have direct real-world effects is job advertisements. Gendered language can negatively affect both the employers and employees alike, by deterring applicants based simply on the wording used to describe the position. Both parties miss out on significant opportunities when qualified applicants don't even apply. To remedy this, several companies provide tools for measuring gendered language in job advertisements, with the specific goal of reducing problematic language and increasing applicant rates. However, initial observations of these tools in the wild suggest they likely rely on outdated data and methods, resulting in inadequate tools. By leveraging the availability of job advertisements posted online and these commercially available tools, it's possible to examine *recent real-world data* that contains and evaluates gendered language. In this case study, I *measure modern gendered language use* in job advertisements, and *confirm that modern commercial tools rely on rudimentary methods and simple tokens derived from very old gender word lists.* From this analysis of results, I **determine the necessity of updating both data and methods used to measure gendered language**.

2.3.1 Case Study: Measuring Gendered Language in the Job Market

For millions of workers, online job listings provide the first point of contact to potential employers. As a result, job listings and their word choices can significantly affect the makeup of the responding applicant pool. Here, we study the effects of potentially gender-biased terminology in job listings, and their impact on job applicants, using a large historical corpus of 17 million listings on LinkedIn spanning 10 years. We develop algorithms to detect

and quantify gender bias, validate them using external tools, and use them to quantify job listing bias over time. We then perform a user survey over two user populations (N_1=469, N_2=273) to validate our findings and to quantify the end-to-end impact of such bias on applicant decisions. Our findings show gender-bias has decreased significantly over the last 10 years. More surprisingly, we find that impact of gender bias in listings is dwarfed by our respondents' inherent bias towards specific job types. Despite recent strides made by women in the workplace, workplace inequality persists [28, 67]. In addition to the widely distributed reports of wage inequality across genders [20], it is also well known that there are significantly fewer women in male dominated positions, across both industry sectors (e.g., technology [117]), and job types (e.g., corporate executives [162]).

A number of contemporary theories have hypothesized the source of these gender imbalances, whether they come from different educational paths and biases introduced early on [151, 110, 40], or whether they are the result of workplace attrition [75, 175, 42]. The job search and hiring process might be another contributor [107, 165, 143, 58, 19, 124, 142]. Hidden biases often creep into job listings [11, 18, 62, 71] , and can either actively or passively discourage certain applicants from applying.

The goal of this study is to understand the role that job postings play in introducing or exacerbating gender imbalance in the workplace. More specifically, we are interested in two general questions. First, how significant is gender bias in today's job listings? How much does gender bias vary over different industry sectors, and how has it changed over time? We hope to track levels of gender bias in job postings through time, to see if any changes are reflected as a result of society's growing awareness of gender bias. Second, we look to measure the end-to-end impact of gender bias in job posts on the actual decisions of potential job applicants. Are applicants aware of such bias in job posts, and do these biases play a role in their decisions to apply for jobs?

To answer these questions, we perform a study with two components, an empirical, data-

driven component that quantifies the presence and magnitude of gender bias in job postings over the last 10 years, and a qualitative user-study that seeks to understand the end-to-end impact of biases on whether applicants apply to a posted position.

On the empirical side, we analyze 17 million job posts collected over the last 10 years (2005–2016). We obtain this dataset of job posts through a near-complete download of job posts maintained by LinkedIn, the largest professional networking site online with 500 million users. To quantify gender bias over large datasets, we develop two scalable algorithms that match the same metrics as online services that evaluate job postings for gender bias, Textio and Unitive. We tune our algorithms and show they can approximate Unitive and Textio in bias classification, generating a raw score and a normalized score of gender bias between feminine and masculine. We use a test sample of key words/phrases to validate our approaches against Unitive and Textio. We then apply our algorithm to the LinkedIn dataset to quantify gender bias in the whole market, specific sectors, and its changes over time.

In our user study, we augment our data analysis with a user survey that captures how gender bias in wording actually affects job applicants. We ask detailed questions to 2 user populations, 469 Amazon Turk workers, and 273 undergraduate college students, to understand their perceived levels of gender bias in our job posts, its correlation to specific gendered keywords or phrases, and the ultimate effect they have on the job application decision.

Our analysis generated several key findings.

1. There is significant gender bias in job listings, but bias has been dropping significantly over the last decade, led by specific job sectors which now trend feminine.
2. Changes in bias levels vary significantly over different sectors, driven by significant changes in usage of a small number of heavily gendered keywords.
3. Our user study shows that users do indeed detect gender bias in job postings, consistent with bias detected by our algorithms.

4. Observed bias still had low levels of impact on user decisions to apply or not apply for a specific position, and there was more sensitivity to bias by men than women.
5. Surprisingly, we observed that users had strong internal biases which played significant roles in their decision on whether they would apply. These biases played a much bigger role than any gender bias language we observed.

To the best of our knowledge, our study is the first large-scale study to look at longitudinal shifts in gender bias in job postings. While our study has clear limitations (lack of historical advertisements from non-LinkedIn job sites, and potential sampling bias in our user study), we believe our results shed light on an important component of the debate on gender equality in the workplace.

Gender Equality in Job Market. Historically, certain industries have been dominated by one gender over the other. Approximately one-third of men and women work in occupations with a workforce comprised of at least 80% males or females, respectively. In the past, men tended to dominate engineering and construction occupations while women consistently dominated clerical assistant and teaching occupations [126, 127]. Although census data shows an increase in overall participation of women in the workforce throughout recent decades, the disparity among genders across particular industries remains over time [60].

One reason for such disparity lies in people's stereotypes of genders and occupations. Research shows that people are most likely to seek out occupations that are compatible with one's sense of self [65], suggesting that people are less likely to seek out occupations in industries dominated by the opposite gender. Moreover, a study of perceptions across 80 occupations found that people assume stereotypes associated with the gender of a worker must correlate with the requirements of their occupation. In particular, both genders perceived that masculine physical and agentic qualities were associated with more prestige and earnings [35]. Across industries, managerial positions have historically been perceived as requiring masculine traits. Men even view women negatively for displaying masculine traits in

a management role because it is regarded as inconsistent with female role expectations [12]. In the field of Information Technology (IT) and Information Systems (IS), general stereotypes skew towards masculine traits, due to men consistently dominating the field [83, 176].

Another source of gender disparity is institutional discrimination in job markets, which has been observed in both traditional [58, 124, 19, 142] and online settings [71]. In a lab experiment that simulated a hiring decision process, male participants displayed a strong tendency to choose male candidates, even if a female candidate appears as a slightly better performer [58]. In another field study, comparably matched men and women are sent to apply for jobs in restaurants, and the study found female applicants were significantly less likely to get an offer from high-end restaurants [124]. Later, a similar study was conducted in a male-dominated occupation (engineer) and a female dominated occupation (secretary). Results show statistical significant discrimination against women in the male-dominated occupation and against men in the female-dominated occupation [142].

Another line of research shows that significant improvement in gender equality has been made over of last two or three decades [34, 115, 20]. The overall wage gap between two genders has declined considerably [20], and no institutional discrimination can now be observed in academia [34, 186].

Detecting Gender Bias by Word Analysis. Gender stereotypes are embedded in language use, i.e., different word usage patterns when writing about males or females. Significant prior research used text analysis to examine gender differences in a number of contexts. Some examined how men and women use language differently in text and conversation [125]. Other work studied how text analysis algorithms express unintentional bias, and detect occupational stereotypes in text, i.e., suggesting sexist analogies such as men are analogous to computer programmers and women analogous to homemakers [23]. Other work showed when writing recommendation letters for faculty positions, more standout adjectives are used to describe male applicants than female applicants [157], and women are described as more

communal and less agentic (assertive or competitive) [104]. In the context of job advertisements, researchers have shown that language used not only reflects such stereotypical views, but also reinforces the imbalance [18, 11], and that a conscious effort toward gender-fair language can help reduce it [160, 76]. Finally, much of the prior work in text classification rely on a supervised model with pre-labeled datasets, and is summarized nicely in a survey by Aggarwal et al [3].

Stereotypes can be captured by *gendered words* – terms describing socially desirable traits and behaviors of male or female genders [17, 147]. Gendered words are usually extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [163]) and Bem Sex Role Inventory (BSRI [17]) are two of the most representative questionnaires in early studies. The words extracted from BSRI and PAQ more typically associate females with more *communal* attributes (i.e., gentle, warm) and men with more *agentic* attributes (i.e., aggressive, competitive). Others generalized gendered words into *expressive* and *instrumental* traits [159]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered wording in job advertisements [62]. Finally, Donnelly et al. found that women in recent years are less likely to endorse traditionally feminine traits in BSRI [48], indicating that gender norms may require an update of the masculine and feminine stereotyped characteristics.

Based on the lists encoded by Gaucher et al. [62], online services like Unitive[1] and Textio[2] use the words and phrases as a baseline to develop gender-neutralizing algorithms with help from machine learning classifiers. The algorithms are trained on internal application and hiring data, and aim at finding gendered wording in job advertisements before recruiters

1. `http://www.unitive.works/`

2. `https://textio.com/`

post online. These services represent the state-of-the-art for identifying gendered wording in job advertisements. Since these services run commercial, proprietary algorithms, it is cost prohibitive to evaluate our large job post dataset through their services. Instead, we designed our own algorithms using similar methodologies, and validate them against these online services using samples of test data.

Comparisons to Prior Work. The focus of our study is using large-scale data analysis to characterize gender bias across a comprehensive, longitudinal dataset. Our work was initially motivated by [62], which studied 4,000 job listings (most in a university setting) to characterize gender bias in job listings as an institutional-level mechanism of inequality maintenance. In contrast, we broadly characterize gender bias at scale, using a large dataset that consists of 17 million online job ads covering more than 140 industries. Our work also focuses on examining shifts in gender bias over ten years, and the impact it has on potential applicants' decision.

More recent work [71] identified gender/race discrimination on (performance) reviews in the online freelance (gig) marketplace, by correlating the review with the worker's gender and race. While their work targets reactions to worker output, ours focuses on job advertisements written by only the employer. While [71] analyzed keywords in review, their analysis was limited to sentiment analysis that identifies the attitude of the review, not gender bias.

2.3.2 Data and Initial Analysis

Data Collection. We collected a large sample of job advertisements from LinkedIn job posts over 10 years (from 2005 to end of 2016). LinkedIn job advertisements are fully public, and accessible online to any user without requiring account registration with LinkedIn. To retrieve a job advertisement, we simply queried known URLs and downloaded the webpages. LinkedIn keeps job advertisements available online for browsing even after the application window has closed. This allowed us to collect a significant longitudinal job advertisement

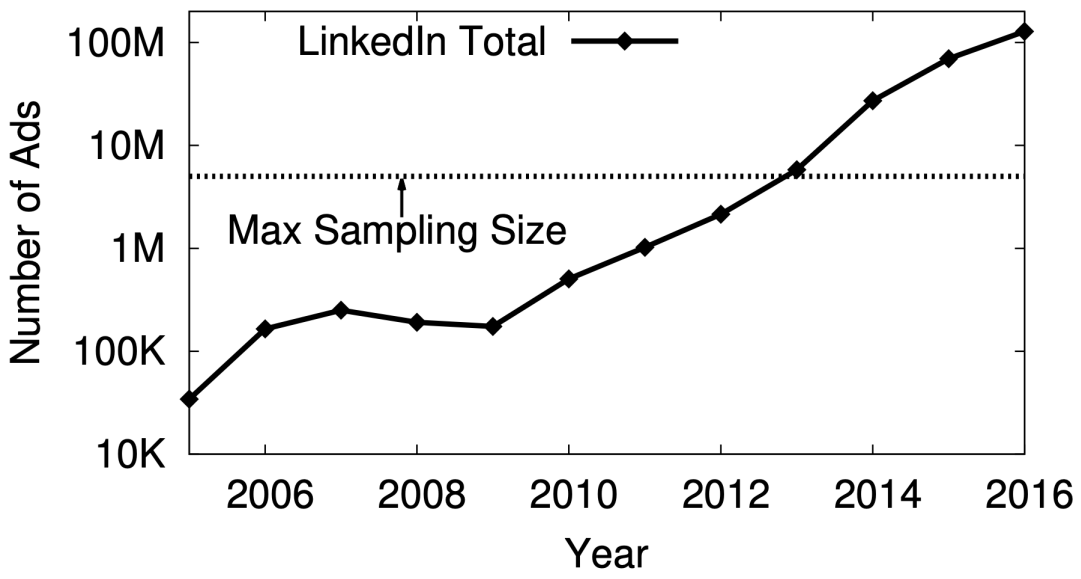


Figure 2.1: Statistics of the number of job ads per year.

dataset.

Job advertisements on LinkedIn are each assigned a unique ID, which increases monotonically over time. By the end of 2016, the maximum possible ID on LinkedIn reached above 253M, which means there are at most 253M job posts on LinkedIn. Since we limited our online query rate to avoid overloading LinkedIn's online services, we did not crawl all 253M job postings. Instead, we randomly sample 5 million job post IDs from each year, and only fetch job advertisements matching these IDs. Note that job listings in each year before 2013 contained less than 5 million ads. For these years, we fetched all available job advertisements. After filtering out job advertisements in languages other than English, our dataset contains 17,376,448 job advertisements in total.

Each job advertisement contains a job title, company name, company location, and the main descriptive content of the advertised position. In addition, LinkedIn also provides metadata, including job industry, job function, employment type, and seniority level. LinkedIn has 147 unique job industries[3], which are then further mapped into 17 sector groups, and 35

3. `https://developer.linkedin.com/docs/reference/industry-codes`

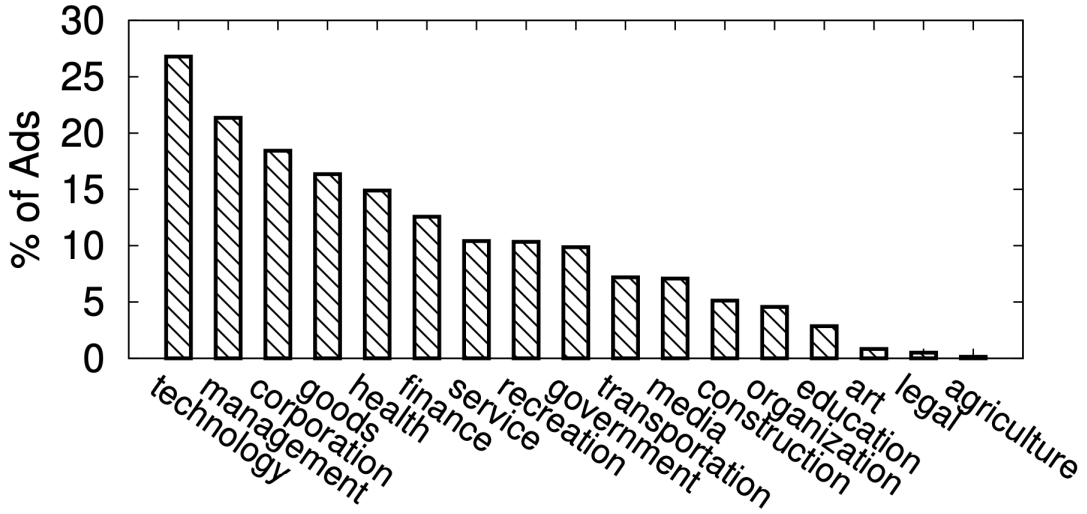


Figure 2.2: Breakdown of LinkedIn data by sector groups.

job functions which describe what activities a person is undertaking. Employment types includes 6 categories: full-time, part-time, temporary, contract, volunteer and other. Seniority level indicates the rank of the position within the business, ranging from entry-level (lowest) to executive (highest).

2.3.3 Preliminary Analysis

Number of Job Advertisements. We plot the number of LinkedIn job advertisements posted per year in Figure 2.1, as inferred by the total number of possible job IDs in LinkedIn matching a given year. For years up to 2013, our dataset closely follows that of the plotted LinkedIn results, with a small number of missing posts due to non-English listings and unavailable data errors for some of the oldest job postings (likely due to corrupted data). For years 2013–2016, we limited our sample set to 5 million postings per year. While our dataset captures only a limited sample from 2013 onwards, we believe a randomized sample set of 5 million ads is sufficient to capture a representative sample of job postings in any given year.

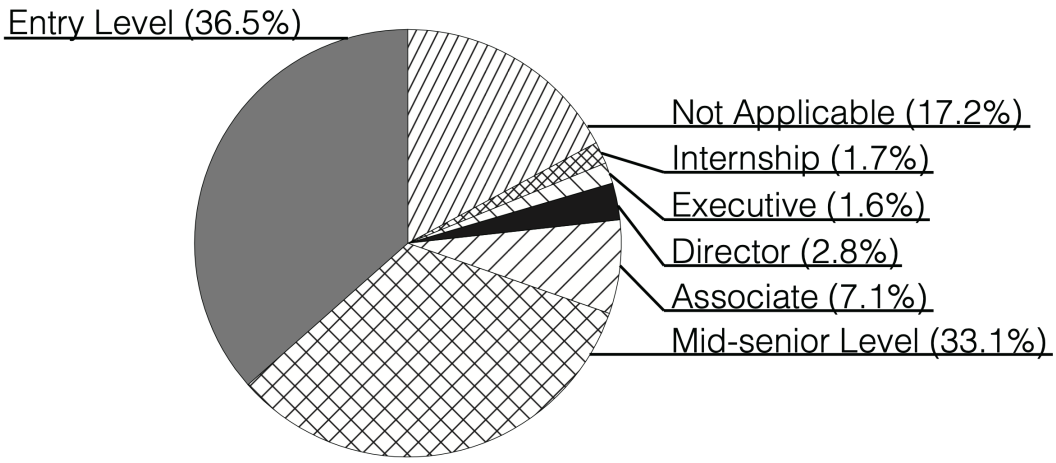


Figure 2.3: Breakdown of LinkedIn data by employment type.

Sector Groups, Employment Type and Seniority Level. Next, we plot distribution of important metadata fields. Figure 2.1 shows the number of job posts in different job sector groups[4]. We found significant variation among the sizes of different groups: over 25% job posts belong to the largest group (technology) while less than 1% job posts in smallest group (agriculture). As for employment type (Figure 2.3), most (91.7%) job listings seek full-time employment, while the rest are mostly split by part-time, contract and temporary (i.e., seasonal). After 2013, LinkedIn introduced Volunteer as a new job sector, which accounts for a negligible portion of our total dataset. For seniority level (Figure 2.2), we found a trend of fewer number of applicable job advertisements at higher levels of seniority. This matches our intuition about hierarchies in the job market.

2.3.4 Quantifying Gender Bias

Our goal is to perform a large-scale analysis of the presence of gender bias over a large corpus of job listings covering the last 10+ years. Our first task is to develop a scalable algorithm

4. If a job belongs to multiple groups, we count the job in each group it belongs to.

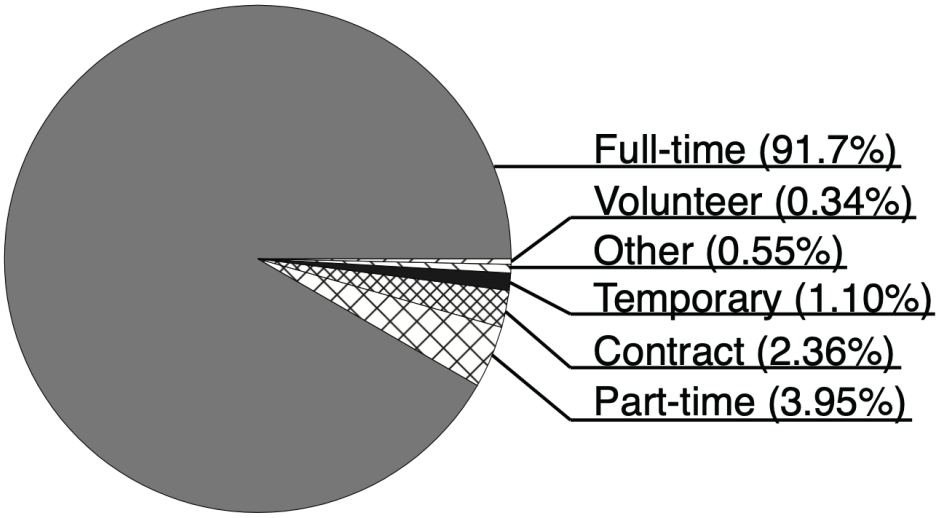


Figure 2.4: Breakdown of LinkedIn data by seniority level.

to accurately quantify gender bias. In this section, we describe how we emulate the gender bias detection algorithms of two state-of-the-art recruitment assistance services, *Textio* and *Unitive*. We validate our approach by comparing our results against theirs on a small sample of 8,000 job ads. We follow up these results in the next section with a confirmatory user study.

Gender Bias Detection Algorithms. To develop a scalable gender bias detection algorithm, we start by developing metrics to accurately capture different aspects of gender bias. For guidance, we look to the two state-of-the-art recruitment assistance services that measure gender bias, Textio and Unitive. Textio and Unitive are the two largest web services today designed to help potential employers write better job advertisements. Each company curates their own algorithm to calculate whether a given job advertisement expresses masculine, feminine, or gender neutral language. The algorithms draw from an established baseline starting with gendered word lists [62]. We observe and adopt the metrics used by these two services, which we refer to as *Gender Target* and *Gender Tone*. Gender target follows Textio's

methodology, which measures the intended audience gender reflected by a job advertisement, and falls into the range of -1 to 1, where -1 means the advertisement specifically targets male applicants, 1 means the advertisement specifically targets female applicants, and 0 means no gender preference is detected. Gender tone follows Unitive's methodology, which captures the extent to which a job advertisement is feminine- or masculine-phrased. It captures a cumulative effect, thus has no fixed range. A gender neutral advertisement has a tone of 0; the more masculine traits stated, the more negative the tone score is, and the reverse for feminine traits. We use the term *gender score* to refer to both metrics.

Gender target highlights feminine and masculine language in job listings. We calculate gender target by first calculating the number of gendered words, with feminine and masculine words canceling each other out. Job ads containing more feminine words than masculine words are considered to be targeting a female audience, and a final score is calculated by applying a sigmoid function on the remaining word count. The same procedure applies when masculine words outnumber feminine words, except that the result of this sigmoid function is reversed to fit into the range of -1 to 0. Finally, when the job ad has a perfectly balanced word count, it is considered to be gender neutral, with gender target score of 0.

In contrast, when calculating gender tone, we first categorize terms as *inclusive* (appealing) or *exclusive* (problematic). Prior research [62, 12] has shown a direct correlation with feminine bias from inclusive language, and between masculine bias and exclusive language. In addition, instead of simply counting words, we weights each words based on how gender specific they are. For example, the word "guy" carries a stronger gender implication than "ambitious." Thus, before calculating a cumulative score, gendered words are assigned weights depending on strength of their tone. A strongly masculine word has a strong negative weight, whereas a slightly feminine word has a weakly positive weight. We then add up the weights of all gendered words used in the ad.

Our first key challenge is obtaining an up-to-date list of biased words. To begin, we

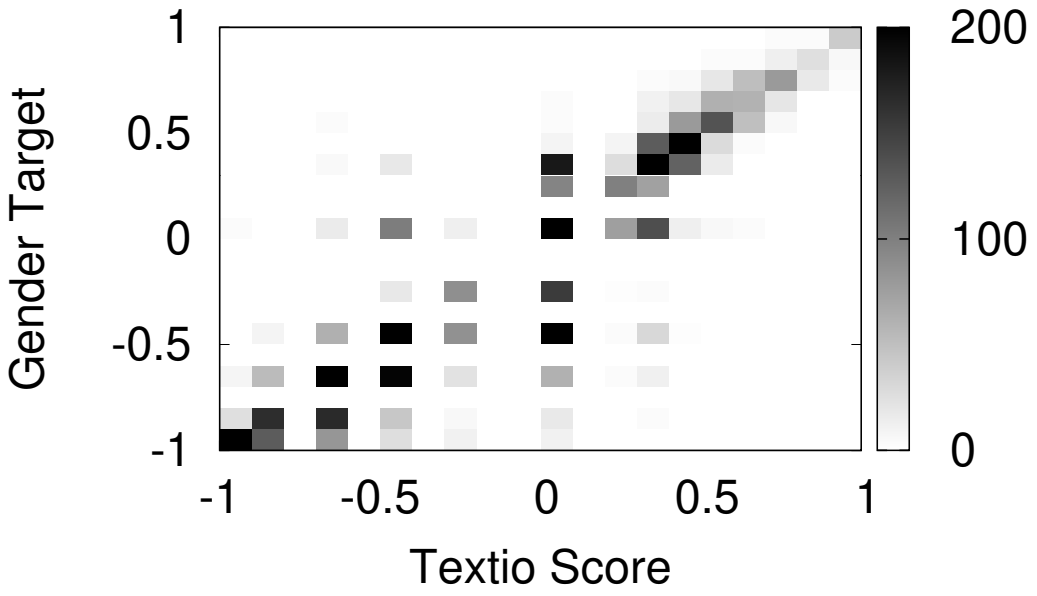


Figure 2.5: Comparison of actual and approximated gender-bias metric for Textio.

extracted 50,000 words with the highest frequency from all English LinkedIn job posts we collected, which cover 97.2% of all word occurrences. Since both services highlight words we can categorize with feminine or masculine bias, we queried the services with these words embedded in text, and examined the feedback. Textio annotated 296 gender-related key words, 150 masculine and 146 feminine. We also obtained 843 weighted keywords along with their weights annotated by Unitive, 445 with positive weights (feminine tone), 398 with negative weights (masculine tone). Note that since these two services picked their keyword independently, only 102 words overlap across services.

Algorithm Validation. We validate how well our techniques emulate these services, by comparing our results against theirs. We randomly selected 8,000 job advertisements from our dataset, and uploaded them to Textio and Unitive using their free online accounts. We compared the results they return to results from our own algorithms. We plot our results against results given by Textio and Unitive in Figures **??**. For Textio, the scores are more scattered due to the use of discrete count before normalization. We found that 71.8% of the

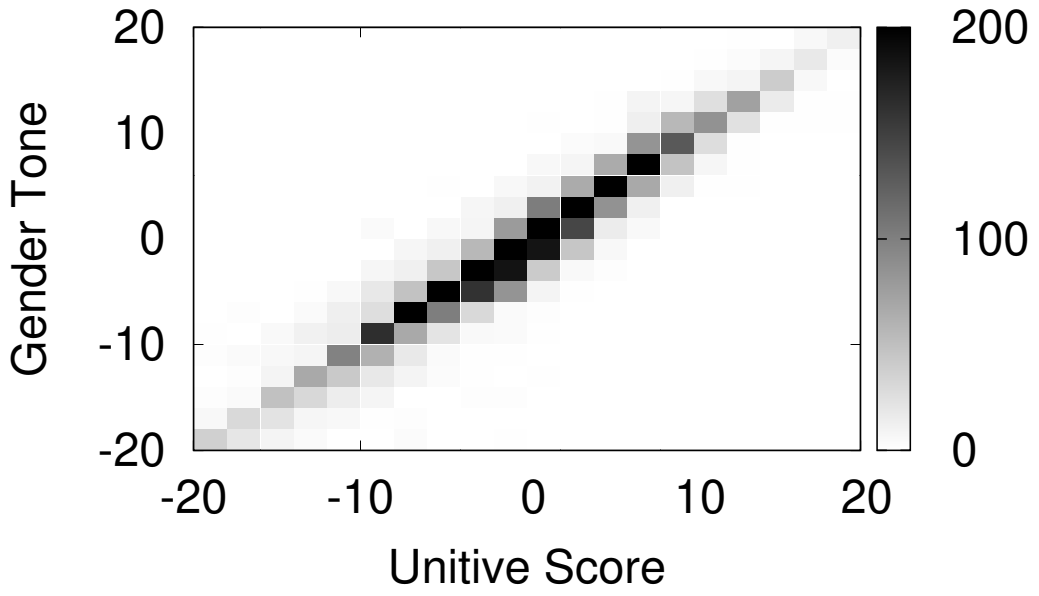


Figure 2.6: Comparison of actual and approximated gender-bias metric for Unitive.

gender target scores are within a difference of 0.1 from Textio scores (an error rate of 10%). In the case of Unitive, the scores match along a straight line with a slight bias of -0.209 towards masculine tone, and 77.5% of the scores have an error of 1 or less. Since Unitive scores varied by as much as 10, this also represents an error rate of roughly 10%.

The error in our scores is due our inability to recover full keyword lists from both services, especially for phrases. Given the highly subjective nature of gender bias, our goal is not to generate "perfect" algorithms, but to obtain general and scalable algorithms with results that approximate public systems.

2.3.5 Longitudinal Analysis

In this section, we apply our gender bias detection algorithms on to our LinkedIn job post dataset. Our results show that in recent years, wording in job advertisements skews masculine, but the absolute level of bias is becoming more neutral. First, we identify a few factors that contribute to the trend. Different job functions across industry sectors distribute un-

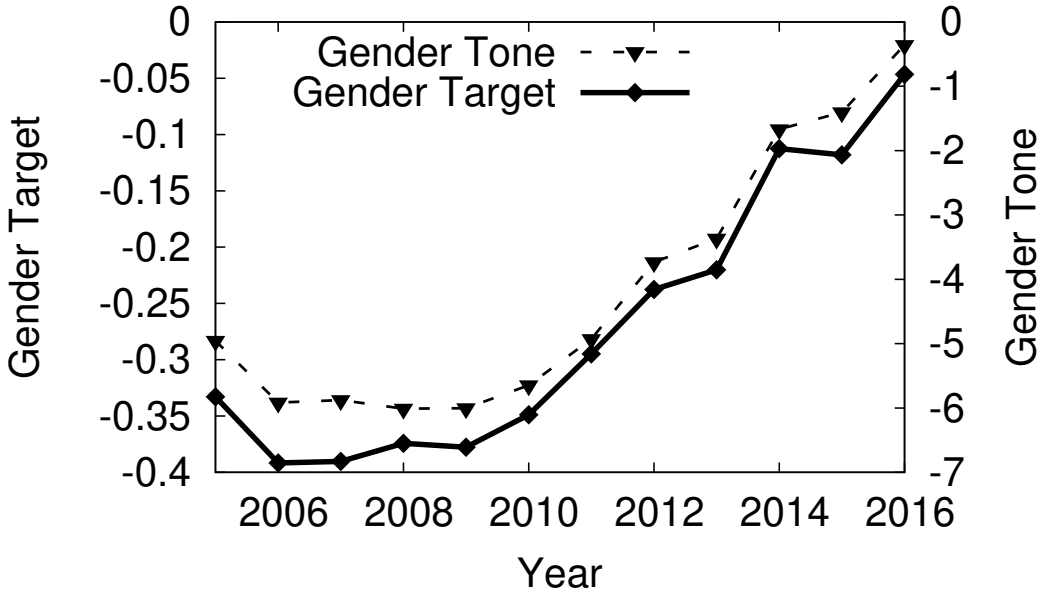


Figure 2.7: Average gender bias score from 2005 to 2016.

evenly in terms of gender score. Although this uneven distribution of jobs across sectors results in an overall averaging of feminine and masculine bias scores, the effect is limited. Second, the masculine bias comes primarily from formal and long-term employment jobs, and appears more severe in senior level positions. Over the 11-year period, the number of entry-level jobs posted is increasing over time, which partially accounts for the decreasing masculine bias, as these positions predominantly skew feminine. Third, to quantify the effect of these factors, we formulate a regression analysis to predict gender score, which shows that the effect of all factors are significant. However, there is still an underlying trend of decrease masculinity after separating out the effect of these factors, indicating possible increasing awareness of using more gender neutral language. Finally, we identify a few gendered words that contribute the most in driving change in levels of gender bias.

Gender Bias Over Time. We begin by studying how gender bias in job postings changes over time. We are interested in whether the market as a whole (and perhaps as a proxy for the general population), is becoming more aware of gender bias. We compute two values

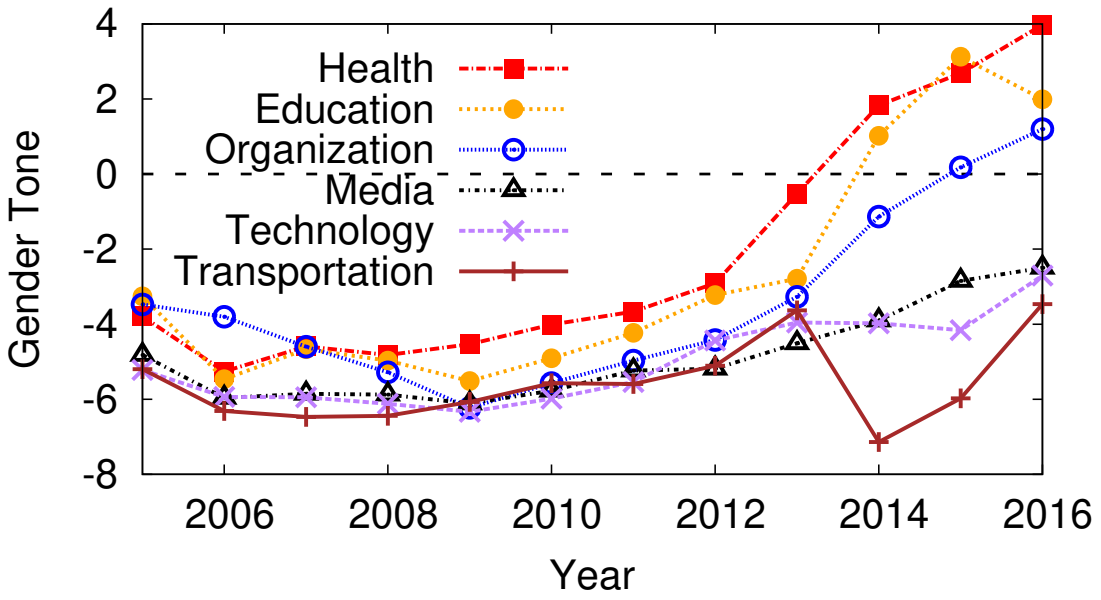


Figure 2.8: Gender tone over time of top 3 and bottom 3 sectors in 2016.

for each job posting: a *gender target* score and a *gender tone* score. For each year, we compute the average scores and standard deviation of all job postings collected from that year. Figure 2.7 uses a dual Y-axis to compare the average score of the two algorithms. The standard deviation values are similar over time and the two algorithms, thus they are omitted.

We make some key observations. First, the average gender scores remain consistently below 0 across all years, indicating that the job market, as captured by LinkedIn postings, is skewed towards masculine appealing positions. Second, an increasing absolute score over time suggests that the market is becoming more gender neutral. Third, while our two metrics use very different algorithms and their absolute scores are not directly comparable, their trends over time are almost identical. We performed another consistency check of these results using the gendered word lists from prior work [62], and the results are highly consistent. The frequencies of the three trends show strong correlation between each other (p-value $<$ 0.0001), with Pearson correlation of more than 0.97. This confirms that the overall trend is fundamental to the job market, and the two metrics capture consistent views of the same

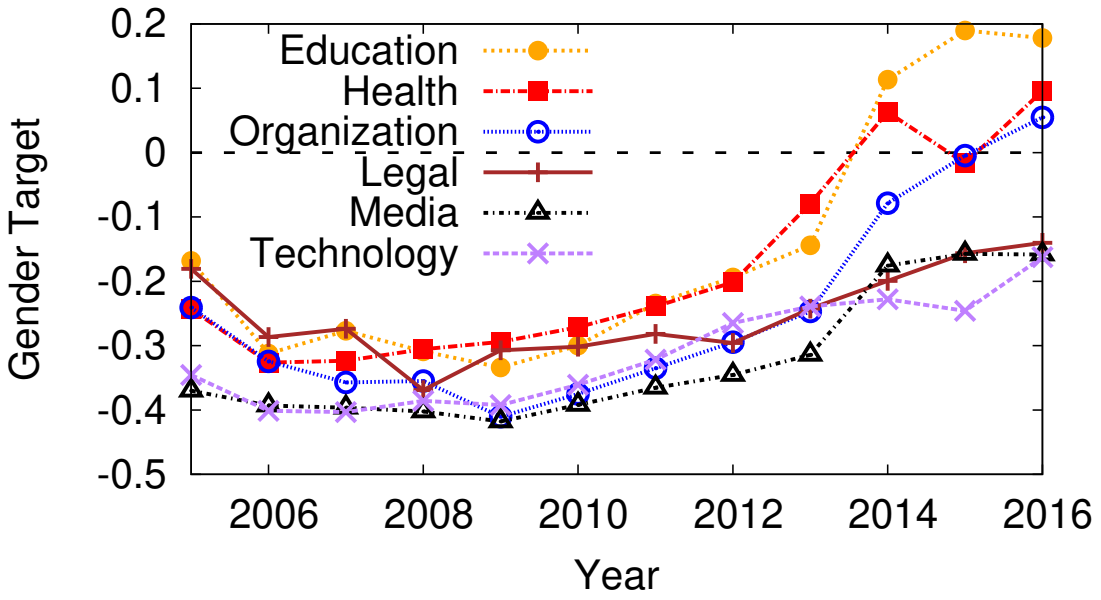


Figure 2.9: Gender target over time of top 3 and bottom 3 sectors in 2016.

phenomenon over time. To get a better understanding of where the masculine jobs originate and to explain the trend over time, we explore a variety of dimensions to better understand the underlying structure of the LinkedIn job market.

Gender Score over Job Sector Groups. We dive down to see how individual sector groups are changing over time with respect to gender bias. Recall that all together we have 147 distinct industries, mapped to 17 sector groups. While we have established that bias is decreasing over time for the entire job market, we want to observe any variance in dynamics across different job sectors.

We begin by looking at the top and bottom sectors sorted by gender scores. In Figure 2.8 and Figure 2.9, we plot the top 3 and bottom 3 sectors sorted by 2016 gender tone and 2016 gender target scores respectively. For each sector, we trace back their scores over past years. First, we note that gender tone and target are remarkably consistent in their top and bottom sectors. In both cases, Education, Health, and Organization are top sectors (more feminine), and all have risen consistently over time to current values above 0 (the dashed line represents

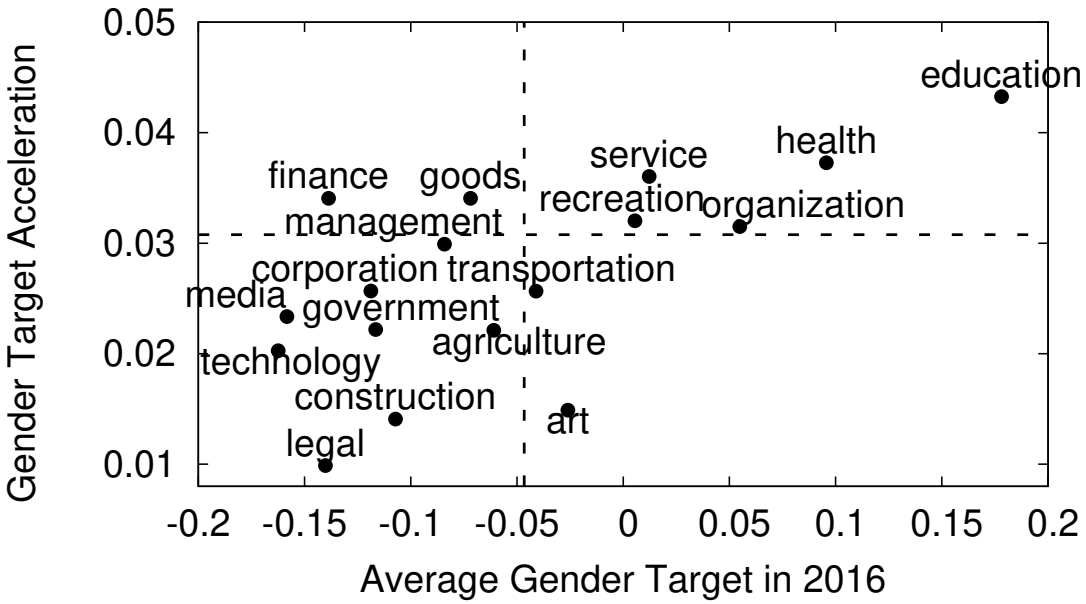


Figure 2.10: Gender target in 2016 vs. gender target acceleration across sector groups.

the value 0). Media and Technology are the two sectors that appear as bottom sectors in both metrics. Their scores are rising, albeit at much slower rates, and occasionally experience short term dips (the start of the great recession 2008–2009). The Tech sector also experiences another dip around 2013–2015, showing that perhaps the most biased sectors might be more sensitive to economic downturns.

In Figure 2.10, we plot each sector's *acceleration* of gender scores over time, against its 2016 gender target value. Acceleration is computed as the slope of a linear regression of a sector's scores over time. The results are intuitive. The sectors slowest to reduce masculine bias (Tech, Legal, Construction) still have some of the most masculine biased gender target scores in 2016. Others like Education and Health have high rates of change towards more feminine wording, and as of 2016, are firmly on the side of feminine bias.

Dynamics of sector groups are correlated with the overall increasing gender score. One reasonable question is how much dynamics between sectors contribute to the overall gender bias trend. To answer this, we first calculate the ratio of gendered job postings inside each job sector. We find that the number of jobs is increasing in several stereotypically feminine

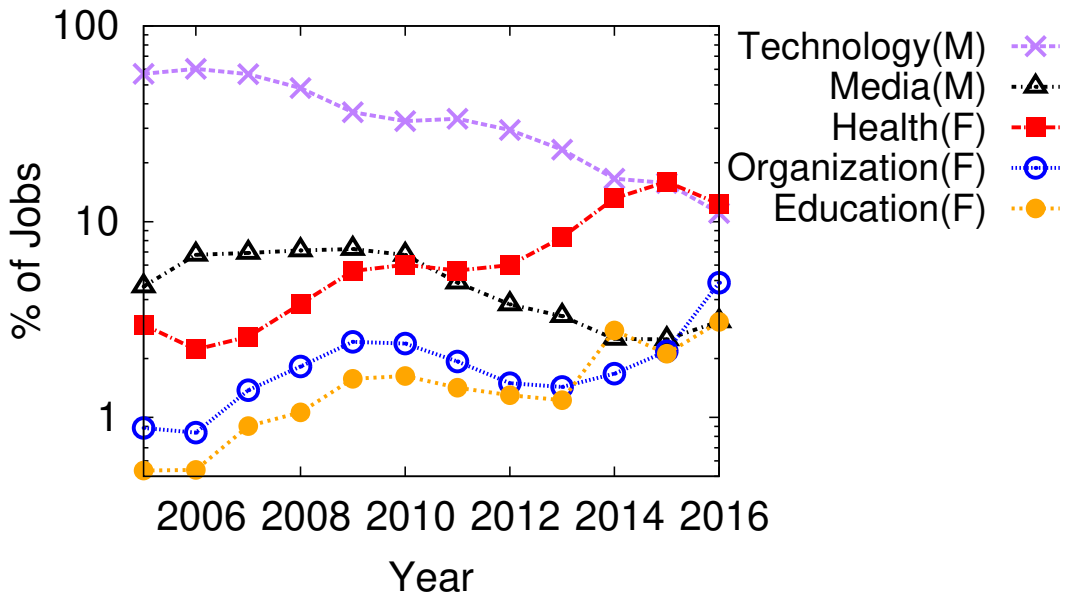


Figure 2.11: Shift in distribution of 2 typically masculine and 3 typically feminine sector groups.

sectors and decreasing in a number of stereotypically masculine sectors (see Figure 2.11). So, it is possible that the overall increasing gender score comes from a changing of sector distribution. To remove such effect, we recalculate the gender score across the entire 11-year period, but weigh each sector based only on the 2016 distribution of jobs across sectors. The result shows that the impact of shifting weights across sectors is small: gender tone only increases at most 0.56 under the new calculation (and gender target only increases by 0.034), much smaller than the overall increasing trend showed in Figure 2.7.

Gender Score over Seniority Levels. Next, we break down all job postings by their *seniority level*, and compute average scores in each category. Figure 2.12 shows the breakdown of results for different seniority levels, where there is a clear correlation between seniority ranking and the masculine tone of the job posting. We omit results of gender target here, since they show very similar trends. These results are consistent with prior work [12] that discusses how men hold an overwhelming majority of top management positions, and thus

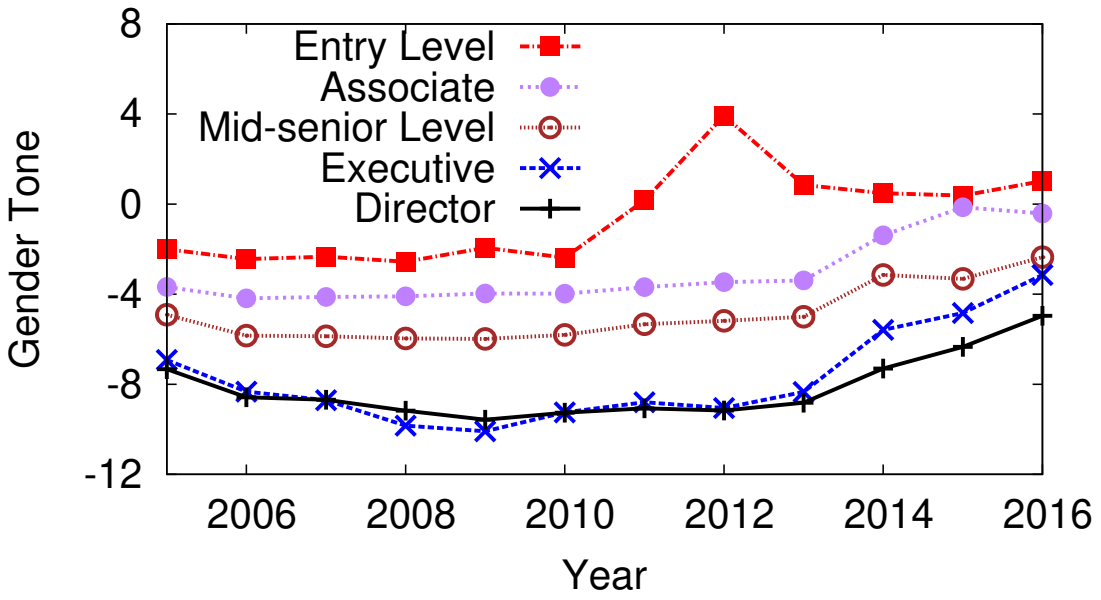


Figure 2.12: Average gender bias score of different seniority level (*gender tone*).

masculine traits are commonly associated with these higher-ranking positions. Due to a phenomenon called *ambivalent sexism* [63], attitudes regarding gender roles presume women historically belong in a domestic setting and are incompetent at holding positions of power. These unconscious biases may persist today, and are likely used to explain the gap in gender participation rates at more senior level positions.

Similar to trends across sector groups, we also find the distribution of different seniority levels changes over time (shown in Figure 2.13). It is clear that the increase in number of entry-level jobs corresponds to a decreasing proportion of mid-senior level jobs over time. Since entry-level jobs tend to be less biased towards masculine, the shift in distributions affects gender score. In Figure 2.14, we show the effect of removing this factor by computing the gender score using a fixed seniority level distribution from 2016. Compared to Figure 2.7, we get a similar increasing score trend, with a much smaller magnitude of masculinity. Thus, we conclude that the overall increasing trend comes from two parts: increasing lower-level jobs and increasing feminine language in each seniority level.

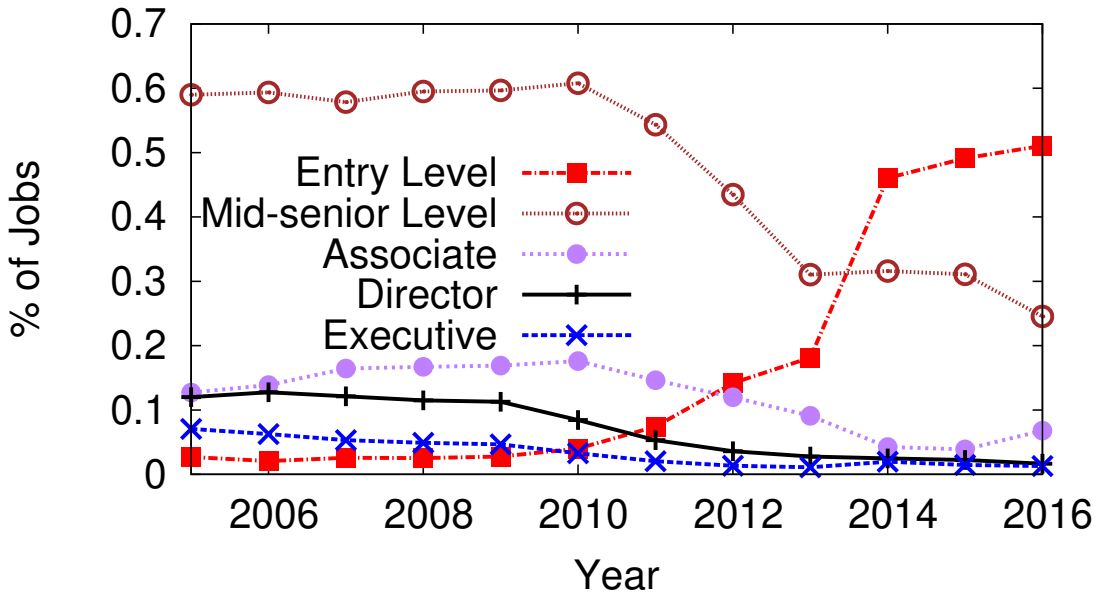


Figure 2.13: Distribution of seniority level from 2005 to 2016.

Gender Score over Employment Types. Figure 2.15 shows how the gender score is distributed over different employment types with respect to *gender tone*. Gender scores show clear and consistent trends in different employment types, and the more formal and long-term the employment, the more masculine tone in the job posting. Since over 90% of jobs are full-time jobs across all years of our dataset, we do not investigate the effect of changing distribution in terms employment types.

Comparison of Gender Bias Contributors. Now, observing how these different factors affect the job market, we aim to quantify the effect of each factor. To do so, we formulate a regression analysis. We use seniority level, year, sector group and employment type as independent variables, to predict the gender score of a job advertisement. The reference categories for seniority level and employment type are "N/A" and "other," respectively. Since a single job can belong to multiple different sectors, there is no redundancy in sector groups.

Our result is shown in Table 2.16 and Table 2.17, for gender tone and gender target,

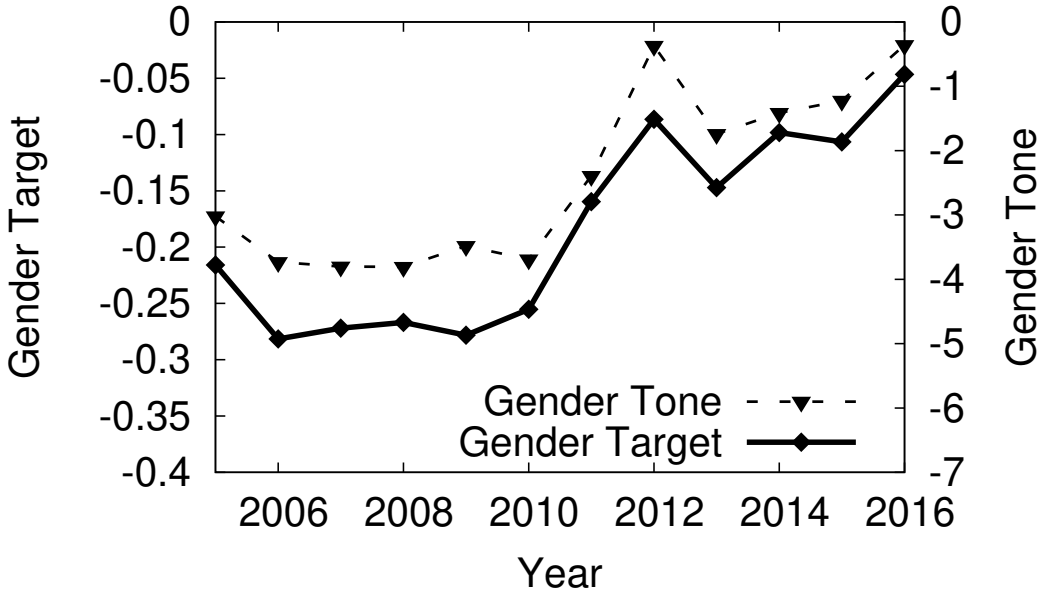


Figure 2.14: Average gender score computed over distribution of seniority level in 2016.

respectively. We find that all the independent variables have statistically significant effects on gender score. The effects are consistent with the previous qualitative analysis, i.e., gender scores vary over groups, and decrease with higher seniority level and more formal employment. However, after ruling out the effect of these factors, we still find an underlying increasing trend that is statistically significant. Although there could be other factors, we believe that awareness of using more inclusive language, and/or using less masculine language, is an important part of the change.

Changes in Word Use. Lastly, we are interested in understanding how different words and phrases vary in their contribution to gender bias over time in job listings. We take the 500+ gender biased words from our dictionary, and plot their frequency of appearance (and therefore impact on gender scores) in Figure 2.18. We find that the frequency distribution of the most popularly used terms is exponential (and therefore it appears linear on a log plot).

Figure 2.19 plots the frequency of usage for top 20 words across the years, where we group the masculine and feminine words separately in the figure. We see that most of the

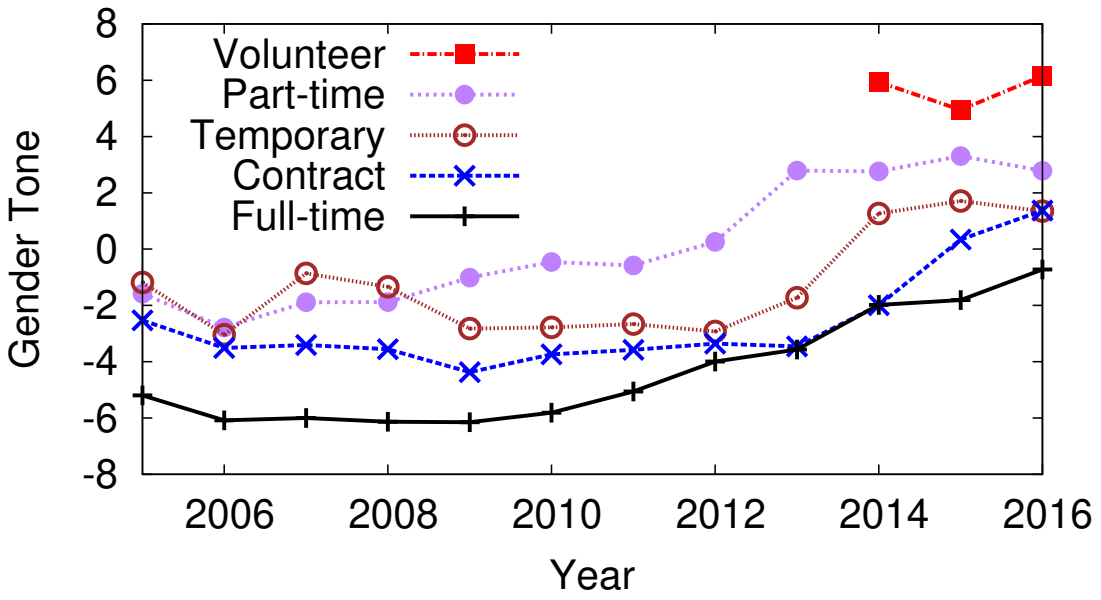


Figure 2.15: Average gender bias score of different employment types (*gender tone*).

masculine words show a stable or slightly decreasing trend[5], while the feminine words display more dramatic changes over time. Specifically, the most used masculine word, "strong", experienced a sharp decrease, potentially due to the growing awareness of biased wording. Among the top feminine words, "care," "patient" and "health" display significant increase, possibly driven by the growth in health-related jobs, while "understand," "develop" and "relationship" show a visible decline over the years. We also repeated our study on words with the largest changes in usage frequency, which produced similar results.

Finally, we compute the frequency of usage for each word across the years, and fit a trend line using linear regression. Of these terms, 145 words show statistically significant trends (p-value < 0.05). 75 of these are masculine toned words, and they are evenly divided between those growing in frequency and those dropping in frequency. Of the remaining 70 feminine toned words, the large majority (84%) showed an increasing trend.

5. The only exception is "driver." This is probably because driver can also be a job title in the transportation sector which experienced a rapid growth since 2014.

Feature	Coefficient
(Intercept)	-1.1380
Year after 2005	0.3064
Entry Level	2.9893
Associate	0.6209
Mid-senior Level	-0.6825
Director	-4.6385
Executive	-3.1562
Volunteer	2.4624
Part-time	-1.8491
Temporary	-2.7402
Contract	-4.4472
Full-time	-5.0084
Transportation	-3.6283
Technology	-1.5921
Media	-1.0557

Feature	Coefficient
Agriculture	-1.1434
Construction	-0.9402
Management	-0.3033
Finance	-0.3763
Art	-0.0245*
Organization	0.2408
Corporation	0.2867
Legal	0.3372
Recreation	0.4573
Service	0.8327
Government	1.2678
Goods	1.6498
Education	2.4565
Health	3.9049

Figure 2.16: Ordinal regression using *gender tone* as dependent variable. $p < 0.001$ applies for all entries except *, which has $p = 0.273$.

2.3.6 User Study: Impact of Gender Bias

So far, we have quantified the level of gender bias in job listings over time, but we do not yet understand how these gender biases actually impact users (potential job applicants). To answer this question, we conducted a user survey, which we describe here. In short, we find that gender scores from our algorithms properly reflect perceived gender stereotypes associated with job postings, but that biased wording has limited effect on the perception of a job, compared to respondents' preconceived notion of the job type. We also find that male respondents are less willing to apply for stereotypically feminine jobs, while the reverse does not hold.

Survey Participants. We recruit survey respondents from two different sources, Amazon Mechanical Turk (MTurk), and undergraduate students. In total, we received results from

Feature	Coefficient	Feature	Coefficient
(Intercept)	-0.0510	Transportation	-0.0458
Year after 2005	0.0208	Corporation	-0.0432
Entry Level	0.1000	Legal	-0.0414
Associate	-0.0193	Construction	-0.0303
Mid-senior Level	-0.0885	Finance	-0.0243
Director	-0.2275	Management	-0.0156
Executive	-0.1589	Government	-0.0041
Volunteer	0.0442	Service	0.0150
Part-time	-0.1497	Organization	0.0203
Temporary	-0.1583	Goods	0.0328
Contract	-0.2670	Recreation	0.0401
Full-time	-0.2848	Art	0.0621
Technology	-0.0731	Health	0.1345
Media	-0.0620	Education	0.1974
Agriculture	-0.0571		

Figure 2.17: Ordinal regression using *gender target* as dependent variable. $p < 0.001$ applies for all entries.

469 distinct MTurk workers and 273 students, each job advertisement is evaluated by at least 20 different workers and 12 different students. Undergraduate students were volunteers who received necessary credit for their course work. Each worker was compensated $1 for finishing the task. To ensure the quality of replies, we require workers to have an 80% HIT approval rate, and have at least 50 HITs approved in the past. We also include a quality check (i.e., gold standard) question in our survey question list, (e.g., "Please answer A and D for this question.") to avoid low-quality/non-responsive workers. For respondents who failed our gold standard questions, their responses are not included in our analysis. In most cases, responses from MTurk workers and the students point to the same conclusion, and we thus combine their answers in such analysis. In cases when the responses differ, we analyze the respondent pools separately.

The demographics of our survey participants are as follows. Among all 469 Mturk work-

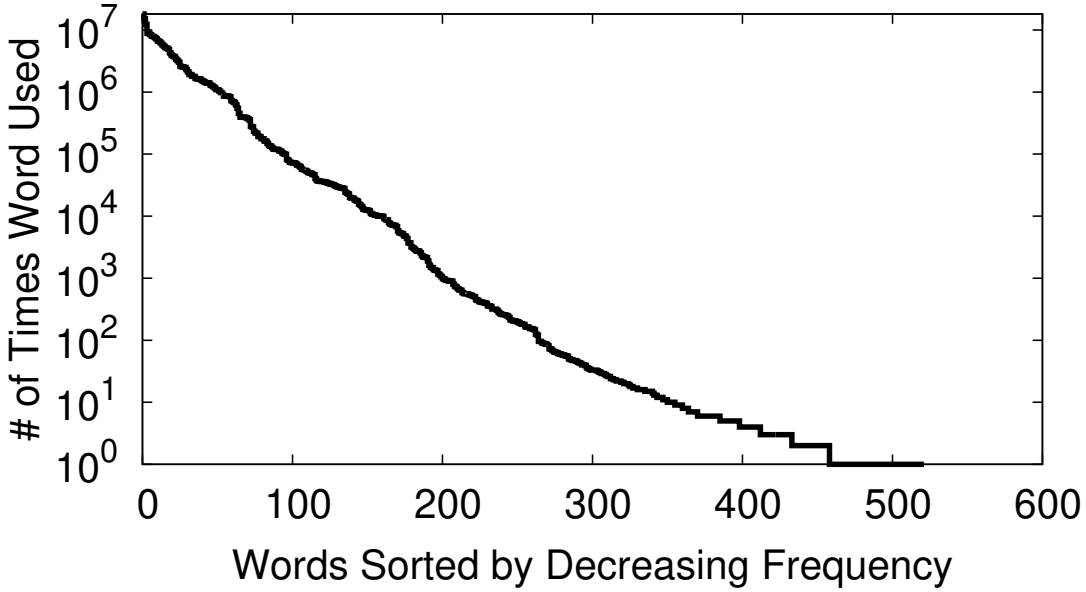


Figure 2.18: Distribution of word frequency.

ers, 54.6% indicated male and 45.4% indicated female. The majority of participant ages fall into the ranges of 21 to 30 (38.8%) or 31 to 40 (34.3%), with 1.49% younger than 21, 11.7% older than 50, and the rest fall between 41 and 50. Most participants work full-time (67.2%); and most hold a Bachelor's (42.9%) or a Master's (24.5%) degree. Among the 273 college students, 201 (73.6%) are female and 72 (26.4%) are male. 209 (76.6%) of the students are younger than 21, and 64 (23.4%) are from 21 to 30.

Methodology. In our user study, we divided job advertisements into 3 categories: *masculine jobs*, *feminine jobs*, and *gender neutral jobs*. Feminine jobs are randomly sampled from job advertisements with the highest 10% gender score, as scored by both gender target and tone. Masculine jobs are similarly sampled from postings with the lowest 10% gender score, and neutral jobs are sampled from advertisements with scores nearest 0. We did not restrict the time-span of our advertisements, since we want to maximize user reaction on potential gendered language. When older job advertisements is included in our sample, we manually check their content to make sure there are no outdated words that will significant

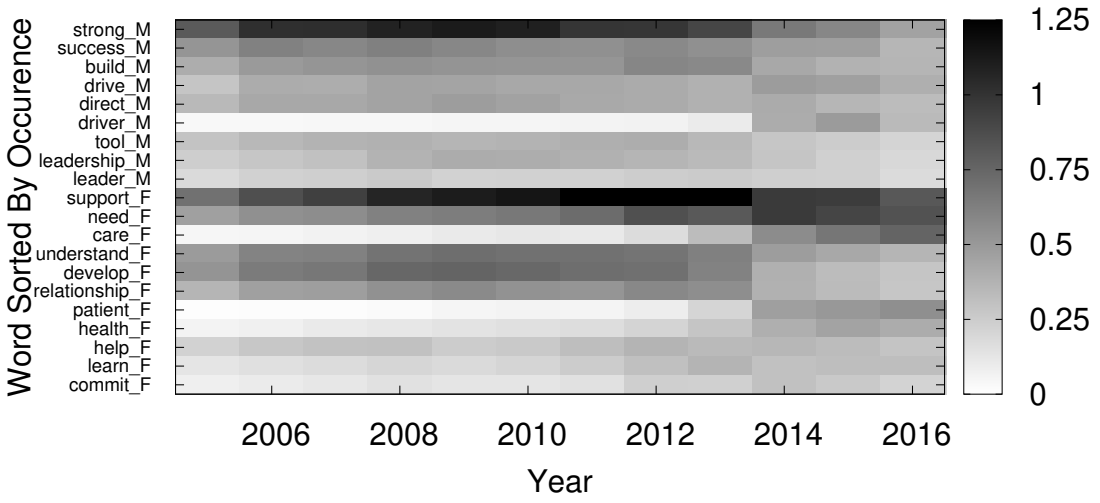


Figure 2.19: Evolution of word frequency for top 20 words.

affect user's reaction.

For each advertisement, we created a second version of the job description by replacing keywords or phrases marked as gendered language from our dataset with more gender-neutral words or phrases not in our dictionary of gendered words. For example, substitutions included "workforce" replaced by "employees," and "collaborating" replaced by "working." Although we made efforts to consistently replace biased words with the same neutral alternative, some instances required more dynamic replacements to retain the readability and intent of the original post. For instance, depending on context, "engage" may be replaced with "participate," "employ," or "work." Other words, like "please," were simply removed. To ensure the biased language was not just replaced with different biased language, we calculated gender scores for the edited descriptions, and verified that the substituted descriptions received gender-neutral scores.

We replaced or removed as much gendered words and phrases as possible without changing the intended meaning of the original job posting. This provides a suitable baseline to isolate the impact of gender wording from people's inherent biases and stereotypes. This

contrasts with prior studies [62] that analyzed masculine and feminine language without comparing against neutral wording as a baseline.

We asked each user to read three job advertisements, one from each category. After reading through the ads, we gathered their responses to the following questions:

- *Q1*: If you were fully qualified to apply for a job like this, how likely is it that you would apply for this particular position? Answers are measured by 5-level Likert Scale (1 indicates definitely would not apply and 5 indicates definitely would apply).
- *Q2*: By looking at the job description, what would you think to be the percentage of women currently working in this type of position?
- *Q3*: While reading the job description, to what extent did you feel that the advertisement would attract more male or more female applicants? Answers are measured by 5-level Likert Scale (1 indicates job attracts mostly males and 5 indicates job attracts mostly females).
- *Q4*: Please mark any words or sentences that you do not feel comfortable with.

Q1-Q3 in the above questions are multiple-choice, and Q4 is open-ended. At the end of the survey, we collected user demographic information. We consulted our local IRB and obtained approval before conducting the user study. Note that while Q3 could be asked differently, i.e., ask users to rate the attractiveness of the job and compare results between male and female respondents, we chose this version so users would focus on the effect of wording rather than allowing other, random or uncontrolled factors to influence their "broad" evaluation of a post.

The three job advertisements were randomly selected from the three categories (masculine, feminine, neutral), with equal likelihood of choosing an edited or raw version for each category. We used a pilot test of Amazon Turk users to determine if question order impacted user response. After controlling for other factors, results showed a decreased likelihood of job application for the second and third ads. Thus we presented the three sampled

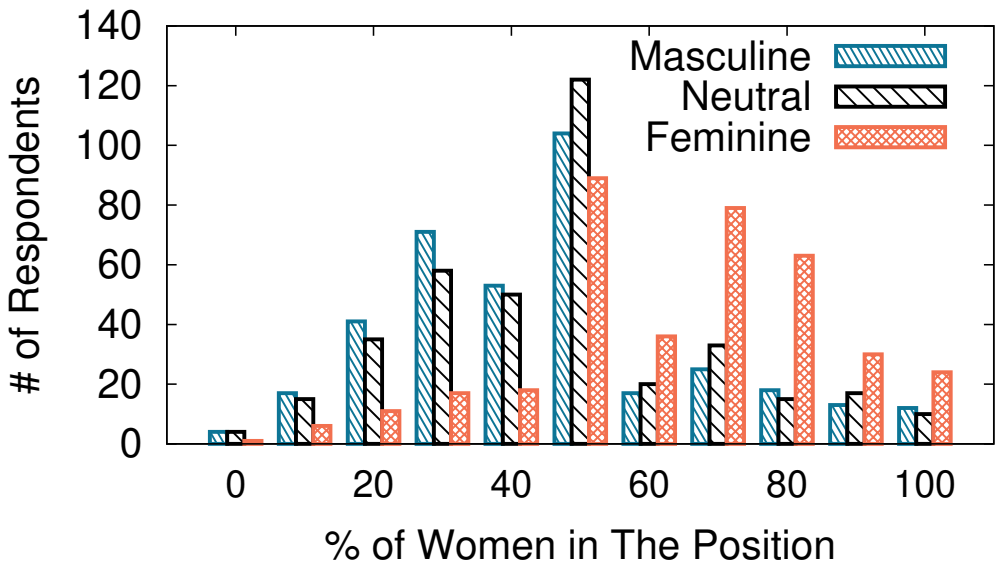


Figure 2.20: Responses on Q2 (percent of female in the position).

advertisements to participants in random order.

Preliminary Task: Assessing whether gender scores accurately reflect user perceptions of gender bias. We first validate our algorithms designs with the user study. Here, we compared user responses to questions Q2 and Q3 for different types of job advertisements before replacing or removing gendered words. The results are shown in Figure 2.20 and Figure 2.21. In each figure, we analyzed responses for each question choice, across all advertised job positions. We find that people presume more female workers in supposed feminine positions, suggesting feminine toned job advertisements appear more attractive to women. Corresponding findings apply to masculine toned job advertisements, as well, thus validating our algorithms.

We also conduct Mann-Whitney U-tests on the distribution of responses between groups, and the difference is statistically significant with p-value less than 0.001 across all types except between masculine and neutral ads. Interestingly, all distributions show an artificial peak at the most neutral answer. Many respondents selected 50% for Q2, where we asked

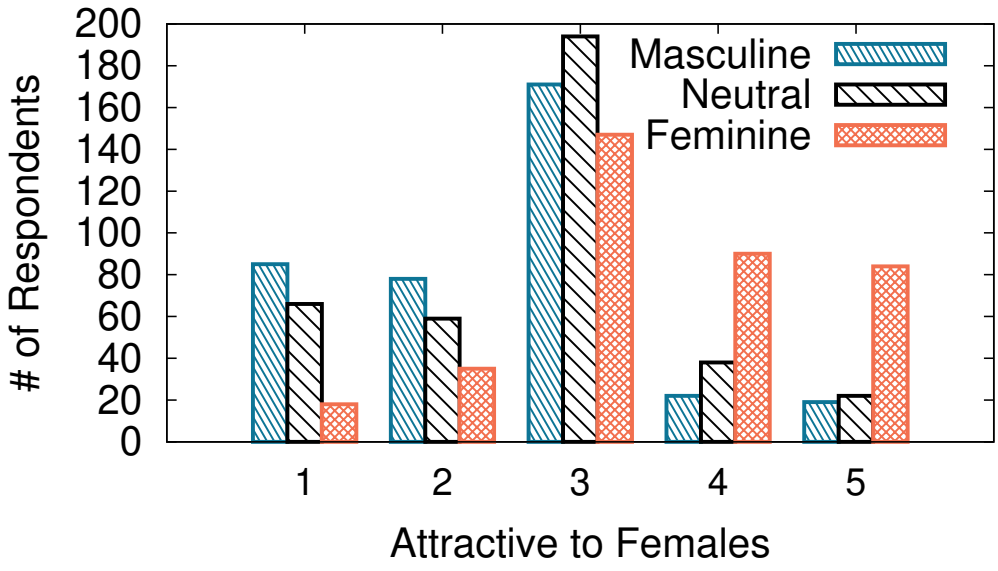


Figure 2.21: Responses on Q3 (attracting female applicants).

about the percentage of women working in the position, and for Q3, many selected the option "attracting male and female applicants equally." Respondents who selected the neutral choice often provided reasons related to "equal chance" or "equal rights," showing a conscious awareness of, or desire for, equal gender participation in the workplace.

Principle Task: Quantifying effects of gendered wording on job application rates. ***Perceived occupational gender bias affects decisions to apply.*** Given the correlation between word choice and perceptions of gender bias, we next sought to examine the extent to which this perceived bias influences one's decision to apply. We show that male respondents express noticeably diminished inclination to apply for jobs perceived to predominately attract females. We quantify the level of perceived gender bias by averaging responses to Q3. In Figure 2.22, we plot the perceived bias of a job against the average tendency of female and male applicants to apply, indicated by average response to Q1. By applying linear regression on both male and female applicants, we discovered that female applicants do not show any preference with respect to gender distribution, with near zero slope

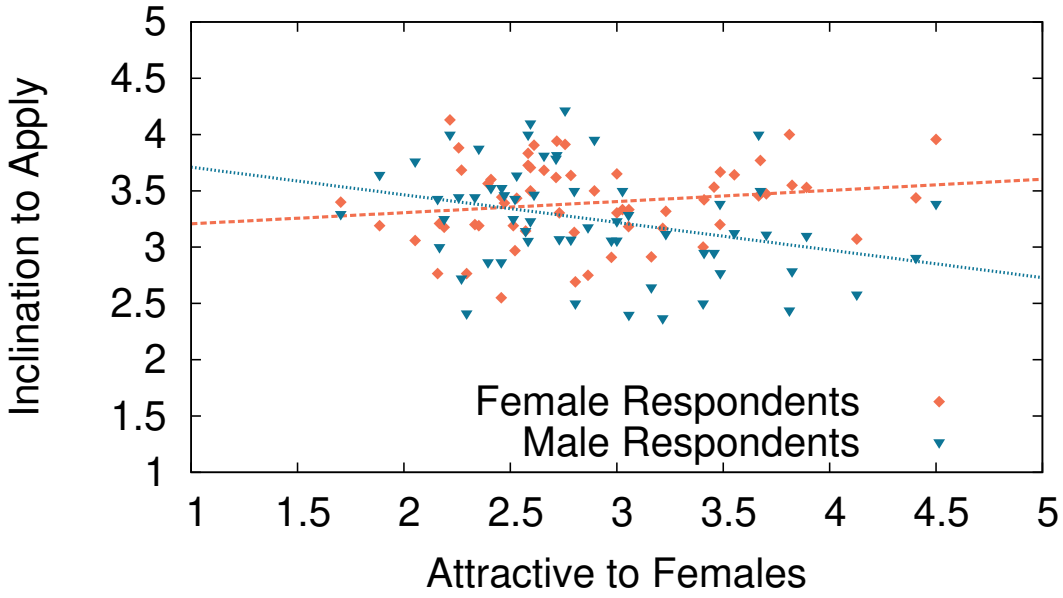


Figure 2.22: Average user response on Q1 against average user response on Q3.

(0.0989) and a p-value of 0.190. Meanwhile, male applicants displayed a preference against applying for female-dominated jobs, with a slope of -0.245 and a p-value of 0.0113. This contradicts prior work [62], where female applicants found masculine worded occupations significantly less appealing. One explanation is that our gender bias is naturally embedded in the job posts, and thus likely to be of a lower intensity than artificial job ads composed specifically to contain gender bias. Additionally, female perception of and reaction to gender bias may have shifted since the 2011 study.

From Figure 2.22, we can observe a high degree of variance, indicating that willingness to apply for a job may be affected by other external factors besides gender neutrality. When we asked their reason for why they will or will not apply for a position, we found a few frequently mentioned reasons, including the anticipated salary, benefits, location, workload, potential of career development, and whether the field of job appeals to the respondent.

Changes in gendered wording have limited effect on predisposition to apply. The ultimate question remains as to whether a recruiter can change the wording in a job

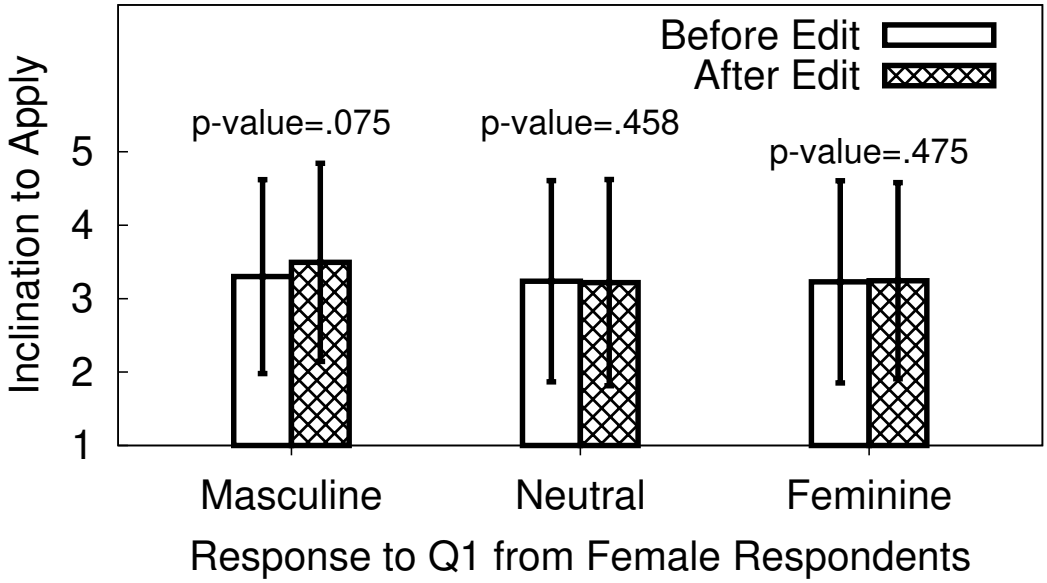


Figure 2.23: Responses from female university students on Q1 (inclination to apply), breakdown by types of jobs. Differences measured with Mann-Whitney U-test.

advertisement and increase the likelihood of potential applicants to apply for the job. Thus, we seek to quantify the causal effect of wording on users' decisions to apply for a job or not. For female and male applicants, we compared the predisposition to apply for a job, measured by averaging answers to Q1, before and after word substitution.

These results are different between the two pools of respondents. For students, wording change in masculine-worded ads does affect application decisions, as shown in Figures **??**. Removing male-biased words from job advertisements leads to less male applicants and marginally more female applicants expressing an inclination to apply. When performing Mann-Whitney U-test on responses of MTurk workers, the p-value are above 0.05 for all three job types with both male and female respondents.

This shows that the effects of word use are observable, but somewhat limited. We then sought to break down the effect, pinning down whether wording actually *causes* a perceived bias, by comparing respondents' reported perception before and after word substitution.

We plotted the average responses to Q3 before and after word substitution. If wording

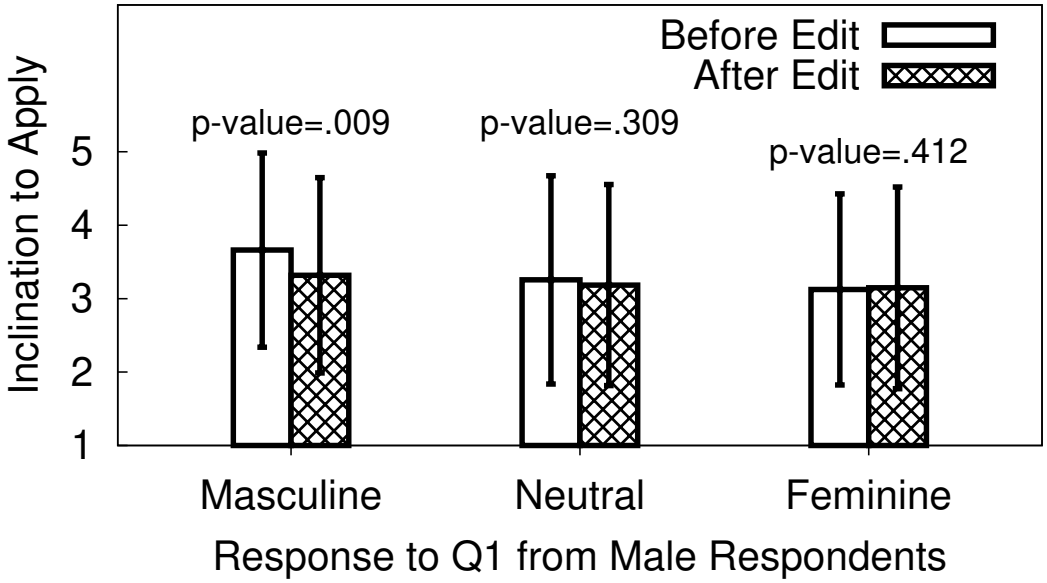


Figure 2.24: Responses from male university students on Q1 (inclination to apply), breakdown by types of jobs. Differences measured with Mann-Whitney U-test.

is the sole cause for gender bias, then by removing the biased words, all jobs advertisements should appear with similar level of perceived gender bias, thus yielding a slope of 0. In contrast, if wording has zero impact on gender perception, it will show a slope value of 1. In Figure 2.25, we can see that the perceived bias persists even after word substitution, with a linear regression yielding a slope of 0.850 and p-value of 0. Similar results are observed for Q2, showing a slope of 0.825 and p-value of 0. This indicates that there certainly exist other properties affecting gender perception more influential than changes in wording.

Preconceived notions of occupations predominately affect user perceptions. To better understand what factors influence user perception, we examined the reasons given in our survey responses. In the survey, we asked users to explain their reasoning and mark any words or phrases in the job advertisements that made them feel uncomfortable. With this exercise, we hoped to gain insights into current perceptions that may be missing from previous studies or even current available services.

We found that many explanations given in our responses include preconceived ideas of

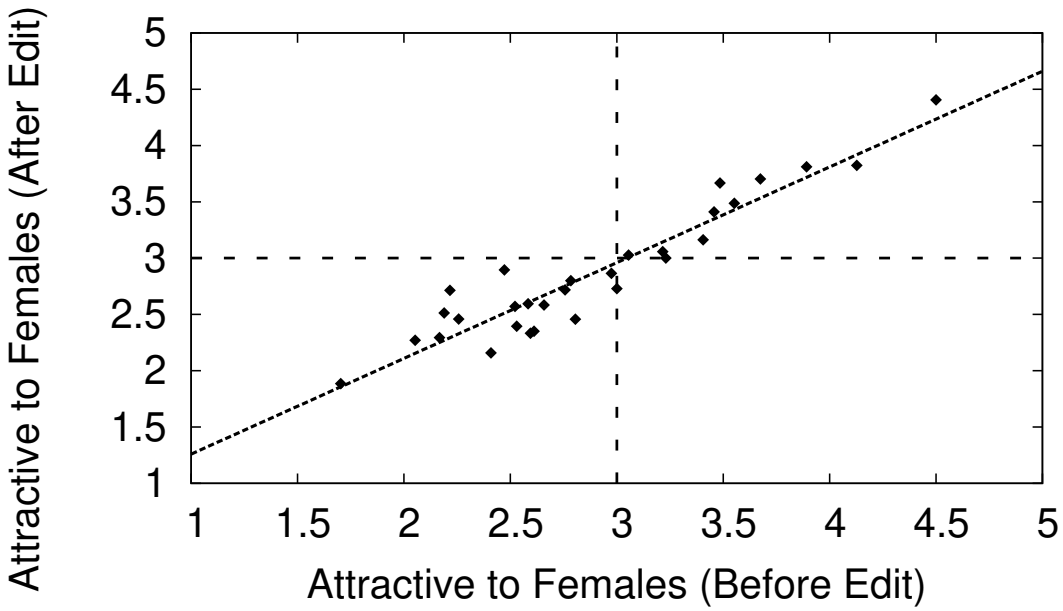


Figure 2.25: Average answer to Q3 (attracting female applicants) before and after word substitution.

the described job function. For example, in response to a job providing technical support for customers in a cable television company, one user believes that 20% of the workers in this position consist of women and therefore presumes the position will attract primarily male applicants, indicating as the reason "It's a technology job." Some respondents even expressed strong gender stereotypes, making statements like, "Low wage jobs tend to hire women, men try to get better jobs." Similar stereotypes also affected users reading posts for jobs perceived to be suitable for female applicants.

Many responses associated a particular gender with specifics characteristics they assumed would best fit the job. For instance, some highlighted phrases such as "bringing accountability, decency, and humor to the job," with explanations stating that these expectations would appeal primarily to male applicants, and women may not like the position. We infer that these users think women are less likely to possess such attributes, making them unfit for the job. These words were not included in our gendered language dataset, indicating modern perceptions of occupational gender bias.

Other responses focused more on preconceived notions of job functions. One response to a position requiring business travel with the company CEO described how they couldn't imagine a man doing this kind of assistant job, demonstrating an inherent stigma against men performing clerical work. Other respondents indicated that non-assistant positions requiring travel or "with little supervision" were better fits for male applicants who may feel more comfortable traveling than women, perhaps due to traditional views that women should or would want to stay at home. Additionally, users suggested that job descriptions requiring an ability to lift up to 50 lbs. or unloading trucks skewed towards male applicants who would be more likely to be capable of such physical activity. On the other hand, many male respondents expressed no interest or consideration for a beauty consultant position because they perceived it as a field of work for females, with some users describing a beauty school degree requirement as simply, "sounds sexist." Most surprisingly, several responses ironically stated that using the phrase "Equal Opportunity Employer" felt insincere and directly singles out females or minorities.

We originally intended to use these questions to better identify specific gendered words or phrases. Surprisingly, we instead gained insights about the role that inherent gender bias plays in the job marketplace.

Limitations. There are potential biases in the survey sampling. This user study recruited 25 users to evaluate each job advertisement, but only studied 30 job advertisements in total. A small number of job advertisements may not represent the largely diverse pool of all the job advertisements. The participant pool was limited to Amazon Mechanical Turk workers and undergraduate college students, neither of which are representative of a highly diverse workforce in the general population. In addition, since the workers do not necessarily evaluate jobs from their own area, some respondents expressed unfamiliarity with terminology (acronyms, corporate jargon) specific to the field of work described. Finally, answers to questions showing gender bias may in fact be reflecting personal familiarity of

respondents with assumed statistics in a given industry.

2.3.7 Discussion

Through our data analysis, we observed an increasing shift away from masculine-biased job postings over the years, and that employers today use less gendered wording than they did 10 years ago. However, the results of our user study also indicate that this trend towards gender neutral wording does not correlate with a perception of gender neutrality in the job market.

Surprisingly, user responses to our survey showed significant gender bias in the participants to specific job positions. Despite the correlation we found between gendered wording and perceived bias, users' explanations show their underlying biases were bigger determinants of their likelihood to apply than any gendered wording. Even after removing all gendered language from the job advertisements, these trends remained in the responses (see Figure 2.25). Gender bias is present in the participant's own perception, independent of the language used in job posts. The implication is that completely removing gendered wording will have limited impact in forming a more gender neutral workforce. This echoes observations made in prior work [71] that inherent user bias was pervasive in the job marketplace. Ultimately, we need to address inherent gender bias in the applicants themselves to significantly improve gender neutrality. While quantifying and understanding these limitations will require further studies, this analysis provides an early empirical perspective on the shifting dynamics of gender bias in the American workplace.

Key Takeaways. The wording used in job advertisements shows us that gendered language remains prevalent today, and does change over time. Despite attempting to reduce biased perceptions by changing the gendered wording, people maintain strong underlying gender associations with certain job positions, irregardless of the wording. Although it's clear that

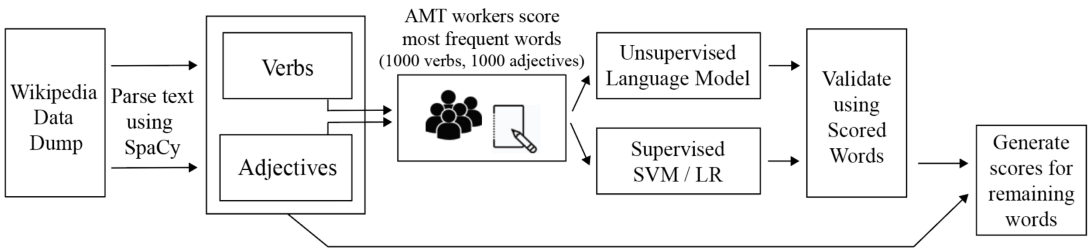
companies desire to reduce gendered language, there remains little effort to improve methods and understanding of real world perceptions of certain words. In addition to preconceived perceptions, this lack of effect may be due to changing the wrong words to start with, or selecting poor words to replace with. To remedy this gap, we see a need to update and expand the gendered word banks and apply modern machine learning methods that can evaluate text beyond single word tokens. This case study results demonstrate how gender scoring methods need to both: include more gender-rated words, and generate gender scores for large bodies of text in other medias (e.g., news articles).

2.4 Updating Methods to Detect Gendered Language

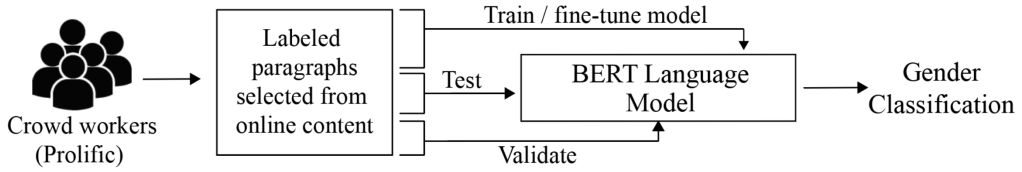
With the knowledge that gendered wording remains present in media today, and has evolved over time, we now seek to update our data and methods to better measure the extent of gendered language used. We first gather a large corpora of text written by and about people to discover how words are commonly used to describe men and women in today's world. Using the gender scoring methodology from the case study, we generate ratings to gather real-world perspectives of how people associate words with male and female stereotypes. Then, to further understand language in context, we crowd-source entire paragraphs or articles that describe people using stereotypical male or female language. This provides insight into current ways people are being described, *and* how those descriptions are perceived by readers. With this, we can then train an end-to-end model to learn from modern language and modern perspectives.

2.4.1 Methodology

To detect gender stereotypes in articles, we implement and compare two approaches: a traditional lexicon-based method that operates on individual words, and an end-to-end method that operates directly on text paragraphs. Both approaches are data-driven and apply ma-



a) Pipeline for building a lexicon-based model



b) Pipeline for building end-to-end supervised model

Figure 2.26: Building gender-stereotype detection models.

chine learning models to scale up our evaluation to arbitrary words/articles.

An overview of the approaches can be seen in Figure 2.26. Specifically, our lexicon-based method starts from breaking down the article to word level tokens, then uses crowdsourced workers to score the perceived masculinity and femininity of a set of most frequently used words. We train supervised models using this data, and apply the result to build the full modern gender stereotype lexicon. The lexicon scores individual words, which when combined, given the overall gender score of an input article. Our end-to-end approach takes crowdsourced text samples illustrating gender stereotypes, and uses them to fine-tune the BERT deep learning language model. The result is a deep learning model capable of gender scoring arbitrary paragraphs and articles.

Crowdsourced Gender Scores. Both the lexicon and end-to-end approaches require datasets that exemplify gender stereotypes at the word or paragraph levels. Currently, no such databases exist that reflect modern perspectives on gender stereotypes in language. Thus, we build our own datasets using crowdsourcing.

Our goal is to create datasets that represent current language use, with minimal bias, and can easily scale up when additional resources become available. We approach this goal in three steps. First, we leverage a large corpora of existing text samples to reflect typical use of language. Second, we use human crowdsourcing to label the data. Finally, these datasets can be iteratively updated and expanded through the methods described, thereby providing practical scalability.

Mitigating Bias. To avoid potential bias, rather than ask respondents to brainstorm original content, we instead focus on gathering data that reflects perceptions of existing written content. To reduce potential biases due to variations across cultures, we limit our respondent pool to US residents. By gathering data points from assorted modern perspectives, we believe this provides representative perspectives of the current US population. For our end-to-end approach, we chose a model not previously trained for any particular language task, which allows us to examine and search for common patterns of language use that may be associated with gender stereotypes.

2.4.2 Detecting Gender Stereotypes: Lexicon-based Approach.

In this section, we introduce our lexicon approach for detecting gender stereotypes in written articles. A lexicon-based approach first analyzes how people associate particular words with common gender stereotypes, and then aggregates these scores to derive a gender score for the entire article.

While gender lexicons have been the preferred approach for detecting bias or stereotypes [17, 147, 164], they have a number of limitations. First, because they are manually constructed and labeled, they are limited in size and coverage. Second, they cannot produce gender scores of arbitrary words, and can lose relevance over time as language describing gender stereotypes evolve. One potential solution is to apply (unsupervised) language mod-

els to automatically estimate gender score of words without human labels. We explore the empirical efficacy of this approach on a dataset of ground-truth labeled data (see below), and report results later in this section.

We also propose a supervised learning solution for computing gender scores of arbitrary words. This solution includes four steps. First, we use existing text corpora (*i.e.* Wikipedia) to identify frequently used, descriptive words as our gender lexicon dataset. Second, we generate our ground truth data by apply human crowdsourcing to label a subset of this dataset. Third, we use our ground truth data to train a supervised learning model that derives *gender scores* (a score reflecting a word's perceived masculinity or femininity) for arbitrary words. We apply this model to label our larger gender lexicon dataset. Finally, we use this gender lexicon dataset to compute the gender score of an article and evaluate how consistent or contradictory the article is with gender stereotypes.

Our Gender Lexicon Dataset. We begin by identifying a large set of words that are are potentially related to gender stereotypes. Here, we restrict our selection to *verbs* and *adjectives*, as stereotypes often manifest in people's behavior and how they are described [55].

We extract candidate words from Wikipedia Datadump[6]. We choose Wikipedia because it is large and diverse, and thus likely to include most of the commonly used English words. We downloaded a snapshot of Wikipedia text on March 4th, 2019, removing all images and links. In total, our dataset includes 5,817,125 documents, 42,653,358 paragraphs and 2,076,621,930 words.

To extract verbs and adjectives that characterize humans, we analyze the word dependencies in each sentence using a parse tree implemented using SpaCy[7]. For verbs, we extract "subject-verb" relationships in each sentence, where the subject is a human-related word like "he," "she," "man," "woman" etc.. For example, in the sentence "He ran away from her,"

6. `https://dumps.wikimedia.org/`

7. `https://spacy.io/`

the subject is "he" and the verb is "ran." So the word "ran" is extracted. We lemmatize all words: we merge variants of a single noun or verb down to its stem, *e.g.* past tense "ran" becomes the lemma "run." For adjectives, we consider two different types: *predicate* and *attribute*. Adjectives are predicates when they are connected to their subject words by a verb, usually "be," e.g., "he is *handsome*." Attribute adjectives are used as modifier before the subject, as in "a *handsome* man." In both cases, we keep the adjective if it is used on a human-related word.

We then filter out words that cannot be found using the Oxford Dictionary API[8]. The removed words are mostly non-English words, non-existent words, or those with the wrong part-of-speech (e.g., a word extracted as an adjective but only used as a noun). In the end, the final lexicon dataset consists of 6,178 verbs and 4,424 adjectives (10602 total).

2.4.3 Ground Truth Gender Lexicon via Crowdsourcing

From our lexicon dataset, we choose the most frequently used 1,000 verbs and 1,000 adjectives, and use a user survey to label them. The resulting labeled words will serve as our ground truth dataset. We manually checked all words to ensure that they are suitable candidates, removing words that depend strongly on context (e.g., "next", "final"). Also, we remove references to race, country, or religion because those biases remain outside the scope of this work.

Survey Design. Like previous studies [17, 164], our survey asks the participants to rate the extent to which they associate each word with a typical man or woman. Specifically, the participants are shown a list of words, and asked to evaluate the statement "I feel that _____ is commonly associated with the characterization of a typical man in US society" or "of a typical woman in US society." The evaluation uses a 7-point Likert Scale, from

8. `https://developer.oxforddictionaries.com/documentation`

"strongly disagree (1)" to "strongly agree (7)." The participant can also select "I don't understand the word." Each participant rates 50 adjectives and 50 verbs "of a typical man" (*male rating*) and another 50 adjectives and 50 verbs "of a typical woman" (*female rating*). We also collect their demographic information at the end. The survey takes on average 15 minutes to complete, and each participant received $3 as compensation.

To ensure that the participants pay attention during the survey, we randomly insert in each survey 4 words that do not exist in English (i.e., gibberish). The participants are expected to select "I don't understand the word" for these words. We include another quality control question when collecting demographics information, which is a multiple choice question asking them to choose both A and D. We removed all the responses that failed these quality check questions.

We recruited our participants on Amazon Mechanical Turk. To reduce potential differences across cultures, we limit our participant pool to US residents over the age of 18. Each participant can answer our survey up to 10 times, but will rate different words each time. We collected a total of 1097 qualified response sessions (HITs), among which 619 are from male participants, 476 are from female participants, and 2 chose not to disclose their gender. Over 99% of the words have more than 50 male ratings and 50 female ratings.

Gender Score Calculation. The ground truth score of a word is measured by the difference between the ratings associating the word with men and the ratings associating the word with women. For example, if a word is perceived as strongly associated with typical men but not associated with typical women, then we evaluate the word as carrying a strong masculine stereotype.

Specifically, we use the T-statistic in a two sampled T-test to measure the difference between masculine ratings and the feminine ratings of a word. The T-statistic reflects the extent to which the average value differs across samples. Like other statistical tests, the T-statistic also maps to a p-value which indicates how likely the average value of the two

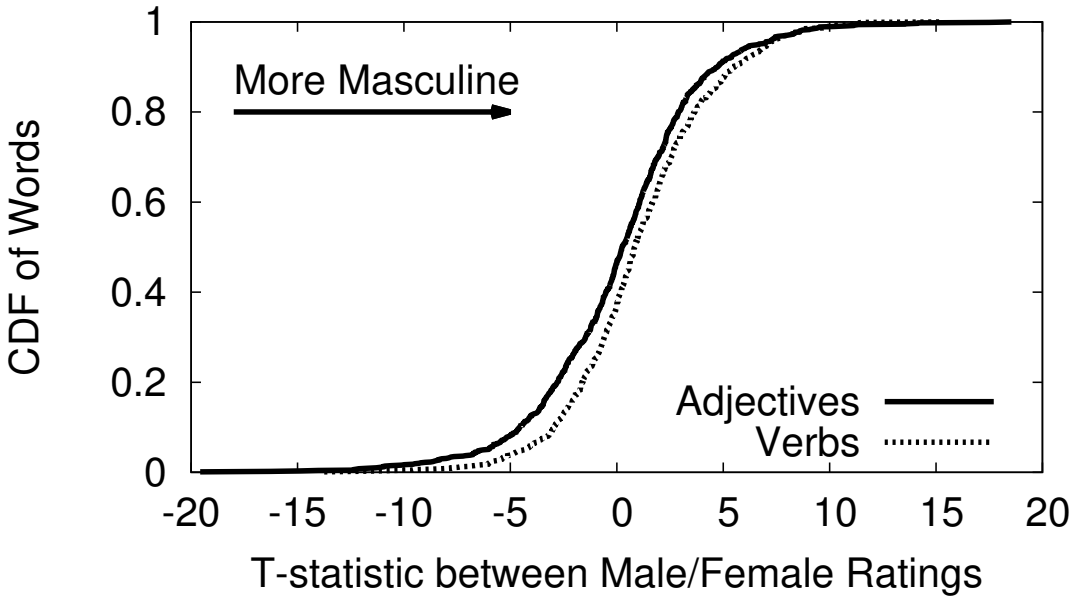


Figure 2.27: CDF of T-statistic between male ratings and female ratings. A higher positive number indicates people more easily associate the word with a typical man than a typical woman.

samples are identical. A small p-value indicates a statistically significant difference between samples [166].

We plot the distribution of T-statistic of each word in Figure 2.27. Most words score around 0, indicating gender neutrality. Figure 2.28 and Figure 2.29 show examples of words with different T-statistics. Except a few words related to appearance (e.g., hairy, beautiful), our highly stereotypical words are consistent with recent work demonstrating that stereotypically men are perceived as strong, active and violent, and women are perceived as weak, emotional and kind [55].

Reliability of User Responses. We perform two tests to examine the reliability of survey responses. The first is *split-half reliability* [118] that measures how likely the data collection is *reproducible*. To do so, we randomly split all our participants into two equal-sized halves, and calculate two sets of T-statistics for each word independently using the two halves. We calculate the Pearson Correlation between the T-statistics of the two sets, resulting in 0.85

Figure 2.28: Wordcloud of adjectives. Red denotes feminine words and green denotes masculine words. Larger font size indicates stronger gender associations (larger T-statistic magnitude).

for adjectives and 0.82 for verbs. This result means that repeating the data collection process is unlikely to significantly change results.

In our second test, we determine to what extent responses from male and female participants agree with each other. We split our responses by the gender of the participant, calculate two sets of T-statistics for each word independently using the two splits. From the two sets, we observe a correlation of 0.82 for adjectives and 0.80 for verbs, similar to the correlations in the split-half reliability. This means no significant difference exists between responses of male and female participants. Thus, in the following analysis, we aggregate all responses (both genders) together for our calculations.

Figure 2.29: Wordcloud of verbs. Red denotes feminine words and green denotes masculine words. Larger font size indicates stronger gender associations (larger T-statistic magnitude).

Labeling Gender Lexicon via Word Embedding. Using our ground truth dataset, we first examine whether existing (unsupervised) language models (e.g., word embedding) can be used to automatically label gender lexicon without human input. For this, we consider four metrics used by prior work to calculate gender information of words.

- *Odds ratio.* It calculates how likely a verb or an adjective is used to characterize a man rather than a woman. If a word is more likely to be used on a man, it may indicate masculinity of the word. Specifically, given a word, odds ratio is calculated as:

$$\frac{\#\ this\ word\ on\ man\ /\ \#\ this\ word\ on\ woman}{\#\ other\ words\ on\ man\ /\ \#\ other\ words\ on\ woman} \tag{2.1}$$

Here, "#" denotes "number of times." Odds ratio reflects gender stereotypes in large language corpus [55]. We calculate the odds ratio using the Wikipedia data.

- *Distance to gender specific words.* It has been shown that word embeddings contain gender biases due to stereotypes in the language [31]. Such biases can be captured by calculating word distance to gender specific words. Specifically, given a word, we calculate its average distance to a set of male specific words (e.g., "he", "man") and its average distance to a set of female specific words (e.g., "she", "woman"), then calculate the difference between the two distances. We test 3 commonly used word embeddings: word2vec [114], GloVe [133], and FastText [22]. Here we do not train our own word embeddings, but apply widely used pre-trained models for each of the three embedding methods: *word2vec* from Google News [9], *GloVe* from the 6 billion token Wikipedia dataset [10], and *FastText* from English Wikipedia [11].
- *Projection on gender direction.* One way to reduce gender bias in word embeddings is to extract a gender direction and remove the vector projection on the direction [23]. Here, the gender direction is the direction parallel to $\overrightarrow{she} - \overrightarrow{he}$ or $\overrightarrow{woman} - \overrightarrow{man}$. For each word, we take its projection on the gender direction as its gender stereotype score.
- *Values on gender dimensions.* Another way to reduce gender bias in embeddings is to encode gender information in a reserved dimension during training [198]. Here, we use the magnitude of the gender dimension as a way to quantify the gender stereotype associated with a word. We use the pre-trained word embeddings provided by [198].

To evaluate these methods, we use them to calculate the gender score for each word in our ground truth data, and compute the Pearson Correlation between the calculated gender score and the ground truth. As shown in Table 2.1, while the scores derived by all these methods are positively correlated with human defined gender stereotypes, the magnitude of

9. https://code.google.com/archive/p/word2vec/

10. https://nlp.stanford.edu/projects/glove/

11. https://fasttext.cc/docs/en/pretrained-vectors.html

Word Embedding Method	Adjectives	Verbs
Odds ratio	0.09	0.29
Distance + word2vec	0.44	0.37
Distance + GloVe	0.47	0.41
Distance + FastText	0.47	0.41
Gender direction	0.40	0.33
Gender dimension	0.20	0.08

Table 2.1: Pearson Correlation of gender scores between predictions from word embedding methods and ground truth.

the correlation is no larger than 0.47.

Labeling Gender Lexicon via Supervised Learning. Our results show that automated lexicon labeling via word embedding produces gender scores with mediocre results. Next, we propose to apply supervised learning to train gender score prediction models using our ground-truth dataset.

Since our labeled training dataset only contains around 2000 words, we cannot use deep neural network models that require large training datasets. Instead we use two classical machine learning models: Support Vector Machine (SVM) and Linear Regression (LR). Our models use word embeddings of each word as features, and the pre-trained word2vec, GloVe and FastText as model inputs.

We train our model using a random subset of 80% of the words from our ground truth dataset, then use the model to predict the score for the remaining 20%. We calculate the Pearson Correlation between the model predicted score and the ground truth. We repeat the experiment 100 times, and compute the average correlation value. Results in Table 2.2 are higher than those produced by word embedding (Table 2.1). Since LR with word2vec produces the highest correlation value, we will use this configuration as our strongest supervised learning approach.

Supervised Learning Method	Adjectives	Verbs
LR + word2vec	0.63	0.57
SVM + word2vec	0.62	0.57
LR + GloVe	0.53	0.52
SVM + GloVe	0.58	0.55
LR + FastText	0.52	0.45
SVM + FastText	0.58	0.53

Table 2.2: Pearson Correlation of scores calculated by supervised learning methods and ground truth.

Domain	Number	Domain	Number	Domain	Number
wikipedia.org	385	npr.org	58	huffpost.com	46
nytimes.com	178	forbes.com	57	washingtonpost.com	45
theguardian.com	78	dailymail.co.uk	54	biography.com	39
cnn.com	78	foxnews.com	48	cnbc.com	37
people.com	63	time.com	47	vogue.com	37

Table 2.3: Top domains and number of articles from each domain.

Computing Gender Score of Articles. Finally, we leverage our gender lexicon to detect gender stereotypical language in articles. Similar to the methodology used to analyze job advertisements, we assign a gender score to an article based on word usage, first extracting all verbs and adjectives in the article, then adding the scores of these words together to get an article's gender score. If the total score is positive, the paragraph is labeled as consistent with masculine stereotypes or contradictory to feminine stereotypes, otherwise it is labeled as consistent with feminine stereotypes or contradictory to masculine stereotypes.

We evaluate the performance of this approach against an end-to-end deep learning model in Section 4.1.2.

2.4.4 Detecting Gender Stereotypes: An End-to-End Approach

We now introduce our end-to-end approach. Different from the lexicon approach, the end-to-end approach operates directly on paragraphs without breaking them down to individual words. Here we take a *supervised* learning approach: first gathers human perceptions of

	Consistent	Contradict
Masculine	championship, ceo, gun, league, player, businessman, top, service, mountain, fight, basketball, win, drive	gay, makeup, gender, singer, fashion, comfortable, mom, youtube, cosmetic, dress, feel, wear, caregiver, beauty, sexuality
Feminine	cook, child, home, beautiful, beauty, care, clean, fighter, daughter, makeup, family, mother, dress, kid, mom	field, champion, history, sport, athlete, fight, martial, force, training, team, technology, institute, lesbian, rank, tech

Table 2.4: Top keywords that distinguish consistent and contradicting stereotypes.

gender stereotypes at the paragraph level to build a moderately sized training dataset, then uses it to train a deep learning classification model based on the BERT language representation tool [47]. The resulting classification model can detect gender stereotypes on arbitrary paragraphs and articles.

We take two important steps to avoid potential bias in our trained model. *First*, we collect our training (and testing) data by crowdsourcing articles with different representation of gender stereotypes. These are existing articles aiming at describing a man (or woman) while the actual human perception can be either consistent with or contradictory to the original intent. Using this ground truth dataset, we formulate our gender stereotype detection problem as *two binary classification problems*: determining whether the description of a man is consistent with masculine stereotypes, and whether the description of a woman is consistent with feminine stereotypes.

Second, our model is built on BERT [47], a unsupervised language representation tool that converts text articles into vectors. Since BERT does not target any particular language task, we can use it to examine and search for common patterns of language use that may be associated with gender stereotypes. Specifically, we use our training dataset to fine-tune BERT by adding one additional output layer to implement the above mentioned binary

classification tasks.

Collecting Labeled Articles via Crowdsourcing. To collect the articles for our tasks, we perform a survey study. Our survey ask users to search the Internet, and copy & paste articles (or a few paragraphs of an article) that meet the following requirements: it describes a man (or woman), and the description is consistent with (or contradictory to) common gender stereotypes. We ask participants to briefly state the reason for choosing each article. Each participant is asked to provide 4 articles, with 4 different combinations of requirements (man or woman, consistent or contradict). The entire survey takes about 25 minutes.

We recruit survey participants from two different sources, Prolific[12], and undergraduate students from our university. Prolific is a crowdsourcing service aiming at providing high quality data that empowers research. The Prolific participants are compensated $3, and the students are compensated with 0.5 research course credits.

In total, we received results from 980 distinct Prolific workers and 110 students. Again, we limit participation to US residents to reduce potential bias due to cultural differences. Among the 980 Prolific workers, 508 (51.8%) are male participants, 457 (46.6%) are female participants, and 15 (1.5%) chose not to disclose their gender. Among the 110 college students, 52 (47.3%) indicated male and 58 (52.7%) female.

We received 4360 articles (4 per participant), and filtered out 27 articles that do not contain any pronouns, named entities or gender specific words, indicating that these articles are not likely to be descriptions of people. When looking at the sources of the articles, most of the articles are from biography pages (e.g., Wikipedia), or news sites (e.g., New York Times). Table 2.3 lists the most frequently used domains.

To understand the content of these articles, we extract top keywords in each category using Chi-square statistics [192], which measures how strongly a word can be used to distinguish articles in different categories, i.e., consistent or contradictory. We calculate Chi-

12. https://prolific.ac/

square statistics for masculine stereotypes and feminine stereotypes separately, and list the top keywords in Table 2.4. We see that our survey participants commonly choose sports and business related terms for men and domestic related terms for women as exemplifying gender stereotypes. Further, some similarities appear between men who contradict stereotypes and women who are consistent with stereotypes (and vice versa).

Building the Classification Model. Our classification model will run two tasks: determining whether the description of a man is consistent with masculine stereotypes, and whether the description of a woman is consistent with feminine stereotypes. Thus we use the articles describing men for the first task and those describing women for the second task. We randomly split up the data into chunks of 8:1:1 for training:validation:testing. We use our training data to fine-tune the BERT model, and use the validation set to identify the optimal hyper-parameters, which is 2e-5 learning rate for 3 epochs. In the following section we use the test data to examine the model performance and compare it to the lexicon approach.

2.4.5 Empirical Evaluation

We now evaluate and compare the lexicon approach and the end-to-end approach using the above mentioned test data. We apply each approach on the test data to predict whether each test article is consistent with or contradictory to its intended gender stereotypes. We then compare these results to the ground truth provided by humans.

Overall, our study shows that the end-to-end approach largely outperforms the lexicon approach, in terms of detection accuracy and robustness. A closer look at these results also offers insights into some fundamental problems facing the lexicon approach. We further confirm that the end-to-end approach does not require a large training dataset to perform well. Finally, we test the end-to-end approach on a practical task of detecting gender bias in job advertisements, which outperforms the industry state-of-the-art.

	Accuracy (M)	AUC (M)	Accuracy (F)	AUC (F)
Lexicon (Full Set)	0.67	0.70	0.68	0.71
Lexicon (Ground truth Set)	0.58	0.61	0.62	0.64
End-to-End	0.77	0.85	0.80	0.87

Table 2.5: Accuracy / AUC of lexicon and end-to-end approaches among articles describing male (M) and female (F).

	PAQ	BSRI	Gaucher et al.
% words overlap	0.22	0.38	0.48
% overlap and matching labels	N/A	0.83	0.85

Table 2.6: Comparison of lexicon coverage against prior work. PAQ does not provide gender labels, thus no direct comparison.

Validating Testing Dataset via User Survey. To ensure that our evaluation (using the testing dataset) is sound, we performed another user study to understand whether the per-user contributed labels in the test dataset can accurately capture public perception of gender stereotypes. Specifically, each survey participant is given 10 articles randomly selected from our testing dataset, and is asked to score on an 7-point Likert Scale, where 1 indicates "strongly contradictory" and 7 indicates "strongly consistent".

We ran the study on Prolific, and each user was compensated $1.10. In total, we received results from 203 distinct Prolific workers, of which 108 (53.2%) are male participants, 92 (45.3%) are female participants, and 3 (1.5%) participants chose not to disclose their gender.

Each article in the testing dataset received at least 4 ratings, from which we computed the average rating and compared it to the actual label of the article. Overall, the new multi-user rating is reasonably consistent with the original rating, indicating that our testing dataset offers a consistent, public view of gender stereotypes.

Comparing Lexicon and End-to-End Approaches. We evaluate how accurately the lexicon and end-to-end approaches can predict gender stereotypes in written articles, by computing prediction accuracy and Area Under the Curve (AUC). The results in Table 2.5 show that the end-to-end approach is much more accurate.

Reason	Lexicon Wrong	Also E-to-E Wrong	Example
Lexicon Coverage	8	0	The first woman I invited to co-author a **publication** was in 2015, four years after completing my **PhD**.
Phrase	10	0	... who **paints his fingernails**, **braids his hair** and **poses for gay magazines**...
Non-human	6	0	Katie Bouman has already worked on looking around corners by analyzing **tiny** shadows ...
Consistent and contradictory	27	4	Even as I regularly work out and **lift weights**, I am a rather **fragile** excuse for a woman, constantly getting sick...
Multiple people	10	3	**My wife** had more earning potential and so I volunteered to concentrate on family and home.
Subtle stereotype, insufficient information	50	123	*American actor Peter Dinklage is labeled as contradicting masculine stereotypes because he is a dwarf, which is not discussed.*
Data noise	30	18	*Random response or failure to meet task requirement.*

Table 2.7: Reasons for lexicon approach making wrong classification. The "Lexicon Wrong" column is the number of cases when the lexicon approach makes a wrong prediction, and the "and E-to-E Wrong" column is the number of cases the end-to-end approach is also wrong among these cases. Bold words are words that are closely related to the reasons provided by the survey participants. Italic words are not exact content from our data, but summarize participant explanations.

In this table we also show the results of the lexicon approach when using the ground-truth lexicon (labeled by our user survey) and the full set lexicon (expanded via supervised learning). We see that the use of full set lexicon effectively improves the detection accuracy, but still cannot match that of the end-to-end approach. Although the two approaches are trained on different data, both datasets are curated from commonly used language in current bodies of text, then evaluated by multiple crowdworkers to generate ground truth labels. As such, we believe these comparisons between the two approaches are fair.

Understanding the Lexicon Approach. To understand why the lexicon approach generates less satisfactory prediction results, we manually examine *all* the incorrect predictions

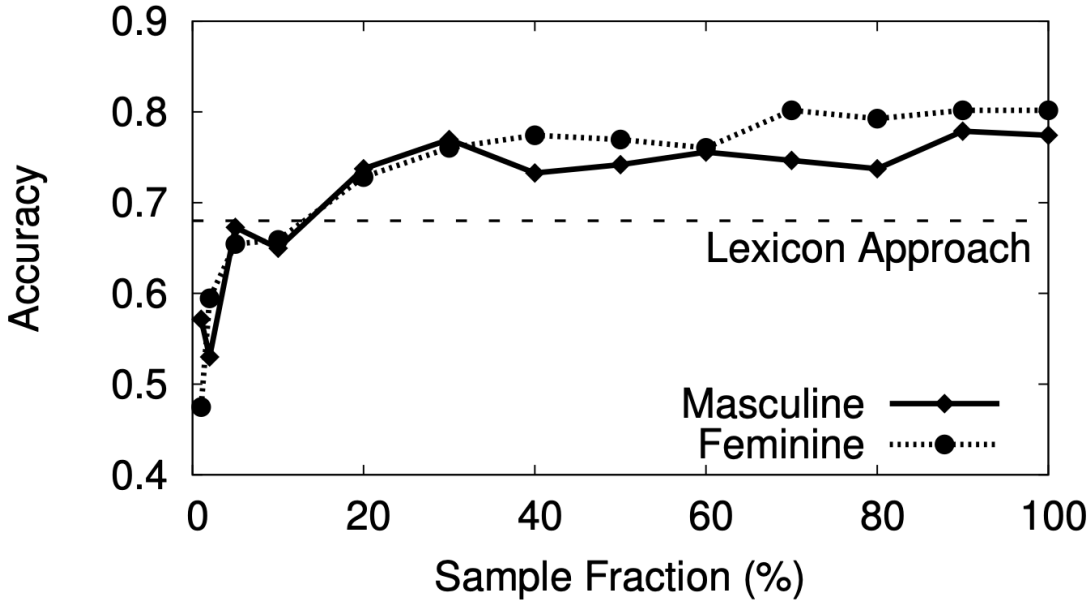


Figure 2.30: End-to-end approach performance with different training data size, compared to lexicon approach (similar number for masculine task and feminine task).

	Textio	Unitive	BERT fine-tune
% of females	0.59	0.54	0.77
Attractiveness to female applicants	0.64	0.54	0.80

Table 2.8: Pearson correlation between user responses and gender bias scores.

the lexicon approach makes in the test set. We summarize the possible reasons behind the misclassifications along with examples in Table 2.7. For each reason, we also calculate how many times the lexicon approach makes incorrect predictions ("Lexicon Wrong" column) and how many times the end-to-end approach makes incorrect predictions among these cases ("Also E-to-E Wrong" column). The detailed explanations are as follows:

- *Lexicon Coverage:* Our lexicon only covers adjectives and verbs, and gender stereotypes can be expressed by words outside of our lexicon. For example, "PhD" could be a word associated with masculine stereotypes.
- *Phrase:* The stereotype is expressed by a multiple-word description, which can not be captured by single words in the lexicon.

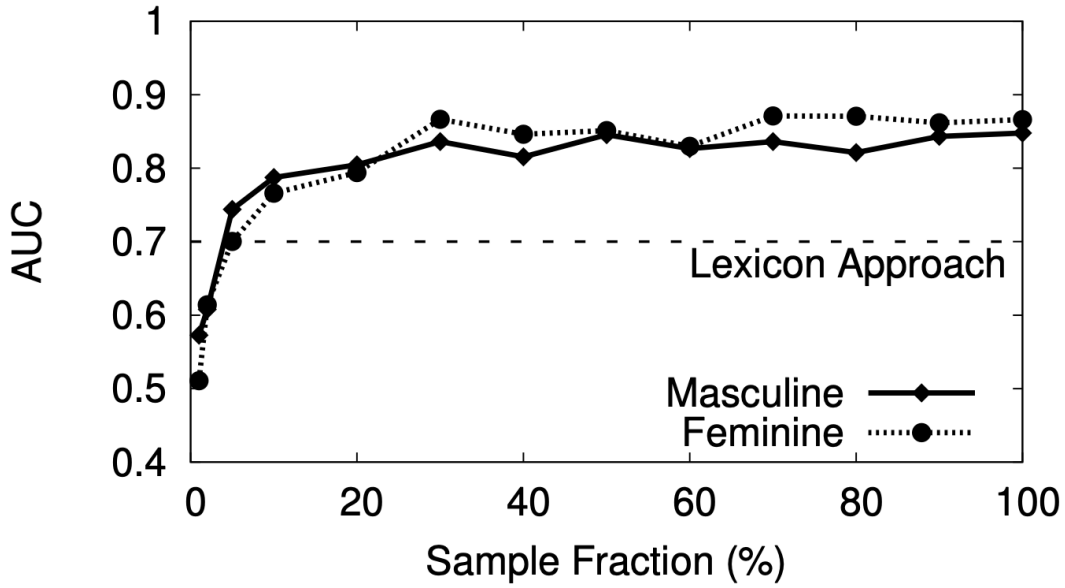


Figure 2.31: End-to-end approach performance with different training data size, compared to lexicon approach (similar number for masculine task and feminine task).

- *Non-human:* The word that indicates strong gender stereotypes is used on a non-human object. For example, in "tiny shadows", "tiny" is labeled as a feminine word but "tiny shadows" does not indicate femininity.
- *Consistent and contradictory:* The article contains a description of a person who has some characteristics that are consistent with stereotypes and some other characteristics that are contradictory. Although the article may focus on one more than other, the lexicon approach can not identify the general focus by word count.
- *Multiple people* The article describes more than one person, usually one person as the main character while the others are supporting characters. The lexicon approach can not isolate the descriptions of the correct person.
- *Subtle stereotyping, insufficient information:* Some users may have different understandings of stereotypes, or insinuate gender stereotypes not explicitly written in the text. For example, an article about American actor Peter Dinklage is labeled as contradicting masculine stereotypes because he is a dwarf, but the fact is not found in the

article.

- *Data noise:* These are the low quality response including cases when the users provide responses that do not fit our task requirement (e.g., the paragraph is not a description of a person, article is consistent with stereotypes when we ask for contradiction).

We also compare our lexicon to those from previous works (PAQ, BSRI, Gaucher), and the overlap with previous lexicons is less than half (see Table 2.6). We found that many of the terms are not often found in current language, and the sparsity of their occurrence in our data makes any comparison of results marginally meaningful. For instance, we found that words such as "aggressive," "assertive," "dominant" or "forceful" are labeled as masculine items in the BSRI, but are not commonly used enough to be included in our lexicon. While PAQ [164] and BSRI [17] include a mix of words and short phrases, our lexicon only considers single words. Phrases such as "acts as a leader," "defends own beliefs," "makes decisions easily" or "willing to take risks" do not directly translate to single words so a direct comparison to such items was not possible. The lexicon generated by Gaucher et al. [62] includes several truncated words, which we found to be an oversimplification, and resulted in some conflicting labels (e.g., for the base "respon*": "response" scored as feminine, but "respond" scored as masculine). Also, since we evaluate adjectives and verbs separately, some words score as feminine in one tense but masculine in the other (e.g., "yield" scores masculine as a verb, but feminine as an adjective). Of those words that overlap, most of our labels are consistent with BSRI and Gaucher et al. , with several exceptions (e.g., previous works labeled "loyal" and "communal" as feminine, and "confident" and "individualistic" as masculine, but our scores are opposite).

Being data-driven, our work is able to evaluate most commonly used verbs and adjectives in *current* bodies of text. The differences in word coverage between our lexicon and previous works may indicate that many of the words from previous lexicons are not commonly used to describe people in modern society. Although some of the terms appear to exemplify

strong gender connotations (e.g., "feminine," "masculine"), such words do not often appear in descriptive language and therefore are not necessary to include in the lexicon. Moreover, our method demonstrates how contextual information, including part-of-speech, plays a significant role in people's perceptions of language, an aspect unaccounted for in previous works.

Impact of Training Data on End-to-End Approach. End-to-end learning approaches often require a large amount of training data. Here, we seek to quantify how much training data is needed to outperform the lexicon approach. We vary the size of the training data, train the BERT fine-tuning model, and evaluate the performance on the same test set. The results are shown in Figure 2.30 and Figure 2.31. We have two observations. First, the performance plateaus when the training data size reaches 40% of our full training data, beyond which further increase in training data size yields no performance gain. Second, the end-to-end approach can outperform the lexicon approach even when the training data is only 10% of current size, which is about 150 articles. This indicates that the end-to-end approach does not need a large corpus to learn typical gender stereotypes.

Application: Gender Bias in Job Postings. Finally, we apply our end-to-end approach to detect gender biased language in job postings, and compare the performance to state-of-the-art tools from industry. In the survey, participants were asked to read 30 job postings and answer questions about the extent to which the job posting is biased by gender. Here, we use their answers on 2 questions to quantify the gender bias in the job postings: 1) the percentage of women they presume are currently in the position (0-100%) and 2) how likely the job posting attracts more female applications or male applicants (7-point Likert Scale, from 7 indicating the post attracts mostly male applicants to 1 indicating the post attracts mostly female applicants).

To convert our male and female stereotype detection models into a single gender bias

indicator, we take the job posting text and use it as input in both the masculine stereotype classifier and feminine stereotype classifier. We take the probability output from both classifiers, and calculate the difference in the two probabilities. For example, if a job posting is predicted as 90% likely to be consistent with masculine stereotypes and 60% likely to be consistent with feminine stereotype, the score of the job posting is 0.3, which indicates a masculine bias.

We calculate the score for all the job advertisements, along with two state-of-the-art services cited by prior work: *Textio* and *Unitive* (now renamed Talent Sonar), both of which are specifically designed to detect gender bias in job posts using a lexicon-based approach [170]. Table 2.8 shows the correlation between the scores and user responses. This shows that although the models in our end-to-end approach are not specifically trained for job advertisements, they still outperform the best lexicon approaches designed for this task.

2.4.6 Discussion

This work seeks to reconcile the traditional lexicon-based approaches for detecting gender stereotypes in language, with modern natural language processing tools almost entirely based on end-to-end deep learning models. The high level question is: what approach should researchers and practitioners take moving forward, an updated version of lexicon-based models (developed in this work), or an end-to-end deep learning model built on existing language models (BERT) and further trained with paragraph-length text samples? This work finds that despite our best efforts to update and strengthen the lexicon-based models, end-to-end models based on BERT provide substantially stronger results, even when trained on our moderately-sized, crowdsourced dataset. In fact, when applied to the context of gender bias in job listings, this end-to-end model significantly outperforms models used by industry services.

Limitations. This work has several limitations. First, the task of gender stereotype detection was simplified as binary classifications for masculinity or femininity. A correct and more inclusive model would include non-binary and trans labels. Also, we only collect a moderate dataset with a few thousand training samples, which is small relative to some popular datasets in the NLP community that contain hundreds of thousands of samples. It is unclear whether the performance can be significantly improved if the training data was orders of magnitude larger. Further, although attempting to mitigate effects from cultural biases by limiting the participant pool to US workers, some biases may remain. Lastly, the end-to-end approach comes from fine-tuning the current state-of-the-art BERT model. It is possible that a more task-specific model can generate a better performance.

Key Takeaways. By leveraging both traditional methodologies and modern deep learning models, we now have a better understanding of how gendered language expressions in modern media. We find that while the expanded lexicon shows significant improvement over the traditional gendered word lists, the end-to-end model allows for better classification of larger bodies of text. Although the end-to-end model shows significant promise, it still misses out on capturing some of the more nuanced concepts in certain examples we tested. As gendered language remains inherently subjective, and language remains extremely complex and evolving over time, there likely could never exist a model that could provide perfect classifications. Nonetheless, this does show the importance of incorporating new data and perspectives over time, to continue striving for better models for identifying gendered language. Overall, simply measuring gendered language does not inherently reduce bias. It may, however, encourage people to be more aware of the wording they use when describing people, and the impact it can have on the subject and readers of the content. By gaining a better understanding of how language may reflect and perpetuate biased perspectives, perhaps next we can enact changes that prevent the biased perceptions.

CHAPTER 3

PERCEPTION AND AUTOMATED SPEECH TRANSFORMATION OF VOICE BIASES IN HUMAN SPEECH

Beyond word choice, our tone of voice can evoke and perpetuate bias as well. Recent developments in generative models for speech allow us to manipulate our tone of voice, content, and accent. However, little research has been done to understand the effects of altering tone of voice, and how it changes perception in the listener. This work seeks to prevent biased perceptions from occurring by changing interactions in real time. By leveraging existing emotion speech labeled datasets and open-source deep learning models, I examine how changing emotional tone of voice can prove beneficial for the speaker. I find that reducing emotional tone improves preference for people in situations when competence and trust are desired. Additionally, I find that although most people can see some benefit to improving a conversational experience with someone by altering their tone of voice, they would never want to do so themselves, and find it dishonest for others to do so.

3.1 Introduction

The way humans express themselves through language includes not only the words spoken, but also *how* they say it. Human speech also carries with it factors such as tone, emotion, and intention. Listeners, on the other hand, may perceive some or all of these factors, but also properties about the speaker, including external characteristics (e.g., race/ethnicity [1], gender appearance [135], socioeconomic status, regionality [172], education) and personal characteristics (e.g., emotion, confidence, trustworthiness [156]). For example, a recent study found that listeners often formed inferences of a speaker's income, education and occupation status after hearing only a few seconds of speech [87].

Correct or not, these implicit speech biases can have significant impact in real life situations, such as job interviews or first impressions in social situations. Specifically for job interviews, some biases may lead to positive outcomes for the speaker (e.g., exhibit confidence) [121], while others may be detrimental (e.g., exude anxiety) [137]. These perceptions about the speaker are likely influenced by existing biases or preconceived notions about the listener [172, 9]. All of this may happen without the listener even being aware of the bias. For instance, someone listening to a speaker with an angry tone can subtly experience increasing tension often without being aware [13]. However, if we can change change someone's tone of voice, we may be able to reduce the chance of making these assumptions in the first place.

Recent advances in speech synthesis shows it is now conceivable to use machine-learning based tools that modify human speech to remove or reduce the factors that contribute to bias in the listener. For example, recent work has shown promising results in transforming speech to reduce emotional intensity [139], or even generate speech from scratch that matches a target speaker's voice with minimal emotional tones [188]. However, these manipulation tools discussed do not include any motivation for their development or discussion on putting these tools to use in the real world.

The limited amount of existing work addressing advantages of altering vocal tones [187] neglects to address how artificial manipulation may not be desirable for use by real people. Recent research [69] shows that although people generally accept the existence of vocal deepfakes, people worry about the truthfulness and potential for deception when manipulating voices in certain ways. Aside from the technological barriers of altering one's voice in a convincing and compelling way, existing ethics research lacks in determining whether this would be acceptable or desirable in society. In this work, we study the *feasibility* and *desirability* of reducing bias by using ML-based tools to modify emotions in human speech.

The goal of this work is to identify changes that would reduce biases and increase positive perceptions of the speaker. To solve this, we identify critical situations (e.g., job interviews)

when speakers could be most affected by biased perceptions. Since our interactions are often dominated by speech, we focus on biases related to the way we speak. More specifically, we evaluate how our interactions could be influenced by modifying human speech, e.g., by changing emotional tone in our voices. We explore the efficacy and impact of techniques to modify emotional tones in human speech, and how they affect perception of the speaker.

Our work focuses on the following research questions:

- Can existing ML models successfully transform emotional tone in human speech?
- Can human listeners consistently detect changes in emotional tone?
- Can changes in emotional tone translate to changes in perceptions about the speaker?
- Do people want the ability to change their emotional tone in the real world?

To answer these questions, we train an emotion conversion ML model, and conduct two user studies. The ML-based emotion conversion model is trained on a labeled emotion dataset, with the task of converting emotion tones towards a more neutral tone. The first user study (n=360) evaluates the feasibility of listeners detecting changes in emotion tone, and if the conversion alters perceptions of the speaker under a variety of scenarios. The second user study (n=100) evaluates the level of interest people have in manipulating their own emotional tone in real life, and their acceptability of others to do so.

We make several key findings:

- We find that ML-based emotion conversion can effectively convert emotionally toned speech to a noticeably more "neutral" tone, but with noticeable degradation in quality.
- Participants do not consistently interpret the same emotion as one another, but do consistently detect emotion tone changes after conversion.

- For key scenarios, participants show strong preference for more neutral emotion tones, suggesting the voice conversion tool does make the voice more desirable in those scenarios.

- We find that participants feel manipulating tone of voice to be generally deceptive and fear that it could be misused. Most participants only showed some acceptance for certain scenarios when such a tool could improve effective communication and overall experience for speakers and audiences.

Emotional tones can significantly impact perception of speakers in our daily interactions. As machine learning tools demonstrate increasing ability to alter how we perceive the world around us, we hope our work provides new insights into how ML can act to reduce bias from emotion tones in voice. Our discoveries show how our voices can be subtly altered to achieve better human communication by reducing some undesirable tones of voice. Additionally, we investigate how people perceive the real world implications for these tools, and find they have significant concerns regarding how they may be potentially misused for deception.

3.2 Background

Perceptions and Speech. The way people speak can seemingly reveal a lot of information about the speaker. Our brains attempt to make quick decisions about people based on simple heuristics [91], ranging from internal characteristics (e.g., competence, boredom [155], sarcasm [36], trustworthiness [156, 66]) to external characteristics (e.g., gender appearance [135], age [194], regionality [172], ethnicity [172], socioeconomic level [88], attractiveness [24]). Some of these inferences create a positive impression of the person (a halo effect), which could benefit both parties. Once these beliefs form in the mind, positive or negative, they difficult to change, regardless of the content being spoken [167].

While individuals maintain many unique characteristics in the way they speak, soci-

olinguistic research identifies common patterns that listeners subconsciously hear. Through analyzing differences in pitch, frequency, and pronunciation of particular phonemes (sounds), patterns emerge that reveal commonalities among particular groups of people or expressions of certain characteristics. Further, listeners make these determinations from *thin slices* of speech (<1sec - 5min) [172, 134], suggesting the perceptions do not at all depend on the content of the speech.

Expectations of how a person (or group of people) should sound can bias a listeners' perceptions of the person speaking. For instance, lower pitched voices are associated with higher dominance [8, 82] and better leadership ability [86]. Typically, females and younger speakers are perceived as more trustworthy [156], and, at the same time, if a woman's pitch is outside an "optimal" frequency range, they may be perceived as less attractive [24]. Collectively, this research shows that the way in which we speak, in a way, speaks for us.

Furthermore, if we can change the way we speak, we may be able to influence those around us. For instance, speaking positively may instill a sense of optimism [94] or trust [156] in those listening. While some vocal mannerisms persist for long periods of time, such as a hometown accent, some are temporary and more malleable, like emotion. For some people, changing the way one speaks may come easily, but others may be unaware of their emotional tone or unable to easily change it on their own.

Emotion in Speech. As of a couple of decades ago, computational researchers started heavily searching for commonalities across populations that would indicate similarities in emotion expression [154, 52]. However, these studies still relied upon traditional methods of measuring emotion recognition, which usually consisted of exhaustive field experiments. Thanks to recent advances in machine learning, researchers can now train models to learn emotion classifications. These tools include Speech Emotion Recognition (SER) models [189, 70, 38, 92] that are now widely used in the computational linguistics community. Such models, though, rely on very large datasets of voice data accompanied by emotion labels.

Many emotion speech datasets exist that have proven incredibly useful for researching emotion and ways to manipulate it [29, 101, 32, 193, 202, 152]. Due to the costly effort to procure such large audio datasets, they vary widely in the source, audio quality, content, and how the emotions are labeled.

For some, the source of speech clips (e.g., YouTube videos) may result in poor audio quality, which poses a problem for models to interpret the voices correctly. Alternately, some datasets use voice actors, which could compromise the integrity of genuine expression of emotion [84]. For most datasets, the emotion labeling either relies on voice actors' portrayals of the emotions or on the accuracy of crowdsource annotations. Most often, though, the labeling process used for any given dataset remains unknown. The VESUS repository [152] attempted to address some of these issues by generating their own dataset and procuring emotion ratings from both the speakers themselves and crowd sourced annotators via Amazon Mechanical Turk. Unfortunately, despite investigating these particular issues, their raw annotator data is not available, and thus unable to be confirmed. The Interactive Emotion Dyadic Motion Capture (IEMOCAP) [29] has been used widely in the past for evaluating speech emotion recognition models [189, 38, 92], and emotion conversion models [144, 120, 202, 129] , and does include its annotator data with the speech dataset. Additionally, the Emotion Speech Dataset (ESD) [202] is a recently released dataset with high quality emotion labeled audio clips, though it's unknown how they annotated the data.

While emotion labeling tasks are often performed in a controlled setting, real world interpretations of emotions of those around us can affect one's own emotions. Studies have found that manipulating emotion expressions of a speaker changes the emotions of listeners, even without the listener becoming aware of the influence [13]. For instance, when participants listened to a subtly altered version of their own voice, they concluded that calmer voices reduce anxiety, and lower pitches increase feelings of power [41]. More broadly, Lerner et al. [94] found that regardless of content or one's current situation, positive emotions correlate

with optimistic judgements, and negative with pessimistic. Knowledge of how we interpret and react to emotions of others has since been proposed to optimize interactions with voice assistants [105, 33], and encourage students to feel enthusiastic about learning [98]. However, outside of digital devices, there currently lacks sufficient research regarding the practicality of systematically altering voices in the real world, and the ethical implications.

Manipulation of Speech. Existing work has explored different ways to manipulate voice, including real-time vocal perturbation via haptics, digital manipulation tools (real-time or post-processing), and deep learning voice conversion models.

Physical voice manipulation methods, such as haptics, proposes reducing voice-related biases without artificially altering the speaker's voice identity in real-time [5]. Similarly, BarryWhaptics [187] uses a haptic device to reduce pitch bias, thereby increasing perceptions of dominance, while maintaining speaker identity and naturalness of human speech. However, this approach requires pre-installed and individually configured physical actuators, and only focuses solely on increasing perceived dominance.

Digital audio analysis and manipulation tools allow for fine-grained adjustments of existing speech through software. Praat [21] is a tool developed for high-quality speech analysis and conversion. With ground truth labels (e.g., regionality, gender, socioeconomic status), this tool allows researchers to identify commonalities among groups, or alter the raw audio features. However, this manual process can be time consuming as it requires individual manipulation for each speech sample. More recently, a tool called DAVID [139] allows for real-time speech transformations for different emotions (i.e., happy, sad, afraid) that can automatically transform vocal emotions at varying levels of intensity. Aucouturier et al. [13] found that the subtle manipulations of emotional tone correlate with changes in the emotion of the listener.

Additionally, deep learning models applied within the audio domain allow for generation and transformations of speech, music, and other sounds [106]. Such models can take raw

audio as input and transform accents [173, 54, 130], speaker identity [138] and, more recently, emotions [50, 201, 203, 199, 200]. Generative models, such as CycleGAN [16, 199, 99] and StarGAN [144, 120] can be trained on existing speech datasets and learn to generate high quality transformed speech. Additionally, EmoCat [158] shows robustness for language-agnostic emotion transfer, but like the AutoVC model it's based on [138], requires a carefully configured bottleneck. Alternately, researchers have applied text-to-speech (TTS) models trained to mimic a target speaker and then apply emotion conversion, such as a recent model from Zhou et al. [200]. Zhou et al. also configured various other models to optimize for non-parallel datasets [199] and transform speech of unseen emotions [202] or unseen speakers [203], among others. The metrics for evaluating each of these models varies greatly, and often focuses on detecting any change in perception of a voice transforming from a neutral tone to an angry, happy, or sad emotional tone. While these models serve as powerful tools for speech conversion, there remains a lack of real world examples of the models put to use.

3.3 Methodology

This work seeks to evaluate how changing the tone of one's voice can affect perceptions of the speaker in a positive way. First, we must find out if we can effectively change the tone of someone's voice with existing voice conversion tools. Then, we must evaluate how these changes in voice correlate with changes in perception. Lastly, we evaluate how people feel about the possibility of changing their voice in real life.

Therefore, to analyze biased perceptions, we first evaluate how perceptions of emotions relate to perceptions of other personal characteristics (e.g., trust, competence, anxiety), and then identify situations where perceptions of these characteristics could be critical.

To determine how perceptions change after altering one's voice, we start by looking at how voice can be analyzed and manipulated. First, we identify existing voice datasets and conversion models. While it would be ideal to examine many aspects of voice that may

affect perceptions of speakers, such as regional accents or age, the availability of such labeled datasets remains extremely limited.

The purpose of this study is to evaluate how changing tone of voice and change perceptions. Instead of using the emotion conversion model to convert "neutral" speech *to* an emotion tone, we attempt to *reduce* the emotion tone. Then, we can measure the difference of a stronger emotion tone compared a more muted tone, and the perceptions of highly emotional speech compared to toned down speech.

First, we evaluate how consistently people perceive the same emotion as the label from the dataset meta-data. We also evaluate if the speech produced by the emotion conversion model is perceived as the intended emotion tone. Then, we attempt to reduce the emotion tone with an emotion voice conversion model, and measure the changes in user perceptions of the altered speech. Lastly, we evaluate perceptions of the tool itself, such as which situations it may be acceptable or desirable to alter tone of voice in these ways.

3.3.1 Data Collection and Model Training

For our study, we use the recently released Emotional Speech Dataset (ESD) [202]. ESD contains audio data from English and Mandarin native speakers (10 for each language, 20 total). We only use the English speaker data, which has 4 female and 6 male speakers. Each speaker records 350 parallel utterances in 5 different emotions (neutral, happy, sad, angry, surprise) for 17500 total English utterances. Each utterance is a sentence (or part of), on average 2.76 seconds. We chose to use ESD over other datasets because it contains high quality audio and consistent data points for evaluation across all speakers. ESD provides clean parallel data, which serves two purposes. First, the utterances are repeated by all speakers, for each of the 5 emotion tones, which reduces the chances that our results will be affected by the content of the speech. Second, high quality parallel data increases the likelihood of achieving higher quality results from the emotion conversion model, which

Content of Utterances
Now quicker the fiddle went.
On the 22nd of last March.
The nine the eggs I keep.
This turn goes to the hill.
Let's make the noise a snake.

Table 3.1: Table shows the content of the audio clips selected. Due to the dataset containing parallel data, both speakers spoke the same content in each of the 5 emotion tones.

reduces the chances of the audio sounding as though it's been digitally altered.

We also compare the IEMOCAP annotation data against our own crowdsourced annotations of the ESD dataset in §3.4.2 to better understand consistency of interpretations of emotion tones. IEMOCAP was selected for comparison because it contains additional data detailing the emotion labels from the original annotators. Upon examination, approximately 25% of its audio clips remain without a marked label because the annotators didn't agree upon a single emotion expression. This suggests a significant amount of variation in how people perceive emotion in speech.

Next, we examined available voice conversion models. Dozens of models continue to be developed and improved each year, and we selected a model [203] that can be trained with unparallel data, and works on unseen speakers. With its ability to be used on unseen speakers, this model allows for the future potential to create a pre-trained pipeline that users could later implement themselves for pre-recorded talks (further discussion in §3.6). We reference the Emotion Voice Conversion model that we implement in this work as EmoVC, and the original authors of the model have made it publicly available on GitHub[1].

To train the model using the ESD dataset, we randomly selected 1 male and 1 female from the English speaking voices. Prior research [156] shows that male and female speakers

1. https://github.com/KunZhou9646/Speaker-independent-emotional-voice-conversion-based-on-conditional-VAW-GAN-and-CWT

vary in pitch and frequency, even when expressing the same emotion. Ideally, we would have included more speakers, but were limited by the cost considerations of conducting user studies, and so we prioritized gathering a sufficient number of responses for each utterance over increasing the number of speakers. We selected 5 utterances (see Table 3.1), for each of the 5 emotions, for each of the 2 speakers, for a total of 50 utterances to be transformed and evaluated. For each utterance (besides neutral), we performed 2 transformations. First, we converted the utterance from its original emotion tone to "neutral" (e.g., angry $\rightarrow$ neutral). Existing research primarily evaluates how well these emotion conversion models translate neutral speech to an emotion tone, rather than an emotion tone to neutral. Converting from an emotion tone to neutral serves 2 purposes: (1) we evaluate how well these tools effectively alter the perceived emotion, and (2) whether the conversion towards the neutral tone results in changes of the perception of the speaker. Second, we also generated utterances that converted to the same emotion as the original utterance (e.g., angry $\rightarrow$ angry). The purpose of the second transformation is to establish a fair baseline for comparison and determine if the transformation towards neutral is significant enough to be noticeable. While the tools produce results of decent quality, there remains significant noticeable differences compared to that of unaltered human speech, which would result in significant confounding effects.

Mitigating Bias. We took steps to mitigate bias during our data selection process. First, although the English speakers in the ESD all speak with North American accents, we found that some speakers had a noticeable regional accent (e.g., Southern accent). As previous research has shown [172], accents can bias the perceptions of a listener, and may affect perceptions of emotion tone. After listening to a few dozen utterances from each speaker we randomly selected, we felt confident the speakers that did not have any noticeable regional accent. Second, we then sampled utterances at random until we found 5 that fit the criteria of: similar length (5-6 words), no offensive or suggestive language (e.g., violence). The length of 5-6 words was selected so that each utterance can be considered a thin slice, such that

they are long enough to perceive emotion tone, but not long enough to potentially provide significant context. Additionally, the participants were instructed before listening to each utterance to ignore the content and only consider the tone of voice. The content of the utterances are in Table 3.1).

User Study Evaluations. For our evaluation, we conduct two user studies. (IRB approved)

1. The first user study is intended to evaluate the *feasibility* of using these tools to alter perceived emotion tone. We measure the quality and perceived emotion tone of each utterance, before and after conversion. Additionally, we compare the consistency of our emotion label results for the non-synthesized ESD data to that of IEMOCAP. Then, to determine how participants associate emotions with certain characteristics (e.g., warmth, anxiety), we compare the utterances in pairs of the original emotion and it's corresponding conversion, and how these preferences may change after emotion conversion. Lastly, we measure preferences for the different emotion tones of voice in various scenarios (e.g., talking on the phone to a call center operator).

   Findings. Despite a degradation in quality following the voice conversion, the EmoVC tool effectively reduces emotional tone of voice, resulting in positive changes of perception and preferences for several scenarios surveyed.

2. The second user study is intended to determine the *desirability* and *acceptability* of manipulating voice in the real world. Desirability measures how much a potential user would *want to use such a tool* to manipulate their own speech, and in what way (i.e., reduce anger, increase happiness, or no change) and under what circumstances. Acceptability measures perceptions of whether such a tool *should be used* by anyone

at all in certain circumstances, particularly with regard to potential deception of the speaker and/or listener(s).

Findings. Participants overwhelmingly disfavor speech manipulation due to the potential for deception, and with few exceptions they much prefer authentic interactions with people.

3.4 Feasibility of Emotion Conversion

Survey Design. We conducted an online user study consisting of multiple choice and free response questions. The goal of the study is to measure the ability of EmoVC to transform emotion tone, and whether it changes the perceptions of the participants. We do this by measuring consistency of emotion recognition across users before and after emotion conversion. Then, we evaluate whether perceptions of a speaker change after emotion conversion.

Participants. We recruited participants from the online crowd sourcing platform Prolific[2]. The survey was designed to take 15 minutes on average, and participants were compensated $3. We only recruited participants ages 18+ from the US, with at least 100 previous studies completed, and 95% approval rate. We collected 360 valid responses (participants who did not complete the full survey or failed the attention check question were removed from analysis). Our participant pool included 232 females, 117 males, 11 selected other or chose not to disclose. The participants spanned multiple age groups: 18-19 (4%), 20-29 (40%), 30-39 (30%), 40-49 (12%), 50-59 (9%), and 60+ (5%) years old.

Task. Participants saw one of two forms of the survey. For both surveys, the questions were the same, but the utterances were either all converted via EmoVC or all original (non-

2. https://prolific.co/

synthesized) speech. Participants were asked a series of multiple choice and free response questions in 3 parts:

1. First, they were presented individual audio clips and asked to rate the quality and indicate their perceived emotion(s) of the speaker. We used the same scale for quality measurement as [144] (see appendix). The emotion labels presented were the same as those used for the IEMOCAP annotations, including the additional option of writing in an unlisted emotion for "other". They were asked to select all emotions they perceived. For each emotion selected, they were then asked the extent to which they perceived that emotion, measured on a 4-point scale ["neutral", "somewhat", "probably", "definitely"].

2. Next, we presented the participants with pairs of audio clips. Each pair uses the same speaker and same utterance, but with different emotions. For the non-synthesized ESD data, one audio clip of the pair is labeled as one of the 4 labeled emotions (happy, sad, angry, surprise) and the other is labeled as "neutral". Both are spoken by the same speaker and contain the same spoken content. For the EmoVC data, one audio clip is one of the non-neutral emotions, and the other is that emotion filtered towards neutral. For each pair, the participants are asked to select (along a 5-point scale, with "neutral" in the middle) which clip they associate more with a series of characteristics (i.e., trustworthy, competent, warm, anxious, polite, positive, negative).

3. Lastly, participants were presented again with pairs of audio clips. Instead of characteristics, we described one of five scenarios and asked which voice they would prefer to speak with for each scenario (along a 5-point scale), as shown in Table 3.10. The scenarios were selected based on previous research regarding changing perceptions in specific situations (e.g., talking to someone at a call center).

For each single or pair of audio clip(s) evaluated by the participants, the speaker and

	Male Speaker	**Female Speaker**
Angry	2.94	3.01
Angry → Neutral	3.12	3.28
Happy	3.25	3.89
Happy → Neutral	3.66	4.04
Neutral	2.57	2.35
Sad	2.63	2.63
Sad → Neutral	2.82	2.82
Surprise	2.97	3.43
Surprise → Neutral	3.29	3.72

Table 3.2: Quality ratings for EmoVC synthesized converted audio.

content remained the same, only change was the emotion tone.

In line with previous works [203, 202, 144, 120], each utterance (or pair of utterances) received at least 15 ratings from user study participants, for each question.

Key Findings

- Synthesizing speech via ML models results in a non-significant amount of degradation in audio quality, but varies significantly for different emotion conversions
- Participants do not consistently agree on a single emotion label
- Happy tones are associated with "positive" characteristics (i.e., warmth, trustworthiness, positive, politeness), while surprise shows no associations with *any* positive characteristics.
- Neutral tones are more closely associated with *competency* than any emotional tones, and also preferable to emotional tones in key scenarios

3.4.1 Audio Quality

The participants evaluated the quality for all audio clips (the original ESD non-synthesized dataset and EmoVC generated data) on a scale of 1-5 (1 indicates very good quality with

no distortion, 5 indicates very poor quality with noticeable distortion). The quality scaling used is the same used to evaluate an existing emotion voice conversion model [144] (see appendix for full text). While there's no information given about the precise recording conditions for these audio clips, we can determine they were likely recorded by voice actors for this task (rather than procured from secondary sources, such as online videos). For the non-synthesized ESD dataset, most clips were rated as a 1 ("very good - imperceptible distortion") or 2 ("good quality - perceptible but not annoying distortion"), with an average of 1.85 ($\pm$1.04). For the EmoVC clips, most received a rating of 3 ("decent - perceptible and slightly annoying distortion"), with an average of 3.14 ($\pm$1.07). As shown in Table 3.2 the EmoVC audio quality varied significantly for each clip, each speaker, and each emotion, with the the model performing best on neutral and sad data, while worst on happy conversions and surprise conversions.

People often exhibit different biases during interactions with virtual agents than they do with humans, so we believe it's key for the voice conversion tools to maintain natural human tones. The concept of the "uncanny valley" [119] addresses interactions with synthetic speech, and how it can alter perceptions of the speaker, particularly with regards to likability and trust. Previous research has shown that humans can be fooled by synthetic speech in real world situations [188], and our results suggest that the converted speech quality is comparable to that of the non-synthesized speech. As we expect this voice conversion tool would be used for situations like tele-communication (e.g., Zoom meetings, recorded presentations), it's likely the degradation would not have a significant effect on the listeners' perceptions of the voice. As such, we believe that the change in quality from the emotion voice conversion does not affect the perspectives of the human subjects.

Key Takeaways. While the synthesized audio received mostly "decent" ratings, the lower quality as compared to the non-synthesized data may have affect perception of the speech. Given the overall variation in quality for different types of conversions, though, improvements

Dataset Emotion Label	Participant Emotion Label (M / F)				
	Angry	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	**53.2% / 65%**	14.3% / 5.2%	20.1% / 19.3%	5.2% / 11.1%	10% / 4.9%
Happy	14.3% / 0%	**36.7% / 67.2%**	20.1% / 20.5%	4.5% / 0%	20.8% / 14.1%
Neutral	12.5% / 22.4%	8.3% / 8.7%	**66% / 62.6%**	14.9% / 8.5%	0% / 0%
Sad	13.7% / 0%	6% 5%	39.1% / 53.2%	**37.5% / 33.2%**	0% / 0%
Surprised	18.7% / 9.7%	4.7% / 12.9%	8% / 5.6%	4% / 4%	**53.9% / 65.7%**

Table 3.3: Participant responses for emotion labels of the ESD non-synthesized audio clips, separated by male / female speakers. The numbers indicate the likelihood of participant labels matching the dataset labels.

Emotion Label	Participant Emotion Label (M / F)				
	Angry	**Happy**	**Neutral**	**Sad**	**Surprise**
Angry	**35.7% / 38.1%**	10.2% / 4.3%	6.8% / 7.7%	3.7% / 6.7%	11.4% / 10.8%
Angry → neutral	**21.2% / 30.3%**	10% / 4.3%	**37.1% / 21.7%**	10.3% / 5.8%	5.1% / 5.9%
Happy	9.7% / 9.9%	**19.4% / 32.6%**	5.5% / 17%	2.8% / 0%	22.5% / 10.4%
Happy → neutral	3.4% / 8%	**12.7% / 10.4%**	**45.8% / 44.3%**	7.5% / 8.2%	13.2% / 10%
Neutral	20% / 20%	9.9% / 6.8%	**51% / 59%**	25% / 4.9%	3.5% / 3%
Sad	6.2% / 3.4%	12.4% / 11.3%	35% / 41.1%	**33.1% / 35.1%**	8% / 5.6%
Sad → neutral	5.1% / 4.4%	7.9% / 13.3%	**40% / 36.9%**	**34.2% / 29.4%**	2.8% / 3.3%
Surprise	13.2% / 13.1%	6.8% / 9.7%	12.5% / 11.8%	4.3% / 4.4%	**47.4% / 36.9%**
Surprised → neutral	7.4% / 9.5%	5.1% / 7%	**16.5% / 34.6%**	8.6% / 5.7%	**35.2% / 20.1%**

Table 3.4: Participant emotion labels for EmoVC synthesized audio, separated by male / female speakers. The numbers indicate likelihood of participant labels matching the dataset labels.

in robustness are needed for models such as this, before they can be deployed as convincing voice conversion tools in the real world.

3.4.2 Emotion Labels

For both non-synthesized and EmoVC surveys, overall participant responses generally favored more towards the dataset labels for "angry", "neutral", and "surprise" than any other label. The responses for the "sad" emotion clips though show less consensus, and were more frequently labelled as "neutral" instead, for both synthesized and non-synthesized. For EmoVC survey, participants were presented a random sample of synthesized audio clips. Thereby, they would remain unaware that any clips were intentionally converted towards a neutral tone. As shown in Table 3.4, participant responses show less consensus of the unconverted emotion clips. However, the labels of the audio clips of the altered speech fol-

Emotion Label	"other" Labels (male speaker)	"other" Labels (female speaker)
Anger	grumpy, apathy, explanatory, relieved, irritated, arrogant, condescending	annoyed, frustrated, hostility, sarcastic, irritated, serious, stern
Happiness	irritated, frustration, wanting, amused, sarcastic	annoyed, passive aggressive, irritated, hyper, fake happy, giddy
Sadness	bored, annoyed, disgust, agitated, upset, dread, exasperated, exhausted, stressed, ashamed, condescending, dejected, disappointed, grateful, sarcastic, monotone, smug, frustrated	bored, apathetic, numb, stilted, pretentious, relaxed
Surprise	confused, inquisitive, irritated, stern, stressed, shocked	confused, questioning, annoyed
Neutral	defeated, annoyed, apathetic, disgruntled, focused, insistent, bored	annoyed, bored, stilted, irritated

Table 3.5: Participant "other" labels for ESD non-synthesized emotion label task.

lowing the EmoVC transformations do show the model effectively reduces the emotion tone. These results coincide with previous work [84], exemplifying how people may not collectively interpret emotion tones the same, making it difficult to categorize short samples of speech into a single emotion tone.

Comparison with IEMOCAP Annotators. To further evaluate how consistently humans recognize emotions in speech, we compare annotation data of the non-synthesized ESD data from our user study with another emotion speech dataset (IEMOCAP). IEMOCAP dataset consists of 10 actors (5 male, 5 female) in scripted and non-scripted (improvised) settings. Each utterance in IEMOCAP is on average 4.46 seconds, (comparable to ESD). The

Emotion Label	Improvised	Scripted	Total
Anger	289	814	1103
Disgust	1	1	2
Excited	663	378	1041
Fear	8	32	40
Frustration	971	878	1856
Happiness	284	311	595
Sadness	608	476	1084
Surprise	60	47	107
Neutral	1099	609	1708
XXX	800	1707	2507
Total	4784	5255	10039

Table 3.6: Detailed breakdown of emotion data in the IEMOCAP dataset. XXX labels indicate the annotators did not agree on the label.

Emotion Label	"other" Labels
Anger	frustration, annoyed, disgust, surprise, neutral, excited, sadness, fear, irritation, sarcasm
Disgust	frustration, excited
Excited	happiness, neutral, surprise, frustration, fear, relief, pride, curiosity, suspicious, intrigued
Fear	excited, neutral, frustration, surprise
Frustration	anger, neutral, sadness, disgust, excited, fear, surprised, annoyed, confused, sarcasm
Happiness	excited, neutral, frustration, surprise, sadness, amused, reminiscent, content, nostalgic, curious
Sadness	frustration, neutral, happiness, anger, fear, concern, excited, disgust, surprise, remorse
Surprise	frustration, excited, happiness, fear, anger, neutral, disgust, disbelief, confused, sadness
Neutral	frustration, excited, sadness, happiness, anger, fear, surprise, concern, disgust, supportive
XXX	neutral, frustration, excited, happiness, anger, sadness, surprise, annoyed, fear, disgust

Table 3.7: Top 10 next most mentioned annotations for each emotion in the IEMOCAP dataset, in order of most mentioned to least.

short utterances are significant because we consider the annotations for both datasets to be determined based on thin slices, meaning that the content of the speech remains irrelevant. A breakdown of the utterances by emotion can be found in Table 3.6.

Each utterance is annotated by 3 or 4 people, to categorize the audio clips into labelled emotions (neutral, happy, sad, excited, fear, anger, frustration, surprise, disgust, other). For each utterance, annotators could choose one or more emotion labels, and were marked "xxx" if less than 3 labels match a single emotion. Additional data provided with the IEMOCAP dataset includes all original annotation labels for each utterance. With this, we can analyze

how consistently annotators perceived the same emotion for any given utterance.

Overall, as shown in Table 3.6, annotators did not collectively agree on a single label for ∼25% of all utterances. Further, since annotators could select more than 1 emotion label for each utterance, most utterances received more labels than the number of annotators. Table 3.7 shows the top 10 "other" labels selected for each emotion category (i.e., utterances labeled "angry" were most frequently also labeled "frustration"). We found that for each utterance of each emotion category (excluding "xxx" labeled utterances), only approximately 50% of the annotation labels matched the final emotion label. This suggests the annotators frequently perceived multiple possible emotions for each utterance.

The results from our user study reflect similar levels of annotation agreement to that of the IEMOCAP annotations. Overall, the results for the non-synthesized data annotations show plurality agreement with the original labels provided by the dataset. However, similar to IEMOCAP annotations, most of our participants did not strongly agree on a *single* emotion label. The "other" labels for the ESD dataset did, though, show significant overlap to the "other" labels for the IEMOCAP dataset, as shown in Table 3.5. From this, we can see that the annotators perceive emotion tones in similar clusters rather than single specific emotion labels, even across different datasets and populations of annotators.

Key Takeaways. For the non-synthesized speech, participants were more likely to choose the emotion label provided with the dataset than any of the other emotions, with the exception of "sad". For the speech converted towards neutral, participants were more likely to label it as neutral after the conversion. However, there was a significant amount of variation and very little consensus overall as a collective regarding a single emotion label. These results suggest that using emotion datasets with only a single label is insufficient and could lead to biased interpretations of emotions by researchers and the models they may train.

Emotion Label	Participant Ratings (M / F)						
	anxious	**negative**	**competent**	**polite**	**positive**	**trustworthy**	**warm**
Angry	0.11 / 0.32	0.25 / 0.31	-0.19 / -0.24	-0.44 / -0.33	-0.12 / -0.19	-0.21 / -0.27	-0.11 / -0.25
Happy	0.01 / 0.02	-0.15 / -0.59	-0.01 / -0.15	0.12 / 0.49	0.26 / 0.74	0.11 / 0.21	0.29 / 0.66
Sad	0.12 / 0.19	0.2 / 0.05	-0.2 / -0.27	-0.07 / 0.07	-0.11 / -0.06	-0.22 / -0.15	-0.13 / 0.07
Surprise	0.57 / 0.55	0.17 / -0.09	-0.54 / -0.51	-0.32 / -0.14	-0.25 / 0.03	-0.39 / -0.29	-0.14 / 0.09

Table 3.8: ESD non-synthesized pair preferences for traits. -1 indicates association with neutral tone, +1 indicates association with emotion tone

Emotion Label	Participant Ratings (M / F)						
	anxious	**negative**	**competent**	**polite**	**positive**	**trustworthy**	**warm**
Angry	0.4 / 0.22	0.18 / 0.27	-0.27 / -0.26	-0.19 / -0.24	-0.07 / -0.16	-0.2 / -0.24	-0.18 / -0.2
Happy	0.42 / 0.26	0.03 / -0.13	-0.31 / -0.18	-0.04 / 0.06	0.14 / 0.19	-0.12 / -0.09	0.02 / 0.1
Sad	0.14 / -0.03	0.03 / -0.13	0.02 / 0.03	0 / 0.11	0.01 / 0.11	-0.09 / 0.14	-0.02 / 0.21
Surprise	0.6 / 0.34	0.35 / 0	-0.42 / -0.34	-0.24 / 0	-0.21 / 0.08	-0.32 / -0.19	-0.25 / 0.02

Table 3.9: EmoVC pair associations for traits. -1 indicates association with neutral tone, +1 indicates association with emotion tone.

3.4.3 Personal Characteristic Associations

The way people *express* emotion often coincides with assumptions of one's personal characteristics [**?**]. However, the way people *interpret* an expression of emotion can vary widely [84]. For instance, a "happy" tone of voice often can also be interpreted as an expression of joy, pleasure, amusement, and/or relief, among others, Similar to the categorization of emotion expressions, in this study, we categorize "competent", "polite", "positive", "trustworthy", and "warm" as *desirable* traits, and associated with positive emotions (i.e., happy, surprised). Conversely, "anxious" and "negative" are characterized as *undesirable* traits, and associated with negative emotions (i.e., angry, sad).

As expected, for both the non-synthesized and EmoVC synthesized data, we find that the "angry" and "sad" tones are consistently associated with undesirable traits, and not at all associated with the desirable traits. An exception to this is the "sad" results for the EmoVC data. This is likely explained by our results in Table 3.4, which shows that "sad" was labeled as "neutral" more frequently than "sad", suggesting the emotion tones were interpreted as very similar. Previous work shows that expression of "sad" and "neutral" tones often present with very similar frequencies, which are also close to frequencies associated with

Scenarios
Suppose you were **talking on the phone to a telephone operator** (e.g., bank, airline, cable company). Which voice would you prefer to speak with on the other end of the line?
Suppose you were **watching a political debate**. Which voice would you prefer to vote for?
Suppose you were **interviewing personal assistants for yourself**. Which voice would you prefer to hire as your assistant?
Suppose you were **going to have surgery**. Which voice would you prefer to be your surgeon?
Suppose you were **going to have a job interview**. Which voice would you prefer to interview you?

Table 3.10: Scenarios presented for part 3 of feasibility user study

Emotion Label	Scenario Participant Preferences (M / F)				
	Telephone Operator	Debate Candidates	Personal Assistant	Surgeon	Job Interviewer
Angry	-0.28 / -0.5	0 / -0.34	-0.26 / -0.51	-0.38 / -0.54	-0.33 / -0.52
Happy	-0.02 / 0.11	0.01 / -0.14	0.03 / 0.25	-0.06 / -0.37	0.17 / 0.34
Sad	-0.35 / -0.17	-0.21 / -0.38	-0.28 / -0.24	-0.06 / -0.39	-0.22 / -0.22
Surprise	-0.43 / -0.37	-0.38 / -0.31	-0.37 / -0.3	-0.63 / -0.65	-0.35 / -0.29

Table 3.11: ESD non-synthesized associations for scenarios. -1 indicates association with neutral tone, +1 indicates association with emotion tone

boredom [128]. Therefore, when the EmoVC model was used, it's possible the attempted conversion towards neutral could have amplified the frequency and resulted in speech that instead was interpreted as even more sad sounding.

Conversely, the "happy" tone of voice is consistently more strongly associated with desirable traits than the neutral toned voice (with the exception of "competent"), particularly for the female speaker. The response to surprise contradicts those for happy, with surprise showing strong correlations to "anxious" and "negative". Upon reflection, this result could be due to an assumption of a "good" surprise. However, in real life it's possible to experience both good and bad surprises, and it remains unclear if the voice actors were given any instructions on how to express the emotion.

Key Takeaways. Participants rated "competent" and "trustworthy" as associated closer to neutral tones rather than any of the emotions (with the exception of happy-trustworthy for non-synthesized speech). This suggests a more neutral tone of voice could enhance the perception of these desirable traits, as opposed to a strong expression of any of these emotions.

Emotion Label	Scenario Participant Preferences (M / F)				
	Telephone Operator	**Debate Candidates**	**Personal Assistant**	**Surgeon**	**Job Interviewer**
Angry	-0.34 / -0.18	-0.07 / -0.23	-0.31 / -0.27	-0.28 / -0.25	-0.39 / -0.16
Happy	-0.07 / -0.21	-0.02 / -0.24	-0.01 / -0.17	-0.31 / -0.26	0 / -0.16
Sad	-0.05 / -0.02	0.14 / 0.07	0.05 / 0.07	-0.01 / -0.1	0.09 / 0.05
Surprise	-0.27 / -0.2	-0.12 / -0.04	-0.34 / -0.09	-0.42 / -0.56	-0.14 / -0.01

Table 3.12: EmoVC associations for scenarios. -1 indicates association with neutral tone, +1 indicates association with emotion tone

3.4.4 Scenario Preferences

To further explore these results, we examined preferences for emotion tones compare to neutral tones in various scenarios. To understand the real world implications of changing one's tone of voice, participants were asked to indicate their preference for either the emotional toned voice or the (more) neutral tone of voice given a scenario. The scenarios, shown in Table 3.10, were selected based on previous work [145, 7, 72, 178, 10, 26] showing how emotion tone may affect the perception and/or outcome in these situations. For instance, emotion tone can affect one's level of trust in their doctor or surgeon, expectations for treatment outcomes [10], or the likelihood a surgeon may be sued for malpractice [7]. The tone of voice of a telephone operator can affect how quickly they assist patrons with an issue [72]. Alternately, anger between a call center operator and caller can escalate and result in distressing experiences for both parties [145]. When watching political debates, voters evaluate candidates' emotion expressions to determine if they possess an ability to lead and work with others [26]. We predict that due to the nature of the scenarios presented, competency and trustworthiness would be expected as desirable traits. Based on the responses regarding character traits, participants associated competency and trust with more neutral tones, so we expect that neutral tones would be preferred for these scenarios as well.

Participants overwhelmingly prefer more neutral voices in almost all scenarios. For the ESD non-synthesized utterances, only 2 instances show a preference for the emotional tone over the neutral tone (happy is preferred for "personal assistant" and "job interviewer"). For the EmoVC converted utterances, in all scenarios the participants show preference for

the (more) neutral tone of voice over the angry, happy, and surprise tones. For the EmoVC converted data specifically, the preference for sad or (more) neutral shows mixed results. As previously discussed, we believe this to be due to the way the model converts sad to neutral, and the conversion process may have resulted in a tone that actually sounds *more* sad rather than more neutral.

On the other hand, while some emotional tones may at times be considered "positive" (e.g., polite) [84], the ratings for these characteristics did not correlate with preferences for any of the given scenarios. Based on our results from the characteristic pair associations in §3.4.2, we found that happy was consistently associated with positive traits, and therefore would be expected to correlate with preferences for these scenarios as well. However, happy was only shown as preferential for the "personal assistant" and "job interview" scenarios for the non-synthesized ESD data, and not preferred at all over the (more) neutral tone for the converted data. This may be related to the nature of the job and/or interaction. For instance, in a high stakes environment, like undergoing surgery, when one may not prefer a surgeon with a happy disposition, but rather a person with more seriousness in their tone of voice. These results coincide with our previous findings that a less emotional tone was considered more competent.

Key Takeaways. We found that tone of voice needs to be appropriate for differing situations. For situations with higher stakes, overtly expressive emotions of any kind may not be preferable, but rather more muted emotions. Following our previous results, more neutral emotional tones (that closer coincide with ratings for competence and trustworthiness) are preferred in situations that could be of higher stakes for those involved (i.e., surgery).

3.5 Desirability of Emotion Conversion

Survey Design. Drawing from methodology of recent work [69], this study is intended to evaluate the *desirability* and *acceptability* of using tools like these to manipulate voice in the

real world, and why (or why not). Guerouaou et al. [69] studied the ethics of changing tone of voice, and discovered participants found it more morally acceptable to reduce negative tones (i.e., anger) than to increase positive tones. Most interesting, they also found that people would consistently not allow any manipulations to be done to others that they would not allow for themselves. This study seeks to expand upon the results of our first user study to determine whether users find using such a tool as desirable and morally acceptable.

Key Results

- Overwhelmingly, participants would not be interested in ever using a tool to manipulate their own voice.
- Approx. 1/5 of participants wouldn't mind others using a tool to improve effectiveness of communication
- Consent is most important consideration, particularly regarding the one who's voice is being altered

Participants. Again, we recruited participants from the online crowd sourcing platform Prolific. The survey was designed to take 5 minutes on average, and participants were compensated $1. We only recruited participants ages 18+ from the US, with at least 100 previous studies completed, and 95% approval rate. We collected valid responses from 100 participants. Our participant pool included 59 females (39 males, 2 other/prefer not to say), covering various age groups: 18-29 (2%), 30-39 (45%), 40-49 (25%), 50-59 (7%), and 60+ (7%) years old.

Task. First, we described a potential tool that would allow people to subtly alter their tone voice. Then, we provided examples of what the voice would sound like before and after emotion conversion (i.e., angry → less angry), and some potential scenarios when these

Reason	Examples
improves experience for all parties	"If it **improves the effectiveness of the speaker**, I dont necessarily see anything morally wrong about it."
reduce negative emotion	"If you are **having a really bad day** it would be helpful to use so you **don't come off quite as angry**"
workplace	"virtual work meeting" "as any customer service person" "for things **work or school related it is fine**...at a job interview and your nervous."
as a joke / for fun	"**playing a guessing game/for fun**" "Pranks to deceive intentionally for comedic affect"
media jobs	" for a **film or video content narration**"
anonymity	"wanted to **remain anonymous**"
with consent of speaker	"It doesn't hurt anyone, and if the person using it is aware, then it's a choice they're making on how they present themselves. **it has nothing to do with the audience**"
help someone with a speech disability	"If the person was registered for disability services, like with their school or with ADA." "When the person has some type of **medical condition causing the individual to have less control of emotion**"
always acceptable	"I think it's **acceptable under pretty much any circumstance**"

Table 3.13: Responses for acceptable circumstances for altering tone of voice.

changes may be desirable (i.e., reduce an angry tone of someone talking to a call center operator, or having a meeting over Zoom). Next, the participants were asked whether they would use such a tool, and what circumstances using such a tool would be morally acceptable. Particularly, we ask if it would be acceptable to use a tool if *only the speaker knew it was being used*, or what (if any) time it would be acceptable to use if the *speaker did not know* the tool was being used to alter their voice. Lastly, we ask specifically if (and under what circumstances) it would be acceptable to (1) reduce anger, (2) reduce anxiety, or (3) increase positivity, and why or why not.

Reason	Examples
deception, manipulation	"to **blackmail** someone"
workplace	"Undesirable for high value conversations, like with investment holders in large companies or government decisions" "it's **unacceptable to falsify anything in a professional setting**."
tool may not work properly	"I personally wouldn't utilize such a tool, mainly for **fear that it could malfunction** while I'm using it"
personal interaction	"would be undesirable if **talking heart to heart** to someone"
children involved	"unacceptable if there were **small children involved**"
if either party is not aware	"people should be aware of such unnatural things...it's **like deceiving and not showing how someone else really acts and/or talks**."
never acceptable	"It's someone's **responsibility to watch their tone**, and this tool would not reduce the impact of the words being used" "totally against voice modulation"

Table 3.14: Responses for unacceptable circumstances for altering tone of voice.

Results. ***People prefer authentic expressions.*** Participants indicated they would "definitely not" (58%) or "rarely" (24%) find it acceptable to use the tool for themselves. Although many participants can see the potential upsides to reducing negative emotions or traits (i.e., anger, anxiety), many still prefer speakers to express their honest emotions, or learn how to control them on their own. Throughout the survey, participants frequently stated that the use of such a tool would diminish the *honesty* and/or *authenticity* of the entire human interaction. When asked under what circumstances it would be acceptable to alter tone of voice, responses suggested low-stakes situations when it may be alright, such as subtle changes that didn't interfere with critical business scenarios or attempt to maliciously deceive the other party (see Table 3.13 and Table 3.14). Overall, participants showed overwhelming disapproval for altering someone's voice without their knowledge (76%), especially when

asked about reducing anxiety to increase perceived confidence. When asked about their reasons for disapproval, most stated it would violate the speakers autonomy if they did not consent.

Participants saw benefits for both sides in reducing negative emotions. When asked if they found it morally acceptable for reducing anger, only 24% found it to be *never acceptable*. Particularly, given the scenario when a call center operator deals with an angry caller, participants expressed sympathy for the operator and were more likely to be accepting of the caller being unaware of reducing the anger in their voice. However, many participants repeatedly stated that reducing the angry tone too much could reduce the effectiveness of the call center operator's ability to understand the frustrated emotional state of the caller.

Additionally, as fear of public speaking affects millions of people around the world, many participants sympathized with the potential benefits of reducing an anxious tone of the speaker. For the speaker, many stated this could increase their confidence, and as potential audience members themselves, they recognized how this could increase the effectiveness of the presentation. However, some also stated anxiety is something that should simply be expected and not artificially manipulated for the benefit of the speaker.

Most participants don't care about increasing positivity. While they were more likely to see this type of manipulation as generally benign with few to no foreseeable downsides, but thereby also unnecessarily manipulative to warrant use of such a tool. These reactions echoed a common theme of avoiding unnecessary lies in favor of authentic human interactions. This result is in line with previous work [69], which found that people were generally more accepting of alterations that would reduce negative emotions than to artificially boost positive ones.

3.6 Discussion

This work collectively evaluates correlations between the way people speak and biased perceptions of the speaker. Specifically, we analyzed emotion speech datasets and how perceived emotional tone of voice relates to assumed characteristics of the person speaking. We found that people interpret tone of voice along a spectrum of many similar/related emotions, rather than any one single emotion. Additionally, we measured the change in perception from altering vocal tone with emotion voice conversion tools. Our findings show that more neutral tones of voice are associated with stronger impressions of competence and trustworthiness. Further, we examined how emotional tones correlate with scenarios involving personal interactions, such as job interviews or interactions with telephone operators. We found that more neutral tones (less emotional expressiveness) were preferential for situations that could be considered high stakes (e.g., medical surgeons, an applicant attending a job interview). This suggests that tampering down emotions could prove to be beneficial in high stakes situations. These results show potential for improving human interactions.

However, participants expressed a marked discomfort in the idea of artificially changing their tone of voice. Particularly, when given the option of using a tool for themselves, participants showed strong preference of authentic interactions over those that could be perceived as more negative, or positive. The results from these studies show how difficult it may be to categorize, and alter, emotional tone of voice and the implications for maintaining authentic interactions, both for the speaker and listeners.

Limitations. Our analysis of voice actor data may not fully represent perceptions of emotional expression, as voice actors purposefully trying to express emotions does not always represent natural emotion expressions. We attempt to remove content and context as variables for consideration, which often prove significant factors in perceptions during real world interactions. This work focused on evaluating the feasibility of these tools to tamper emotional tone towards neutral, but there still remains a significant unknown space regarding how these tools can be applied for emotion to different combinations of emotion conversions.

The number of combinations for all possible emotions was prohibitive in this study, as it would have resulted in far too many utterances and too costly to be evaluated in a user study. Also, our user studies only gathered responses from participants in the United States, and may not fully generalize to all other populations.

Future Work. Further research is needed to understand which transformations may be most beneficial for particular scenarios. In most instances someone may want to transform their voice, they are most likely not starting at a neutral tone and would therefore want to convert a "negative" tone to a "positive" one (e.g., sad $\rightarrow$ happy). After determining the likely desired transformations for various scenarios, a tool such as these could be implemented as part of a pipeline. Starting with a speech emotion recognition model, unlabeled speech paired with a scenario could be input into a conversion model to achieve the desired outcome. Many companies today record and analyze job interviews, such as HireVue. Potential applicants may not express confidence, or express too much excitement, and could be rejected based simply on their tone of voice, despite being fully qualified. Additional research in the ways tones of voice are analyzed and how that information is used could benefit both potential job applicants and hiring companies.

CHAPTER 4

PERCEPTIONS AND SECURITY OF ARTIFICIALLY GENERATED CONTENT IN TEXT, VOICE, AND IMAGES

Deep learning presents great potential, while also posing a great threat to people and society at large. As these models continue to become more sophisticated, the content created becomes more and more difficult for humans and systems to distinguish as artificially generated. In the domains of text, voice, and images, artificially generated content continues to become more pervasive in society, with too little consideration from developers regarding the implications. Here, I present several user studies that evaluate how users perceive artificial generated content, and the implications for society.

4.1 Automated Crowdturfing Attacks and Defenses in Online Review Systems

Malicious crowdsourcing forums are gaining traction as sources of spreading misinformation online, but are limited by the costs of hiring and managing human workers. In this paper, we identify a new class of attacks that leverage deep learning language models (Recurrent Neural Networks or RNNs) to automate the generation of fake online reviews for products and services. Not only are these attacks cheap and therefore more scalable, but they can control rate of content output to eliminate the signature burstiness that makes crowdsourced campaigns easy to detect.

Using Yelp reviews as an example platform, we show how a two phased review generation and customization attack can produce reviews that are indistinguishable by state-of-the-art statistical detectors. We conduct a survey-based user study to show these reviews not only evade human detection, but also score high on "usefulness" metrics by users. Finally, we develop novel automated defenses against these attacks, by leveraging the lossy transforma-

tion introduced by the RNN training and generation cycle. We consider countermeasures against our mechanisms, show that they produce unattractive cost-benefit tradeoffs for attackers, and that they can be further curtailed by simple constraints imposed by online service providers.

4.1.1 Attack Methodology

We focus our study on Yelp, the most popular site for collecting and sharing crowdsourcing user reviews. Yelp's review system is representative of other review systems, e.g., Amazon or TripAdvisor.

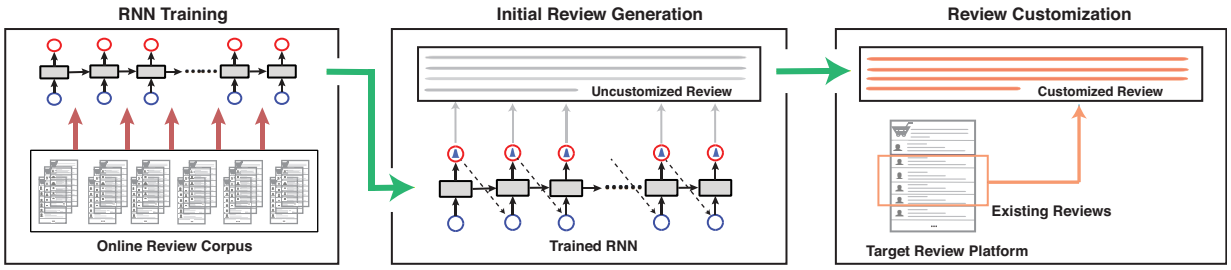


Figure 4.1: Overview of our attack methodology.

Our attack methodology is illustrated in Figure 4.1. At a high level, the attack consists of two main stages: (1) The first stage starts by training a generative language model on a review corpus. The language model is then used to generate a set of initial reviews. (2) In the second stage, a customization component further modifies these reviews to capture specific information about the target entity (e.g., names of dishes in a seafood restaurant), and produces the final targeted fake review. In our experiments, the customizable content is extracted from a *reference dataset*, composed of existing reviews associated with the target entity. If there are no existing reviews, an attacker can build a reference dataset using reviews of entities in the same category (e.g., seafood restaurants) as the target. Restaurant category metadata is available on Yelp and similar sites, and can be used to identify similar

entities.

Generating Initial Reviews. First, the attacker chooses a training dataset that matches the domain of the target entity. For example, to generate reviews targeting restaurants, the attacker would choose a dataset of restaurant reviews. Next, the attacker trains a generative RNN model using the dataset. Afterwards, the attacker generates review text using a sampling procedure. Note that the attacker is able to generate reviews at different *temperatures*.

Review Customization. In general, there is no control over the topic or context (e.g., name of a food in a restaurant) generated from the RNN model, since the text is stochastically sampled based on the character distribution. To better target an entity (e.g., restaurant), we further capture the context by customizing the generated reviews with domain-specific keywords. This is analogous to crowdsourced fake review markets, where workers are typically provided additional information about the target entity for a writing task [141]. The information consists of specific nouns (e.g., names of dishes) to be included in the written review. Based on this observation, we propose an automated noun-level word replacement strategy.

Our method works by replacing specific words (nouns) in the initial review with new words that better capture the context of the target entity. This involves three main steps:

1. *Choose the type of contextual information to be captured.* The attacker first chooses a keyword C that helps to identify the context. For example, if the attacker is targeting a restaurant, the keyword can be "food," which will capture the food-related context. If the target is an online electronic accessories store, then the keyword can be "accessory" or "electronics."

2. *Identify words in reviews of the reference dataset that capture context.* Next, our method identifies all the nouns in the reference dataset that are relevant to the keyword C.

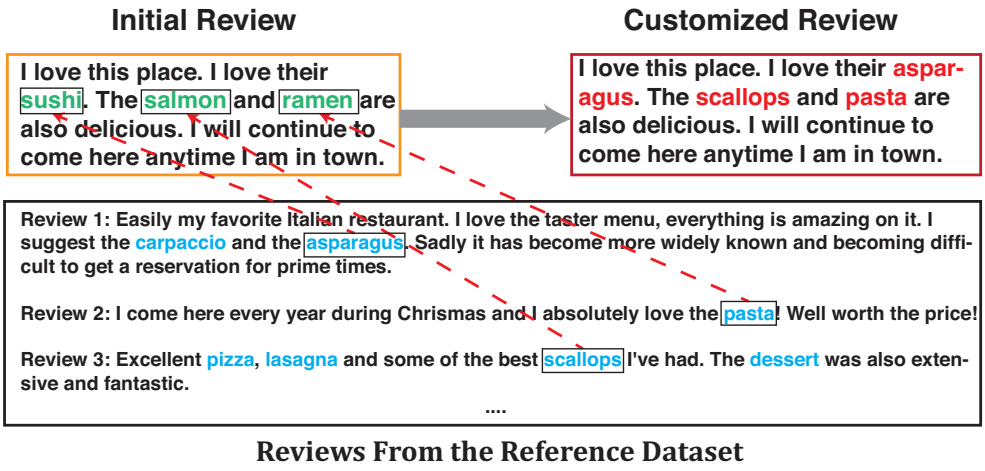


Figure 4.2: Example of review customization.

Relevancy is estimated by calculating lexical similarity using WordNet [116], a widely used lexical database that groups English words into sets of synonyms and measures their concept relatedness [132]. We identify a set of words p in the reference dataset that have high lexical similarity with the keyword C, using a similarity threshold MIN_{sim}. The set of words p captures the context of the target entity.

3. *Identify words in initial reviews for replacement.* Finally, we find all the nouns in the review set R that are also relevant to C using the same method in Step 2. We replace them by stochastically sampling words in p based on the lexical similarity score.

Figure 4.2 shows an example of customizing an initial review that has language more suitable for a Japanese restaurant, to a review more suitable for an Italian restaurant. The nouns to be replaced in the initial review are marked in green, and replacement nouns are marked in blue. Note that we choose this noun-level replacement strategy because of its simplicity, and there is scope for further improvement of this technique.

4.1.2 Evaluating Quality of Machine-Generated Reviews

We evaluate the quality of machine-generated reviews along two dimensions. First, we investigate whether generated reviews can bypass detection by existing algorithmic approaches. Second, we conduct an end-to-end user study, by presenting restaurant reviews containing both generated reviews and real reviews to human judges. Our goal is to understand whether humans can distinguish generated reviews from real reviews.

Detection by Existing Algorithms. We train a linear SVM classifier on the Yelp ground-truth dataset, composed of real reviews (Yelp unfiltered reviews), and fake reviews (Yelp filtered reviews). After training with all 77 linguistic features, we tested the performance of the classifier on the Yelp attack dataset, composed of real reviews and machine-generated reviews. We run 10-fold cross validation and report the average performance.

Evaluation of attack performance uses *precision* (percentage of reviews flagged by the classifier that are fake reviews), and *recall* (percentage of fake reviews flagged by the classifier). Overall, we observe high performing attacks at all temperatures. The best attack is at temperature 1.0, with a low precision of 18.48%, and a recall of 58.37%. Low precision indicates the inability of the ML classifier to distinguish between real reviews and machine-generated reviews.

We observe that attack performance increases with temperature. To further understand this trend, we analyze how the linguistic features of the generated text vary as we increase temperature. We compare the average value of a linguistic feature of generated reviews with real reviews at different temperatures. In general, feature values of the machine-generated reviews diverge from real reviews at low temperatures, and converge as temperature increases, thus making it harder to distinguish them from real reviews at high temperatures.

Evaluation by User Study. Regardless of how well machine-generated reviews perform on statistical measures and tests, the real test is whether they can pass for real reviews

when read by human users. We conducted an end-to-end user study to evaluate whether human examination can detect machine-generated reviews. In practice, service providers are known to involve human content moderators to separate machine-generated reviews from real reviews [174]. More importantly, these tests will tell us how convincing these reviews are to human readers, and whether they will accomplish their goals of manipulating user opinions.

User Study to Detect Machine-Generated Reviews. To measure human performance, we conduct surveys on Amazon Mechanical Turk. Each survey includes a restaurant name, description (explaining the restaurant category and description provided by the business on Yelp), and a set of reviews, which includes machine-generated reviews and real reviews written for that restaurant. We then ask each worker to mark reviews they consider to be fake, using any basis for their judgment.

For our survey, we choose 40 restaurants with the most number of reviews in our ground-truth dataset. For each restaurant, we generate surveys, each of which include 20 random reviews, out of which some portion $(0-5)$ are machine-generated reviews, and the rest are real reviews from Yelp. We show an example of our survey in the Figure C.1 in Appendix C.1.

Figure 4.3 shows the human performance results as we vary the temperature. First, we observe that machine-generated reviews appear quite robust against a human test. In addition, similar to algorithmic detection, attack performance improves as temperature increases. This is surprising, since we would expect that reviews at the extreme high or low temperature parameters would be easily flagged (either too repetitive or too many grammatical/spelling errors). We saw earlier that higher temperature produced reviews more statistically similar to real reviews, but expected errors to make those reviews detectable by humans. Instead, it seems that human users are much more sensitive to repetitive errors than they are to small spelling or grammar mistakes.

Helpfulness of Machine-Generated Reviews. Previously, we showed that humans tend to mark many machine-generated reviews as real. This raises a secondary question: *For machine-generated reviews that are not caught by humans, do they still have sufficient quality to be considered useful by a user?* Answering this question, takes us a step further towards generating highly deceptive fake reviews.

In each survey, we first asked the workers to mark reviews as fake or real. Additionally, for the reviews marked as real, we asked for a rating of the usefulness of the review on a scale from 1 to 5 (1 as least useful, 5 as most useful). An example of the survey is shown in the Figure C.2 in Appendix C.1.

The average usefulness score of false negatives (unflagged machine-generated reviews) is close to that of true negatives (unflagged Yelp real reviews): machine-generated reviews have an average usefulness score of 3.15, which is close to the average usefulness score of 3.28 for real Yelp reviews. That is to say, workers think of unflagged machine generated reviews almost as useful as real reviews.

Overall, our experiments find machine-generated reviews very close to mimicking the quality of real reviews. Furthermore, the attacker is incentivized to generate reviews at high temperatures, as such reviews appear more effective at deceiving users.

4.1.3 Discussion

This work focuses on the potential for misuse of deep learning models in the context of attacking online review platforms. Our work shows how RNNs can generate deceptive yet realistic looking reviews targeting restaurants on Yelp. An extensive evaluation of the quality of generated reviews indicates the difficulty in detecting such reviews using existing algorithmic approaches, and even by human examination (which serves as an end-to-end test of our attack).

Due to the information loss, generated reviews diverge from real reviews when comparing

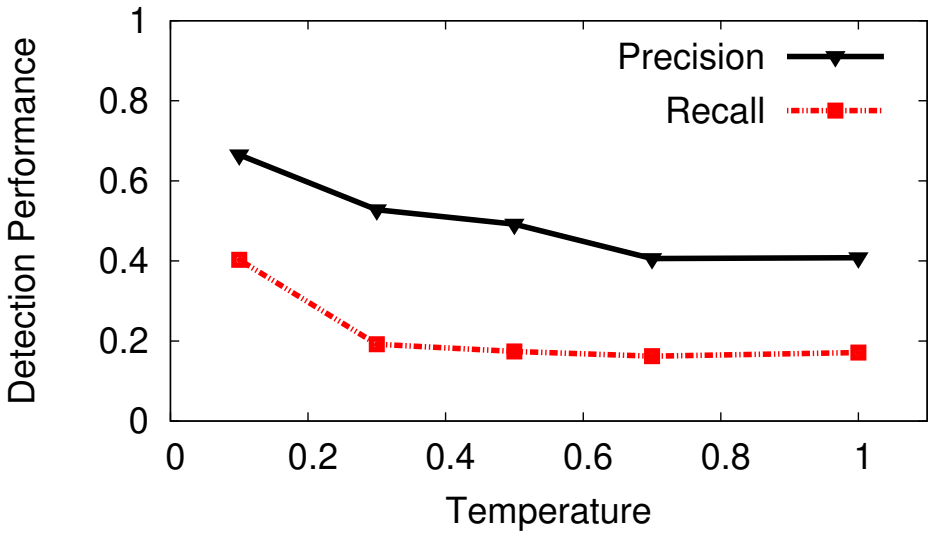


Figure 4.3: Performance of human judgment on detecting machine-generated review.

their character-level distribution, even when higher level linguistic characteristics are preserved. Our scheme, based on supervised learning, can detect machine-generated reviews with high accuracy (F-score ranging from 0.8 to 0.98 depending on the amount of available ground-truth) and outperforms existing ML-based fake review filters.

In terms of potential future work, one direction is to consider the role that user and content metadata can play in both the attack and defense perspectives. Metadata can be crucial in terms of deceiving users (e.g., by increasing the number of friends/contacts on the site) and in assisting defenses [179, 79, 56, 80, 122, 95, 182] (e.g., by analyzing the patterns in timestamps of user activities). Orchestrating the general behavior of user accounts using deep learning to bypass metadata based defenses could be an interesting research challenge. Second, while we limit ourselves to the domain of online review systems and fake review attacks, deep learning-based generative text models can be applied to launch attacks in other scenarios as well. We highlight two of these possible application scenarios.

Strengthening Sybil Attacks. Attackers can use our techniques to generate realistic looking text-based user behavior patterns [25], e.g., posting, commenting and messaging. This can

help attackers make Sybil (fake) accounts indistinguishable from legitimate accounts based on textual content. A special case of this involves launching an impersonation attack in online social networks [64].

Fake News Generation. Identifying fake news, i.e., "a made-up story with an intention to deceive" [150], currently remains an open challenge [53]. The research community has started to explore the possibility of automating the detection process by building an AI-assisted fact-checking pipeline [180, 113, 185]. We now know that AI can not only assist fake news detection but also generate fake news. Given the availability of large-scale news datasets [171], an attacker can generate realistic looking news articles using deep-learning models. Due to its low economic cost, attackers can pollute social media newsfeeds with a large number of fake articles.

Ideally, these results will bring more attention to the problem of malicious attacks based on deep learning language models, particularly in the context of fake content on online services, and encourage the exploration and development of new defenses.

4.2 Deep Learning-based Speech Synthesis Attacks in the Real World

Our voice conveys so much more than the words we speak. Advances in deep learning have introduced a new wave of voice synthesis tools, capable of producing audio that sounds as if spoken by a target speaker. If successful, such tools in the wrong hands will enable a range of powerful attacks against both humans and software systems (aka machines). These speech synthesis attacks, particularly those enabled by advances in deep learning, pose a serious threat to both computer systems and human beings. Yet, there has been – until now – no definitive effort to measure the severity of this threat in the context of deep learning systems.

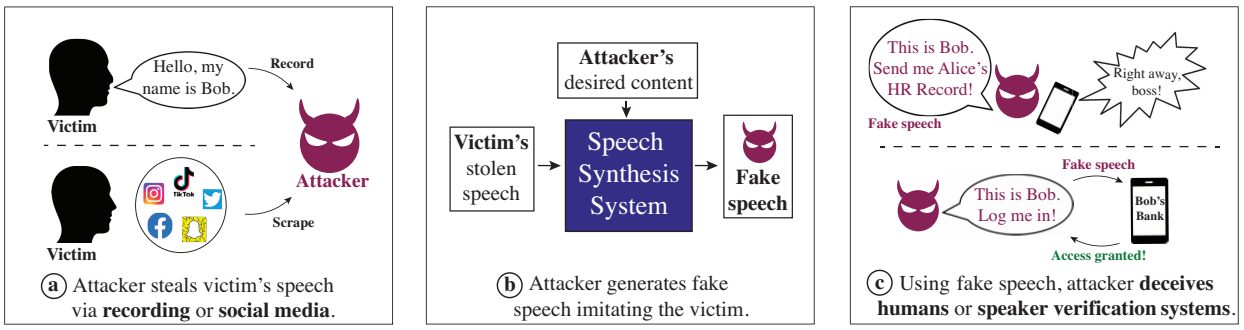


Figure 4.4: Workflow of synthesis-based voice spoofing attacks: (a) the attacker obtains voice samples from the victim, either by secretly recording them or by downloading available media; (b) the attacker then uses a speech synthesis system to generate fake speech, which imitates the victim's voice but contains arbitrary, attacker-chosen content; (c) the attacker uses this fake speech to impersonate the victim, e.g., attempting to access personal or financial information or conduct other attacks.

We believe there is an urgent need to measure and understand how deep-learning based speech synthesis attacks impact two distinct entities: *machines* (e.g., automated software systems) and *humans*. Can such attacks overcome currently deployed speaker recognition systems in security-critical settings? Or can they compromise mobile systems such as voice-signin on mobile apps? Against human targets, can synthesized speech samples mimicking a particular human voice successfully convince us of their authenticity?

In this work, we describe results of an in-depth analysis of the threat posed to both machines and humans by deep-learning speech synthesis attacks. We begin by assessing the susceptibility of modern speaker verification systems (including commercial systems Microsoft Azure, WeChat, and Alexa) and evaluate a variety of factors affecting attack success. To assess human vulnerability to synthetic speech, we perform multiple user studies in both a survey setting and a trusted context. Finally, we assess the viability of existing defenses in defending against speech synthesis attacks. All of our experiments use publicly available deep-learning speech synthesis systems, and our results highlight the need for new defenses against deep learning-based speech synthesis attacks, for both humans and machines.

Key Findings. Our study produces several key findings:

- Using a set of comprehensive experiments over 90 different speakers, we evaluate and show that DNN-based speech synthesis tools are highly effective at misleading modern speaker recognition systems ($50 - 100\%$ success).
- Our experiments find that given a handful of attempts, synthesized speech can mimic 60% of speakers in real world speaker recognition systems: Microsoft Azure, WeChat, and Amazon Alexa.
- A user survey of 200 participants shows humans can distinguish synthetic speech from the real speaker with $\sim$50% accuracy for unfamiliar voices but near 80% for familiar voices.
- An interview-based deception study of 14 participants shows that, in a more trusted setting, inserted synthetic speech successfully deceives the large majority of participants.
- Detailed evaluation of 2 state-of-the-art defenses shows that they fall short in their goals of either preventing speech synthesis or reliably detecting it, highlighting the need for new defenses.

It is important to note that speech synthesis is intrinsically about producing audible speech that sounds like the target speaker to humans and machines alike. This is fundamentally different from adversarial attacks that perturb speech to cause misclassification in speaker recognition systems [37, 96, 90]. Such attacks do not affect human listeners, and could be addressed by developing new defenses against adversarial examples.

4.2.1 Synthesized Speech vs. Machines

We begin by asking "*how vulnerable are machine-based SR systems to synthetic speech attacks*?" While prior work has explored this question using classical (non-DNN) synthesis systems, the efficacy of DNN synthesis attacks against real-world SR systems remains unknown. In this section, we answer this question by evaluating the robustness of four modern

SR systems to DNN-based synthesis attacks.

Specifically, our study consists of the following experiments:

- As a baseline, we **recreate prior classical synthesis attacks** and find that they fail against newer SR systems.
- An **attack on the widely used SR model** (Resemblyzer), shows that DNN-based synthesis attacks reliably fool such systems.
- **Attacks on three real-world SR deployments** (Azure, WeChat, and Amazon Alexa), show that all three are vulnerable to DNN-based synthesis attacks.

Performance is measured using the *attack success rate (AS)*, which denotes the average percent of synthesized samples identified as the target speaker. We design our experiments to not only evaluate the attack success rate against various SR systems, but also explore whether a target's speech samples and personal attributes (e.g., gender/accent) impact the attack outcome.

All four modern SR systems tested are vulnerable to DNN-based speech synthesis attacks, especially those generated by SV2TTS. It is alarming that for three popular real-world SR systems (Azure, WeChat, Alexa), more than 60% of enrolled speakers have at least 1 synthesized (attack) sample accepted by these systems. This clearly demonstrates the real-world threat of speech synthesis attacks.

Another key observation is that the attack performance is speaker-dependent, e.g., the number of synthesized samples that successfully fooled the SR systems varies across speakers. For Resemblyzer and Azure, the attack success rate is consistently higher for female and native English speakers.

Limitations and Next Steps. Our experiments, especially those on WeChat and Alexa, involved a moderately-sized set of target speakers to demonstrate the real-world threat of speech synthesis attacks. To further evaluate the attack dependence on target human speakers, we believe viable next steps include expanding the speaker pool and testing

more operational scenarios. With these two changes, we could more closely examine how an individual's vocal characteristics (e.g., pitch, accent, tone) affect the attack success rate, and whether their impact can be reduced by improving the underlying speech synthesis systems.

Similarly, due to our focus on low-resource attackers, our experiments used two publicly available speech synthesis systems (SV2TTS and AutoVC) that were trained only on publicly available datasets. These two systems will likely underperform advanced synthesis systems trained on larger, proprietary datasets, and consequently our reported results only offer a "conservative" measure of the threat. As speech synthesis systems continue to advance, the threat (and damage) of speech synthesis attacks will grow and warrant our continuous attention.

4.2.2 User Study A: Can Users Distinguish Synthesized and Real Speech?

We begin our human perception experiments with the critical question: "can human listeners distinguish synthetic speech of a speaker from the real thing?" We deploy a survey to assess users' ability to distinguish between real and fake speakers.

Participants. We recruited 200 participants via the online crowd source platform Prolific. All self-identified as native English speakers residing in the United States. Of our participants, 57% identified as female (43% male). The participants are all 18+ years old and cover multiple age groups: 18-29 (43%), 30-39 (32%), 40-49 (14%), 50-59 (8%), 60+(3%). The survey was designed to take 10 minutes on average, and participants received $2 as compensation.

Procedure. The participants completed an online survey consisting of several speech samples presented in pairs for side-by-side comparison. Each pair of samples contains one of the three following combinations: two real speech samples of the same speaker (referred to as "Real A/Real A" in this section); one real speech sample from a speaker and one real speech sample from a different speaker ("Real A/Real B"); or one real speech sample from a

speaker and one fake speech sample imitating the speaker ("Real A/Fake A"). We generate fake speech using SV2TTS, using 30 seconds of clean speech samples from the speaker.

Types of Speakers: We included speakers whose (real) voices have varying levels of familiarity with our participants:

- *Unfamiliar speakers:* Speakers from the VCTK [190] dataset whose voices have (most likely) never been heard by the participants.
- *Briefly familiar speakers:* Inspired by [123], we included a set of speakers whose voices the participants hear only briefly. For each speaker, we provided participants with a short audio clip to familiarize them with the speaker's voice. There are four briefly familiar speakers, and for each one we provided a different length audio clip – 30 seconds for the first speaker, 60 seconds for the second, 90 seconds for the third, and 120 seconds for the fourth.
- *Famous speakers:* We used the voices of two American public figures: Donald Trump and Michelle Obama. We asked participants whether they have heard these voices outside the context of this survey, and over 90% responded "yes."

Task. Participants listened to pairs of speech samples and reported if both samples were spoken by the same person.

Conditions. We deploy two versions of the survey. Both versions ask participants to assess the identity of the speaker and quality of speech samples. The first version does not mention fake speech at all. The second version of the survey mentions fake speech, both in its title and in its description of the task.

Results. We seek to answer the following questions:

1) Do participants think generated fake speech was spoken by the original speaker? As shown in Table 4.1 (bottom row), about half of participants were fooled, i.e., they responded "yes" or "not sure," when asked this question about *unfamiliar* or *briefly*

	Unfamiliar			Briefly Familiar			Famous		
	Yes	*Not Sure*	*No*	*Yes*	*Not Sure*	*No*	*Yes*	*Not Sure*	*No*
Real A / Real A	**90.9**%	9.1%	0%	**69.6**%	17.5%	12.9%	**76.4**%	16.7%	6.9%
Real A / Real B	0%	6.7%	**93.7**%	0%	3.3%	**96.7**%	0%	9.5%	**90.5**%
Real A / Fake A	17.3%	32.7%	**50.0**%	18.5%	31.2%	**50.3**%	4.1%	16.0%	**79.9**%

Table 4.1: Participants' answers when asked if two voice samples were from the same person. We use this to gauge their ability to correctly discern if speech samples were spoken by the same speaker (Real A/Real A), different person (Real A/Real B), or a synthesized (fake) speaker (Real A/Fake A).

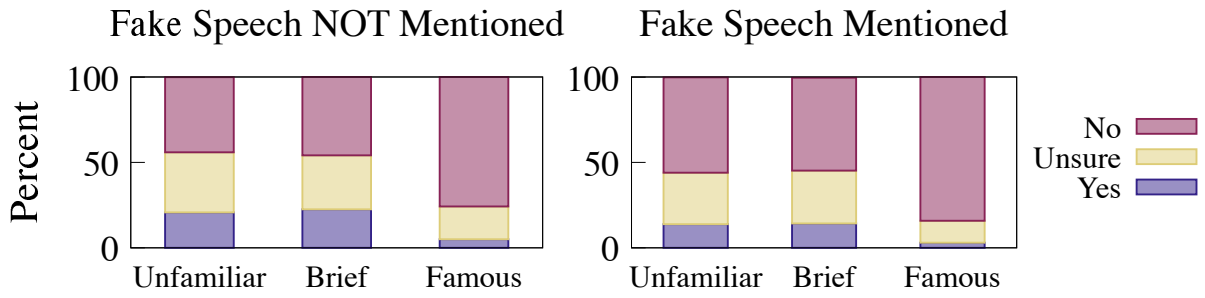


Figure 4.5: User responses to the question "are these two voice samples from the same person?" (Left) when users are not told synthesized speech is used in the survey; (Right) when users are told this.

familiar speakers. For *famous* speakers whose voice participants are generally familiar with, this number drops to 20%.

2) Does hearing more samples from a speaker (i.e., knowing a speaker better) make fake speech more detectable? Results in Table 4.1 suggest that greater familiarity with a speaker will lead to increased skepticism of a fake voice. Compared to a similar user study performed 6 years ago [123], proportion of participants who correctly identified the fake voice for unfamiliar or briefly familiar speakers is consistent (50% for our work vs. 48% in [123]). However, participants in our survey were more accurate at identifying fake speech from famous speakers (80% vs. $\sim$ 50% in [123]), perhaps reflecting a higher general awareness of speech synthesis attacks.

3) Does mentioning fake speech in the survey description change partici-

pants' perceptions of the fake speech samples? Mentioning fake speech in the survey description showed a statistically significant effect on survey responses. Figure 4.5 shows how the responses to the survey version mentioning fake speech reflect an apparent increased skepticism of fake speakers.

Using a chi-squared test for independence, we compared responses from each speaker familiarity category between the two survey versions to see if this change is statistically significant.

- For *unfamiliar* speakers, all but one speaker has a significant ($p < 0.05$) difference in responses.
- For *somewhat familiar* speakers, again all but one speaker has a significant ($p < 0.05$) difference in responses.
- For *famous* speakers, only Trump has a statistically significant difference in responses.

4) Do participant demographics (age, gender) affect responses? Women and younger people were more likely to correctly identify real and fake speakers. Using a chi-squared test for independence, we compared the responses of men vs. women and younger people (age < 25) to older people (age > 45). For *unfamiliar speakers*, there were statistically significant ($p < 0.05$) differences in response between genders and age groups. For somewhat familiar and famous speakers, statistically significant differences are observed for some, but not all, speakers.

4.2.3 User Study B: How Do Users Interact with Synthesized Speech in Trusted Settings?

User study A confirms that DNN-synthesized speech fails to *consistently* fool humans in a survey setting. Looking beyond this, we wonder how, if at all, the *context* in which users were exposed to fake speech influences their susceptibility to these attacks. Specifically, how

would participants act in a setting where they were predisposed *not* to think critically about the voices they hear? Examples of such "trusted settings" include phone or Zoom meetings with colleagues or calls with one or more people they know (or think they know). Human behavior in these so-called "trusted settings" may differ from behavior in a survey-based setting. When humans are primed by their setting to think they are speaking to a real person, they may be more likely to accept fake speech as real.

Study Design. To understand the impact of trusted settings on human interactions with fake speech, we conduct a user study involving deceptive interviews.

Ethics: This study was approved by our institutional IRB. Participants submitted a signed consent form prior to the interview and received a full debriefing afterwards to inform them of the deception and true purpose of the study. No personal information about participants was retained after the interview, and interview recordings were anonymized to protect participant privacy.

Participants. Interviewees were recruited from among the students in our institution's computer science department. We conducted a total of 14 interviews. Twelve interviewees were male (2 female). All were between the ages of 20 and 35 years old, with varied ethnic/racial backgrounds (American, Chinese, Indian, Indonesian, Turkish). The interviews were approximately 10 minutes, and participants were compensated with a $10 Amazon gift card.

Procedure. The recruitment call asked for participation in an interview study about use of speech recognition systems (e.g., Siri) and their perceptions of privacy with respect to these systems. Each interview took place over a Zoom call, with two paper authors functioning as "interviewers." One of the interviewers (hereafter referred to as the *real interviewer*) used their real voice throughout the staged interview, while the other (referred to as the *fake interviewer*) used only fake speech samples based on their real voice. All fake speech samples were generated using the SV2TTS synthesis system and less than 5 minutes of real

voice samples from the fake interviewer. Throughout the call, the fake speech samples are played from an iPad Pro, held close to the fake interviewer's computer microphone.

After the conclusion of the deceptive portion, we revealed the use of fake voice samples and disclosed our research objectives before asking a few additional questions. Participant responses were categorized and coded separately by each interviewer, who later met to combine the codes and resolve discrepancies. All themes in responses described below were expressed by >= 3 participants, unless otherwise noted.

Because all of the interviewees were members of the authors' academic division, they had different levels of familiarity with the interviewers, ranging from general knowledge to frequent social interaction. At the end of each interview, participants were asked to rank their familiarity with the interviewers' voices prior to the interview on a scale of 1 ("not at all familiar") to 5 ("extremely familiar"). Table 4.2 lists the distribution of familiarity rankings.

	Not at all Familiar	*Slightly Familiar*	*Moderately Familiar*	*Very Familiar*	*Extremely Familiar*
Real Interviewer	9	1	2	0	2
Fake Interviewer	7	2	3	1	1

Table 4.2: # of participants and their declared familiarity with the two interviewer's voices before the Zoom interview.

Task. The staged interview itself consists of 8 questions about use of automatic speech recognition systems and perceptions of privacy (see Table 4.3). Five are asked by the real interviewer, and three are asked by the fake interviewer. The three fake interviewer questions are designed to solicit three different types of behavior from participants: conversational response (Q2), website access (Q5), and personal information (Q7).

Conditions. Participants are not told that the study is actually about perceptions of fake speech, and they do not know that one of the interviewers is using a fake voice. When the participant joins the Zoom call, the real interviewer informs them that everyone in the

#	Interviewer	Question
1	Real	Do you use automatic speech recognition systems in everyday life?
2	*Fake*	*How often do you use these systems in your daily life?*
3	Real	What do you do in your interactions with these systems?
4	Real	Do you ever think about your privacy during your interactions with these systems?
5	*Fake*	*Can you visit this website? I'll put the link in the chat.*
6	Real	Have you ever used the "voice profiles" feature of these systems?
7	Real	Are you ever concerned about privacy if/when you use voice profiles?
8	*Fake*	*We need your student id to track your participation in this study. Can you leave it in the chat?*

Table 4.3: Questions asked by real and *fake* interviewers.

call is keeping their video off to preserve interviewee privacy. In reality, keeping videos off prevents the participant from observing that the fake interviewer is using a fake voice. We also asked the interviewees if we could record the interview to maintain a record of their responses.

Because of the relatively low quality of the fake interviewer voice, interviewees are primed to expect a low quality voice from the fake interviewer. For 10 of the 14 participants, before beginning the interview questions, the real interviewer notes that the fake interviewer is feeling unwell and will only chime in intermittently during the interview. We examine the effect of excluding this *priming statement* from the interview in later sections.

Results. *None of the participants exhibited any suspicion or hesitancy during interactions with the fake interviewer's voice.* All 14 responded without hesitation to the three questions asked by the fake interviewer, visited the requested website, and even gave their school ID number to the interviewers. After the interview concluded and the deception was revealed, only four of the 14 participants stated that they thought something was "off" about the fake interviewer's voice. Importantly, these four participants had (intentionally) not been given

the "priming" statement that the fake interviewer "had a cold." Below, we explore the most interesting results from this study and highlight several key limitations.

1) Reaction to fake voice: Several themes arose during the post-deception interviews, as summarized below.

- *Complete surprise:* Four participants were visibly and audibly astonished when the deception was revealed. **P5** noted that, "I really thought it was you – like 100%," while **P10**, after a shocked moment of silence, said "computers just won the Turing test."
- *Satisfied with "sick" excuse:* Seven participants noted explicitly that the "sick" excuse squashed any concerns about the fake interviewer's voice. **P4** said that "I think it totally worked – I thought you were terribly sick," while **P2** noted that "it was really kind of worrying [how sick you sounded]."
- *Silently suspicious:* Four participants (**P9, P12, P13, P14**) expressed suspicions *after* the deception was revealed. **P12** and **P13** said it "sounded like *speaker* had a cold," and **P14** supposed "it was a poor quality microphone."

2) Why participants didn't voice their concerns: After the deception was revealed, participants were asked to identify elements of the interview structure that increased their trust in the fake interviewer. Some, of course, were completely unsuspecting and did not think to question the fake interviewer. However, others noted that the presence of a second (obviously human) interviewer, social convention, and the origin of the interview request (from within our department) bolstered their trust.

- *Presence of real interviewer:* Several participants credited the "tag-team" nature of the interview, with real and fake interviewers colluding, as making the deception more believable. "I feel like [the real interviewer's] obviously human presence played a big factor in [my not saying anything]." (**P9**).

- *Polite social convention:* Multiple participants noted that they felt it would be uncomfortable or wrong for them to say something about the fake interviewer's voice during the interviewer. When asked why they didn't say anything about the quality of the fake interviewer's voice, **P12** exclaimed, "well that would be quite the insult!"
- *Provenance of interview request:* Since we recruited from within our department, the recruitment was sent out through trusted channels only accessible by members of the department (i.e., email list-serv, Slack). **P9** expressed suspicions during the debriefing, but credited the "provenance of the study... seemed like a legit source" as a reason to fully participate with our questions.

3) What would have made participants suspicious: When asked to articulate what would have made them more suspicious, participant responses varied.

- *Nothing:* Participants most surprised by the deception claimed that nothing would have caused them to question the credibility of the fake interviewer: "I'm glad you guys didn't ask me for a bank account, because [...] I would have given it to you" (**P5**).
- *Requesting more personal information:* One participant noted "I don't think the information you wanted was very sensitive [so] I don't see why I need to be concerned about this" (**P6**). IRB constraints prevented us from soliciting anything more personal than a student ID, used to access services at our university. While not public, this information is not inherently sensitive.

4) Effect of familiarity with interviewers: Seven of the participants rated their familiarity with both interviewers' voices as a 1 out of 5 (e.g., not at all familiar with either). Their responses, though, were consistent with the other participants who had some previous familiarity with one or both of the interviewers' voices. Only one participant (**P8**) mentioned that "the voice did seem pretty weird, but since I trust you both, I just went [with] it." These

Category	Defense	Method	Limitations
Liveness Detection	[195]	Measures human vocal tract movement using Doppler radar.	Requires precise static calibration during enrollment/testing.
	[183]	Detects presence of human breath on mic.	Speaker must be <4 inches from mic.
Loudspeaker Detection	[39]	Detects presence of magnetic fields produced by loudspeakers.	Requires careful motion of smartphone during recording.
	[191]	Compares audio environment to previously enrolled speaker environment.	Requires precise static calibration during enrollment/testing.
Artifact Detection	[61, 97, 6, 184, 4, 93]	Trains models to recognize spectral characteristics of synthetic speech.	Only effective when audio directly played from speaker.
	[161]	Measures speech-to-text error rate of speech samples against ground truth.	Requires knowledge of ground truth audio content.
Preventing Synthesis	[77]	Corrupts speech samples to prevent unauthorized speech synthesis.	Degrades quality of defended speech.

Table 4.4: Taxonomy of defenses proposed to prevent speech synthesis attacks.

results suggest that the trusted setting and presence of a human likely play a larger factor than prior familiarity with the speaker's voice.

5) Effect of priming statement: To examine the effect of the "sick" excuse on the believability of the fake voice, we conduct four interviews in which participants are not told that the fake interviewer is sick. In these interviews, participants exhibit an increased level of skepticism about the fake interviewer during the debriefing. One claimed "it was very obviously a fake voice," (**P11**) but said that based on their experience in other deception studies they decided not to say anything. Others did not see through the deception but did note that "I was feeling weird" (**P13**) and "I just feel your voice is very strange" (**P14**).

4.2.4 Key Takeaways

Our two user studies (A & B) show that context and demographics impact the credability of synthesized speech for human users. In study A, we found that mentioning fake speech increased participants' skepticism of the fake speakers they heard. Additionally, women and younger participants in study A were more likely to correctly identify fake speakers.

Our key takeaway from study B is that *a fake voice fooled humans in a trusted interview setting.* Of particular interest is that all our study B participants were graduate students in computer science, some of whom actively research security or machine learning. Our starting hypothesis was that computer science graduate students would be among the hardest targets

to fool with a fake voice. Yet, none of them expressed suspicion about the fake voice during the interview.

Limitations & Next Steps. Our participant pool for study B was largely homogenous in gender, age, and educational background. To conduct a "trusted" interview, our participants were drawn from our academic department (computer science). The gender breakdown of our participants matches that of the department, which skews heavily male. It is possible that the observed effect of gender and age on responses in study A could also extend to study B. Therefore, a viable follow-up work is to conduct larger, more-diverse user studies to provide a more nuanced understanding of synthesized voice attacks in trusted settings.

On a related note, our trusted interview in study B followed a voice-only format, where voice is the only medium for interaction. Yet in real-world scenarios, interviewees could use two-factor authentication mechanisms to verify the trusted setting, e.g., requesting the interviewers to turn on their video feed, or challenging the interviewers with some verbal tests. These combined verification methods could make the attacks much more difficult, allowing human users to effectively defend against speech synthesis attacks. We believe this is an important direction for follow-up work.

4.2.5 Discussion

This work represents a significant step towards understanding the real-world threat of deep learning-based speech synthesis attacks. The results demonstrate that synthetic speech generated using publicly available systems can already fool both humans and today's popular software systems, and that existing defenses fall short. As such, both humans and machines require new defenses against speech synthesis attacks and further research efforts for exploring subsequent challenges and opportunities, while providing a solid benchmark for future research.

4.3 Enabling Personalized Protection against Unacceptable Face Editing

Today, face editing is widely used to refine/alter photos in both professional and recreational settings. Yet it is also used to modify (and repost) existing online photos for cyberbullying. Our work considers an important problem: "*How can we support the collaborative use of face editing on social platforms while protecting against unacceptable edits and reposts by others*?" This is challenging because, as our user study shows, users vary widely in their definition of what edits are (un)acceptable. Any global filter policy deployed by social platforms is unlikely to address the needs of all users, but hinders social interactions enabled by photo editing.

Instead, we argue that face edit protection should be implemented by social platforms based on individual user preferences. When posting an original photo online, a user can choose to specify the types of face edits (dis)allowed on the photo. Social platforms use these per-photo policies to moderate future photo uploads, i.e., edited photos containing modifications that violate the original photo's policy are either blocked or shelved for user approval. Realizing this personalized protection, however, faces two immediate challenges: (1) how to accurately recognize specific modifications, if any, contained in a photo; and (2) how to associate an edited photo with its original photo (and thus the edit policy). We show that these challenges can be addressed by combining highly efficient hashing based image search and scalable semantic image comparison, and build a prototype protector (*Aletheia*) covering nine edit types. Evaluations using user studies and data-driven experiments (on 839K face photos) show that *Aletheia* accurately recognizes edited photos that violate user policies and induces a feeling of protection to study participants. This demonstrates the initial feasibility of personalized face edit protection.

Our Contributions. This work targets the critical challenge of user-specified modera-

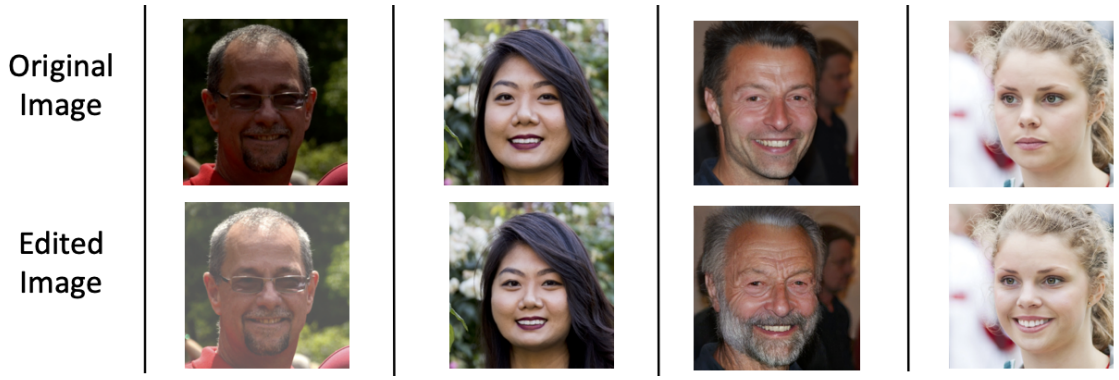


Figure 4.6: *Examples of face edits done by today's low-cost or free edit tools (Photoshop, PortraitPro, FaceApp).*

tion of how others make face edits on our online photos and repost them. This work makes three contributions:

(1) a user study to explore user tolerance of face edits on their photos when done by others. Our results show significant variance across users and edit types, motivating the need for *personalized* face edit protection;

(2) *Aletheia*, a prototype moderation tool for photo sharing sites to implement user-specific face edit protection on their photo posts. We address the key challenge of recognizing the type of edits contained in a photo x by combining highly efficient hash based image search that locates x's original, unedited version, and scalable semantic image comparison between x and its unedited version. This "reference-based" methodology differs from existing solutions for detecting deepfakes, which assume the absence of an original image.

(3) User studies & data experiments on 839K photos evaluate *Aletheia* for protection effectiveness, scalability, and users' perceptions of *Aletheia*'s protection on their online photos. Results show i) *Aletheia* successfully identified 93.8% of edited photos marked as unacceptable by user study participants; ii) *Aletheia* operates at scale with high accuracy and low latency, e.g., $> 97\%$ accuracy and $<1s$ latency per photo in detecting edited images, while existing works offer only $9.5-55.6\%$; and iii) study participants had generally positive views of protection provided by *Aletheia*. Altogether these results suggest that our approach could

Category	Example face edit types
Global retouch	change photo brightness; add filter effect
Insert sticker	add sunglasses/emoji
Change facial attributes	increase/decrease age, change gender appearance, add/remove hair, change face shape; add makeup
Change expression	non-smile → smile, smile → crying
Change identity	swap two faces

Table 4.5: *The face edit types considered by our study.*

be an effective method for protecting online face photos from being improperly edited and reposted.

4.3.1 Understanding How Users Perceive Face Edits Done by Others

Despite existing studies on self photo editing, deepfakes, and photo privacy, there is little work on understanding users' perspectives and reactions on face edits that others have applied to their online photos. To answer this question, we conducted an online survey about users' tolerance for others editing their selfies and perceptions of privacy when posting photos online. More details in appendix D.1.

Participants. We recruited 100 participants via the crowdsource platform Prolific. Participants were required to be 18+ years old, live in the US, and have 95% approval rating on Prolific. The survey was designed to take 15 minutes on average, and participants received $3 as compensation. We collected 99 valid responses (one participant timed out), among which 53 identify as male (46 female). The age distribution is 18-29 years (60%), 30-39 (22%), 40-49 (16%) and 50-59 (2%).

Task. We first presented the concept of face editing to participants, and asked whether they have observed face edits done by others in online images/videos (not necessarily their own). We then asked participants to suppose they shared a photo of their faces online to

a platform similar to Facebook or Instagram, and asked a series of multiple choice and free response questions about their perceptions and opinions regarding others editing the posted photo and reposting it (e.g., what edits can/cannot be tolerated), and their preferences for how the platform should act regarding these edited versions. For our study, we categorized common types of face edits (offered by today's tools) into five groups by their effects [44], from which we produced 15 edit types (see Table 4.5) used for our study.

The goal of our user study is not to develop (or apply) a method to precisely collect a per-photo edit policy from participants, but to explore the pattern and diversity, if any, in participants' responses to others applying face edits to their photos.

Conditions. We presented two scenarios (in random order): the edited photo is viewable to *friends and family only* (similar to Instagram's "close friends" option) or viewable to the *public*. For each scenario, we surveyed participants in two steps. In step 1, we described each edit type and then illustrated its high-level effects using example photos (we explain the photo choices below). We then asked each participant to imagine such edit type (with varying spectrum and style) is applied to their own photos by another person, and rate how likely they would allow the edited photo. The default rating is 5-point Likert scale, i.e., never (1), rarely (2), sometimes (3), usually (4), always allow (5). For edits (e.g., age, brightness, face shape) that can be measured on a spectrum (e.g., increase/decrease), we also presented examples of five edit levels (0%, 50%, 100%, 150%, no limit) on both increase and decrease, and asked participants to which extent they would allow the edit. Next in step 2, we presented several new edited images and asked participants to select the ones they would allow. To verify responses, we included attention check questions and applied both time check and manual inspection to detect straight-lining and false input. Appendix D.1 lists examples of survey images.

Photo Choices. To precisely collect a participant's opinion on face edits, one could present them samples of edited photos of themselves. However, seeing certain edits on their own photos could lead to negative emotional effects that cannot be predicted before the study [89, 102]. Also the remote/one-direction nature of our user study meant we could not debrief our remote participants. Thus to minimize potential harm, we did not collect or alter personal photos from our remote participants. Instead, we showed participants sample photos of other people, before-and-after edits, to help illustrate possible effects of different edit types. We asked participants to visualize edits applied to photos of their own faces when answering the study questions. In our opinion, doing so achieves the desired goal of impressing the impact of different face edit types to individual participants while minimizing any potential negative emotional effects on them.

Face edits by others are commonly observed in today's online platforms. Most participants (75%) reported having observed face edits done by others in online shared images/videos. When asked about how frequent they observe such editing, 31.3% reported 'Somewhat often (a few times a week)' and 12.1% reported 'Very often (at least once a day)'. Also, 19.2% indicated that they themselves have edited other people's face images/videos.

Users vary significantly in tolerance for different types of face edits. Participants showed significant variation in their tolerance of others editing their online face photos. This can be observed from the raw scores on the edit tolerance (rated on a scale of 1-5) in Figure 4.7, where we show the scores of all 99 participants for each of the 15 edit types. Here the color represents the raw score, white = 5 (always allow) and black =1 (never allow). On average, participants would allow half of the editing types presented, with 8 participants (8%) allowing *all* types of edits (i.e., a score of 5 for all 15 types) and 3 participants (3%) allowing *none* for either scenario (i.e., a flat score of 1). We also measured the level of variation as the standard deviation (std) across edit types and participants. For each edit

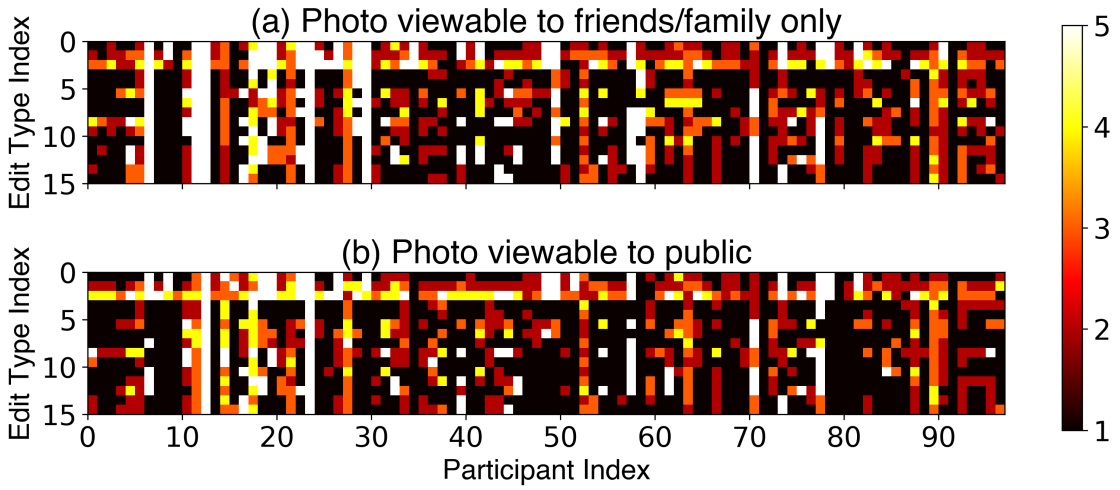


Figure 4.7: The raw score distribution across our study participants (99 users), who provided a score (1-5) for each of the 15 edit types. 5 = always allow, 1 = never allow.

.

type, the std across participants is high and comparable to the mean (std $\in [1.17, 1.58]$, mean $\in [1.62, 3.4]$). For participants allowing some edit types, the std across edit types is similarly high (std $\in [0.25, 2.0]$, mean $\in [1.07, 4.0]$), suggesting that their choices of the edits are highly personalized.

To explore the impact of context (i.e., photo viewable to public or friends/family only), we computed the difference between the mean tolerance of two scenarios per participant. Again the results vary across participants: 52.6% showed indifference, 20.6% would allow more edits for public view, and 26.8% would allow more edits for friends/family view.

To understand the reasoning behind their individual selections, after evaluating both scenarios, we asked participants to explain in their own words. As shown in Table 4.6, the reasons expressed fall into 5 general categories: prefer no edits ever, prefer only specific edits (regardless of the audience), would allow more edits among friends/family, would allow more edits for public photos, or general indifference. These responses also indicated that who the editors are is also an important factor. Overall, we can clearly observe that users differ largely in their tolerance of face edits.

Reason	Example
None/ Very Few Edits	"I **wouldn't want anybody editing my photos, whether I know them or bot** [*sic*]. It feels intrusive."
Specific Edits Only	"It doesn't matter to me who can see it, I **just don't want specific edits** done to me." "Sometimes **some edits end up making the pictures weird**"
Allow More Edits among Friends/ Family	"Well I would find it **funny if my friends did some of those edits** to me **but I would be a bit annoyed if a random person** did some of those to my photo."
Allow More Edits for Public View	"would have **more fun doing more extreme edits for public viewing** because potentially more people will be seeing them rather than friends and family who already know what I look like."
General Indifference	"I feel like **whatever you show in private for the most part you should be able to show in public**"

Table 4.6: *Participant reasons for their edit preferences.*

Many users prefer aggressive identification and notification of policy violations. Our study showed that participants preferred a more aggressive approach to detecting unacceptable face edits. Two-thirds of participants felt the platform should **flag as many potentially edited images** as possible, even at the cost of some false positives. When the system detects an edited image that violates user preferences, 87% of participants wanted proactive notifications. Finally, 60% of participants expressed concern about the development of new face-editing methods, and the need to adjust their preferences accordingly over time.

The need for personalized face edit protection. Overall, our user study shows that users are heavily concerned about others editing and reposting their face photos and want the ability to protect their online photos; but since users hold very different definitions of

what face edits are unacceptable, the protection against face edits must be *personalized*.

4.3.2 Aletheia

To address the unfulfilled need for personalized face edit protection, we propose and design Aletheia to address this gap. *Aletheia* is an image moderator system to protect original face images on photo sharing services.

Usage Scenarios. Here we make two assumptions:

- *Aletheia* focuses on selfie photos (front-shot of a single face), which are the main target of malicious face editing.
- We design *Aletheia* to protect a user's face photos after they are posted online. To receive protection, an original photo must be registered into *Aletheia* before its edited versions. Specifically, when a user posts an original photo into an online service employing *Aletheia*, the photo is verified by *Aletheia* as an original and then registered into the system. A user can fill a claim with *Aletheia* if their original images are registered by someone else, and prove ownership by verification via face recognition or camera-generated stamps [2].

Threat Model. We are motivated by the need to prevent the use of face edit tools for cyberbullying, and design *Aletheia* to resist "standard manipulators" who are familiar with everyday technology (i.e., those who can use commodity tools to modify photos, and delete/modify a photo's metadata), but not security experts or strongly motivated adversaries (i.e., resourceful attackers who analyze *Aletheia*'s internal design and craft adversarial attacks to bypass *Aletheia*'s detection).

Design. Different from existing efforts, we design *Aletheia*to effectively detect if and how an image has been edited, by applying *reference-based* face edit detection and recognition.

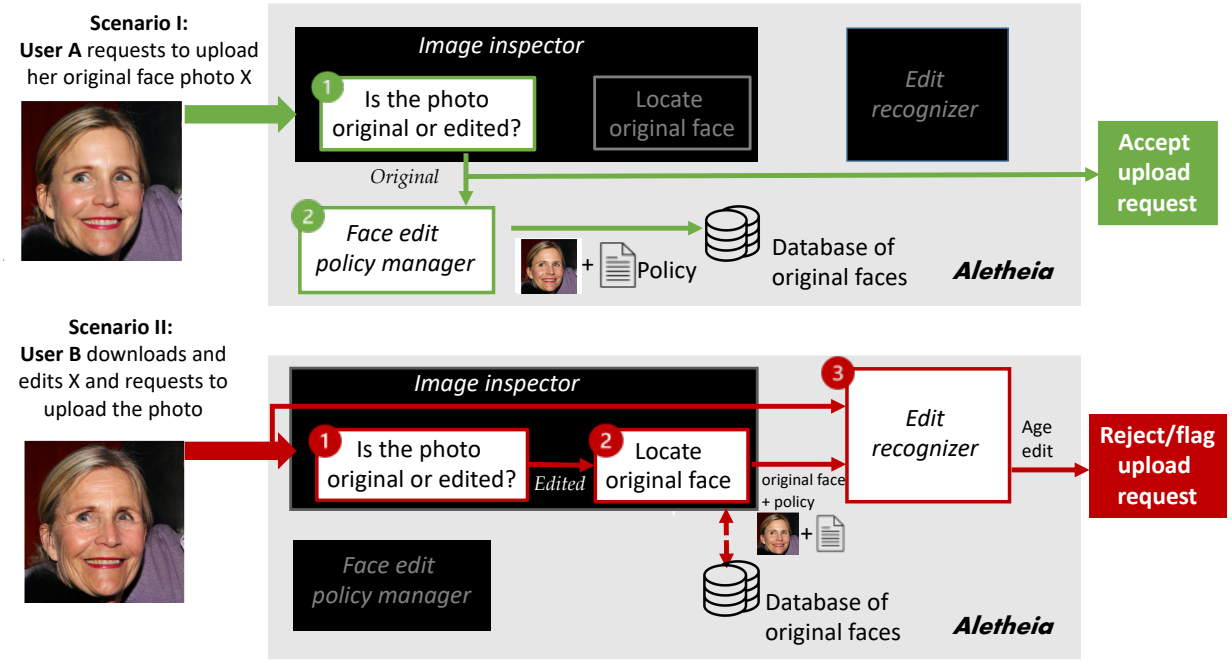


Figure 4.8: *Overview of* Aletheia*'s operation when users request to upload original (scenario I) and edited photos that contain unacceptable edits (scenario II).*

Aletheia consists of four components: **(1) a face edit policy manager** that allows each user, when uploading an original face photo, to specify their policy that defines unacceptable face edits and the subsequent system action upon detecting such edits; **(2) an image inspector** that for each incoming image x, inspects the image to determine whether it is an original image; if so, the inspector asks the user to input policy, and if not, it locates x's original version x_0; and **(3) an edit recognizer** that compares x and x_0 to determine whether x contains any unacceptable edits defined by x_0. In addition, *Aletheia* maintains an internal **(4) database** to store registered original face photos and their edit policy.

Figure 4.8 illustrates *Aletheia*'s operation pipeline for two scenarios. In scenario I, the input image to *Aletheia* is an original face photo. The image inspector verifies the input is original, prompting the user to define an edit policy on this photo via the policy manager. It then registers the photo (and policy) into the database, and accepts the upload request. In scenario II, the input image is an age-edited face photo. The image inspector first identifies

the input as an edited photo and proceeds to locate the original face photo (and edit policy) in the database. Then the edit recognizer compares both photos to identify edits, and uses the edit policy to determine existence of any unacceptable edits. If so, the upload request is either rejected or flagged for user review (per user's policy). If not, the upload request is accepted. In the example of Figure 4.8, the image violates the user policy that disallows age edit.

4.3.3 Aletheia*'s Decision vs. Human Decision*

Using data from the user study in §4.3.1, we examined how *Aletheia* flags edited images that violate user policies, and whether such decisions match human decisions. For each participant, we used their responses to (1) define an edit policy per edit type, i.e., acceptable or unacceptable, and (2) obtain a set of human decisions on the edited images, which we use to evaluate *Aletheia*. The policy was generated from user data collected in step 1, and decision data from step 2. In total, we have 99 participants and 406 valid human decisions (156 unacceptable, 250 acceptable). Next, for each edited image labeled by humans, we ran *Aletheia* based on each participant's policy to determine whether *Aletheia* accurately detects those violating the policies. Our experiment produced two key findings.

***Aletheia* can accurately flag edited images that users disallow (93.6%).** We found that *Aletheia*'s decisions match the participants' decisions well. It successfully flagged 93.6% of edited images (146 out of 156) that participants labeled as unacceptable, and accepted 87.6% of images (219 out of 250) that participants labeled as acceptable.

Decision mismatch came from subtle edits and overlap of edits. We studied mismatch between *Aletheia* and participants' decisions, and found two dominating trends. First, the "unacceptable" images not detected by *Aletheia all* came down to a single skin tone edited image, which contains very subtle change of skin tone. *Aletheia* failed to spot

the change because it uses a common human skin tone palette that "ignores" such subtle changes. Second, when *Aletheia* falsely flagged an acceptable image as unacceptable, the error came from overlap of edits. For example, an age edit often changes hair color, and face swap often changes expression, age, and face shape. When a participant's policy contains conflict across overlapped edits, e.g., allowing age edit but not hair color edit, those false alarms are inevitable.

Insight: the need for precise edit policy. Our results are encouraging and demonstrate an initial feasibility of *Aletheia*. They also confirm the observation that the current definition of edit types is likely too broad to build accurate edit recognizers. *Aletheia* could largely benefit from more precise characterization and interpretation of edit types, so users can clearly define fine-grained policies that are free of conflicts and can be implemented as decision rules.

4.3.4 User Perception of Aletheia*'s Protection*

We conducted an additional online survey to assess how users perceive the protection offered by *Aletheia*, and to submit, if any, suggestions on improving *Aletheia*. More details in appendix D.2.

Participants. We recruited 100 participants via Prolific. The survey was designed to take 10 minutes on average and participants received $2 as compensation. We received 97 valid responses (3 responses failed attention check question). Of those, 7 indicated they did not feel concerned at all about privacy online. Since those privacy-insensitive users are not our target users, we filtered their responses from our analysis. In the end, we analyzed 90 responses (44% identified as female, 66% male). The age distribution is: 18-29 years old (75%), 30-39 (16%), 40-49 (7%) and 50-59 (2%).

Task. We asked participants to imagine using *Aletheia* when posting an image online to a site like Instagram. We first show examples of each edit type, and demonstrate how the system would enforce potential policies when an unacceptably edited image is detected. We then asked multiple-choice and free response questions about the usability of the system, users' sense of protection, and their perceptions of privacy when posting images online.

This *conceptual approximation* of the *Aletheia* system captures its essence and demonstrates the potential value of the service. We used it to help our study participants understand the protection offered by *Aletheia* and determine whether they would want or need such protection. Also, since the remote/one-direction nature of our user study meant we could not debrief our remote participants, we chose to not collect or alter personal photos from our remote participants, in order to protect their privacy and minimize potential negative emotional effects.

Many participants showed appreciation for the protection offered by *Aletheia*. We asked participants how they felt about the protection *Aletheia* would provide for their online photos. Table 4.7 shows a summary of the responses. Nearly half (48%) of the participants felt that *Aletheia* protected their images, especially since they can define personalized protection policy. 15% of the participants were neutral. They questioned the full effectiveness of the protection, but still viewed *Aletheia* as a step in the right direction. 13.3% of the participants did not feel protected by *Aletheia* because they worried that the system could be bypassed, such as posting edited images elsewhere online, or were not convinced *Aletheia*'s technology could accurately detect most edits. 23.7% of the participants expressed that posting images online is never safe and the only way of protection is not uploading any.

Many participants would like to use *Aletheia*. Regarding whether they would use *Aletheia* to protect their online images, we observed considerable differences between *edit-concerned* and *edit-unconcerned* participants (see Table 4.8). Note that at the beginning of

Response	Reason	Example
Protected (48%)	Trust system works to detect disallowed edits	"I would feel my images **are protected by the system as I can specify** whether I would like them to be modifed [*sic*] in a way I would not like."
Neutral (15%)	Can't 100% guarantee protection	"**it may miss when a photo has been edited**" "i think they protect the images **to a certain extent however not fully**"
Not Protected (13.3%)	The system can be bypassed (8.5%)	"I think it **could be cheated** easilly [*sic*]" "Pictures can still be extracted and **posted somewhere else**."
	Don't trust system (4.8%)	"I **don't think the system is advanced enought** [*sic*] to detect these images."
Never (23.7%)	Posting images online is never safe	"I think it's **never safe when we post pictures of ourselves** because they never really leave the internet."

Table 4.7: *Participant responses for whether they feel their images were protected with* Aletheia*.*

User group	Yes	Neutral	No
edit-concerned	68%	21%	11%
edit-unconcerned	42%	27%	27%

Table 4.8: *Participant responses for whether they would use* Aletheia*when posting images on social media sites.*

the user survey, we asked each participant whether they are concerned about their image being edited and reposted by others, and the result was a near-even split (49%/51%) across participants. From Table 4.8, we see that 68% of *edit-concerned* participants were interested in using *Aletheia* and 21% were neutral. Of the 11% (5 participants) who said no, 4 expressed that they **never** shared images on social platforms and did not feel protected even with *Aletheia*. Another interesting observation is that even among those not concerned with edit, 42% indicate they would use *Aletheia*.

Overall, our survey results are highly encouraging, showing that most participants express interest in using *Aletheia* to increase protection online. More efforts like *Aletheia* should be made to provide more privacy-friendly services and to educate users on ways to achieve their privacy goals.

Participants want to configure and adapt their edit policy, despite the overhead. We asked how users consider the tradeoff between time spent setting up their own policy,

and achieving protection. Most participants were either not concerned (45%) or neutral (32%), deeming the protection worth the initial setup time. The rest 23% expressed concern about the time spent, with one participant feeling this may leave many users reverting to default settings. Also, 75% of participants indicated they would prefer a single policy for all images, for simplicity and efficiency. Additionally, similar to the first study, we found that participants want flexibility to change their preferences over time, and expect the system to adapt and new editing methods are developed. Together, this feedback suggests that the design of *Aletheia*'s edit policy management should serve to spare the users' efforts, whilst affording personalized control.

Suggestions on improving *Aletheia*. We asked participants what changes, if any, they would make to improve *Aletheia*. While most participants did not submit any response, there are a few notable ones. Four participants indicated they would like notifications for *any* edits detected, so they could decide whether to remove them. Eleven participants care about *who* makes the edit, such as "set certain friends to have edit permissions" or "allow users to ask for permissions to the original author of the image." Finally, several participants brought up a desire to implement *Aletheia* on all possible platforms, providing ultimate protection against any edited face images posted online.

4.3.5 Discussion

Limitations and Future Work. As the first work on face edit protection for online photos, *Aletheia* faces a number of limitations, much of which will be the targets of future work in this space.

(1) Deeper study on users' tolerance for face edits: Our user study is limited in that we collected tolerance of different face edits when participants evaluated others' face photos (to protect our participants). This tolerance may change when participants evaluate their

own individual photos, which needs to be considered when collecting the specific edit policy from a user seeking protection.

(2) Edit policy definition and management: Our current edit policy specification adopts a simple (default) policy on several common types of face edits. We recognize three broad challenges in clearly defining and deploying face edit policies.

- Current tools and literature define broad and vague "types" of face-edits, and many edits are naturally correlated. These have affected the accuracy of *Aletheia*'s edit recognition. We need a systematic approach to interpret and decompose face edit types, and an interactive interface to guide users in defining usable policy. Here a related question is how to effectively illustrate the edit effect to users while minimizing/addressing potential negative emotional impacts.
- Defining certain edit types such as gender appearance and age may rely on common stereotypes that fail to properly capture real world diversity. Much work remains in developing a more nuanced and powerful policy specification that better reflects user diversity.
- The third challenge is how to automate policy configuration. One can explore the use of machine learning tools to learn users' preferences, and help them set their edit policy automatically.

(3) Expanding edit recognizer: So far our prototype employs nine attributor extractors built from public models. We plan to add new ones to cover a broader range of edits, leveraging ongoing efforts on semantic face analysis. This effort needs to be integrated with the policy component to meet the needs of real-world users.

(4) Integration with multiple photo-sharing platforms: So far *Aletheia* targets a single photo-sharing platform. While this can be effective to protect users if deployed by a very large platform like Instagram, we could achieve much more impact if multiple platforms collaborate. Thus a natural extension to this work would consider privacy-preserving ways

to share personalized user policies and data across platforms, so that unacceptable edits of images from one platform can be detected on others.

(5) Addressing detection errors: Like any practical system, *Aletheia* may occasionally make mistakes. Here we discuss two main types of errors and ways to mitigate them. The first type is wrongly recognizing a new image as an edited one and forwarding it to the wrong owner to review. One way to reduce the likelihood of such errors is to add a verification step to check whether the face identities of the edited image and its original copy match, i.e., the two images are photos of the same person. When the two images display different identities, it could be a detection error or caused by a "faceswap" edit. Such cases could be reviewed by the platform's moderator before taking further actions.

The second type of errors is wrongly identifying an edited image as original, or failing to detect the disallowed edits, so the image is posted online. A user affected by this type of error can mark the photo and submit a complaint. *Aletheia* can verify the complaint and remove the image post if necessary. In addition, *Aletheia* can use this data point to diagnose and improve its detection algorithms. Thus real-world deployments of *Aletheia* need to include a mechanism for users to report errors.

(6) Verifying photo ownership: *Aletheia* protects each original photo based on its edit policy. Intuitively, the legal owner(s) of an original photo should be the one who defines the policy. This leads to the issue of how to define the legal owner(s) of a photo [108], e.g., the person who took the photo, or the person who owns the copyright to the photo. This ownership issue should be addressed by each photo-sharing platform before deploying *Aletheia*, e.g., via their term-of-service or copyright agreement.

Our work seeks to address the threat of online face photos getting edited and reposted by others for malicious purposes. Our user study shows that users are concerned about this threat and want actions taken to protect their online photos. But realizing such protection is challenging because users vary widely in their definition of what edits are (un)acceptable.

This motivates us to develop an image moderation tool that online platforms can deploy to provide personalized protection against unacceptable face edits. In this work, we design and prototype *Aletheia* to address two immediate challenges of personalized face edit protection: detecting and recognizing individual edits on a photo and also identifying its original version (and thus its edit policy). Overall, our work demonstrates the initial feasibility for online platforms to support social interactions via photo editing and sharing, while giving users agency over how their photos can be altered by others.

4.4 Protecting Artists from Style Mimicry by Text-to-Image Models

Recent text-to-image diffusion models such as MidJourney and Stable Diffusion threaten to displace many in the professional artist community. In particular, models can learn to mimic the artistic style of specific artists after "fine-tuning" on samples of their art.

Beyond open questions of copyrights [27], ethics [136, 59], and consent [49, 68, 57], it is clear that these AI models have had significant negative impacts on independent artists. For the estimated hundreds of thousands of independent artists across the globe, most work on commissions, and attract customers by advertising and promoting samples of their artwork online. A model that mimics this style profits from that training without compensating the artist, effectively ending their ability to earn a living. Also, as synthetic art mimicry continues to grow for popular artists, they displace original art in search results, further disrupting the artist's ability to advertise and promote work to potential customers [153, 73]. Today, all of these consequences have indeed occurred in the span of a few months.

This work presents *Glaze*, a system that allows an artist to apply carefully computed perturbations to their art, such that diffusion models will learn significantly altered versions of their style, and be ineffective in future attempts at style mimicry. We worked closely with members of the professional artist community to develop *Glaze*, and conduct multiple

Figure 4.9: Sample AI-generated art pieces from the Midjourney community showcase [112, 146].

user studies with over 1,000 participants from the artist community to evaluate its efficacy, usability, and robustness against a variety of active countermeasures.

Intuitively, *Glaze* works by taking a piece of artwork, and computing a minimal perturbation (a "style cloak") which, when applied, shifts the artwork's representation in the generator model's feature space towards a chosen target art style. Training on multiple cloaked images teaches the generator model to shift the artistic style it associates with the artist, leading to mimicry art that fails to match the artist's true style.

This work makes several key contributions:

- We engage with top professional artists and the broader community, and conduct user studies to understand their views and concerns towards AI art and the impact on their careers and community.
- We propose *Glaze*, a system that protects artists from style mimicry by adding minimal perturbations to their artwork to mislead AI models to generate art different from the

targeted artist. 92% of surveyed artists find the perturbations small enough not to disrupt the value of their art.

- Surveyed artists find that *Glaze* successfully disrupts style mimicry by AI models on protected artwork. 93% of artists rate the protection is successful under a variety of settings, including tests against real-world mimicry platforms.
- In challenging scenarios where an artist has already posted significant artworks online, we show *Glaze* protection remains high. 87.2% of surveyed artists rate the protection as successful when an artist is only able to cloak 1/4 of their online art (75% of art is uncloaked).
- We evaluate *Glaze* and show that it is robust (protection success $> 85\%$) to a variety of adaptive countermeasures.
- We discuss *Glaze* deployment and post-deployment experiences, including countermeasures in the wild.

4.4.1 Collaborating with Artists

Our goal is to help artists disrupt AI models trying to mimic their artistic style, without adversely impacting their own artwork. Because "success" in this context is highly subjective ("Did this AI-art successfully mimic Karla's painting style?"), we believe the only reliable evaluation metric is direct feedback by professional artists themselves. Therefore, wherever possible, the evaluation of *Glaze* is done via detailed user studies engaging members of the professional artist community, augmented by an empirical score we develop based on genre prediction using CLIP models.

We deployed two user studies during the course of this project (see Table 4.9). Both are IRB-approved by our institution. Both draw participants from professional artists informed via their social circles and professional networks. The first (Survey 1, §4.4.1, §4.4.3), asked participants about their broad views of AI style mimicry, and then presented them with a

Survey	# of artists	Content
Survey 1	1156	1) Broad views of AI art and style mimicry(§4.4.1) 2) Glaze's usability, i.e. acceptable levels of cloaking (§4.4.3) 3) Glaze performance in disrupting style mimicry (§4.4.3)
Survey 2 (Extension to Survey 1)	151	1) Additional performance tests (§4.4.3) 2) Robustness to advanced scenarios (§**??**) and countermeasures (§**??**) 3) Additional system evaluation (Appendix **??**)

Table 4.9: Information on our user studies: the number of artist participants and where we report the results of the studies. We sent Survey 2 to some specific participants from survey 1 who volunteered to participate in a followup study.

number of inputs and outputs of our tool, and asked them to give ratings corresponding to key metrics we wanted to evaluate. We select a subset of participants from the first study to participate in a longer and more in-depth study (Survey 2) where they were asked to evaluate the performance of *Glaze* in additional settings.

Artists' Opinions on Style Mimicry While we expected artists to view style mimicry negatively, we wanted to better understand how much individual artists understood this topic and how many perceived it as a threat. Here we describe results from Survey 1 to gather perceptions of the potential impact of AI art on existing artists.

Survey Design. Our survey consisted of both multiple choice and free response questions to understand how well people understand the concept of AI art, and how well the models successfully imitate the style of artists. Additionally, we asked artists about the extent to which they anticipate the emergence of AI art to impact their artistic activities, such as posting their art online and their job security. A handful of professional artists helped disseminate our survey to their respective artist community groups. Overall, we collected responses from 1,207 participants, consisting primarily of professional artists (both full-time (46%) and part-time/freelancer (50%)) and some non-artist members of the art community who felt invested in the impact of AI art (4%). Of the participants who consider themselves artists, their experience varied: <1 year (13%), 1-5 years (49%), 5-10 years (19%), 10+

years (19%). Participants' primary art style varied widely, including: animation, concept art, abstract, anime, game art, digital 2D/3D, illustration, character artwork, storyboarding, traditional painting/drawing, graphic design, and others.

Key Results. Our study found that 91% of the artists have read about AI art extensively, and either know of or worry about their art being used to train the models. Artists expect AI mimicry to have a significant impact on artist community: 97% artists state it will decrease some artists' job security; 88% artists state it will discourage new students from studying art; and 70% artists state it will diminish creativity. "Junior positions will become extinct," as stated by one participant.

Many artists ($>$ 89% artists) have already or plan to take actions because of AI mimicry. Over 95% of artists post their artwork online. Out of these artists, 53% of them anticipate reducing or removing their online artwork, if they haven't already. Out of these artists, 55% of them believe reducing their online presence will significantly impact their careers. One participant stated "AI art has unmotivated myself from uploading more art and made me think about all the years I spent learning art." 78% of artists anticipate AI mimicry would impact their job security, and this percentage increases to 94% for the job security of newer artists. Further, 24% of artists believe AI art has *already* impacted their job security, and an additional 53% expect to be affected within the next 3 years. Over 51% of artists expressed interest in proactive measures, such as personally joining class action lawsuits against AI companies.

Professional artists thought AI mimicry was very successful at mimicking the style of specific artists. We showed the artists examples of original artwork from 23 artists, and the artwork generated by a model attempting to mimic their styles. 77% of artists found the AI model *successfully* or *very successfully* mimic the styles of victim artists, with one stating "it's shocking how well AI can mimic the original artwork." Additionally, 19% of participants thought the AI mimicry is somewhat successful, leaving only $<$ 5% of artists

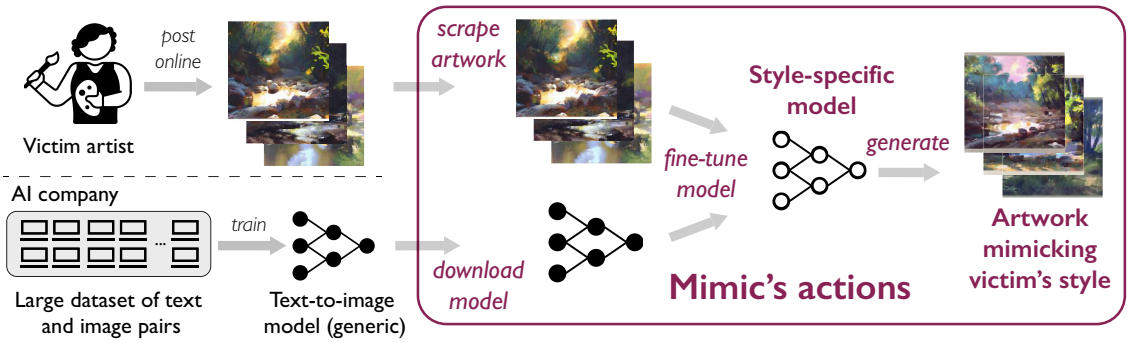


Figure 4.10: Overview of *Glaze*, a system that protects victim artists from AI style mimicry by cloaking their online artwork. (**Top**) An artist V applies the cloaking algorithm (uses a feature extractor Φ and a target style T) to generate cloaked versions of V's art pieces. Each cloak is a small perturbation unnoticeable to human eye. (**Bottom**) A mimic scrapes the cloaked art pieces from online and uses them to fine-tune a model to mimic V's style. When prompted to generate artwork in the style of V, mimic's model will generate artwork in the target style T, rather than V's true style.

rating the mimicry as unsuccessful. Several artists also pointed out that, as artists, upon close inspection they could spot differences between the AI art and originals, but were skeptical the general public would notice them.

A significant concern of most participants, surprisingly, is not just the existence of AI art, but rather scraping of existing artworks without permission or compensation. As one participant stated: "If artists are paid to have their pieces be used and asked permission, and if people had to pay to use that AI software with those pieces in it, I would have no problem." However, without consent to use their artwork to train the models, "it's incredibly disrespectful to the artist to have their work 'eaten' by a machine [after] many years to grow our skills and develop our styles."

4.4.2 Disrupting Style Mimicry with Glaze

We propose *Glaze*, a tool that protects artists against AI style mimicry. An artist uses *Glaze* to add small digital perturbations ("cloak") to images of their own art before sharing online (Figure 4.10). A text-to-image model that trains on cloaked images of artwork will learn an incorrect representation of the artist's style in feature space i.e., the model's internal

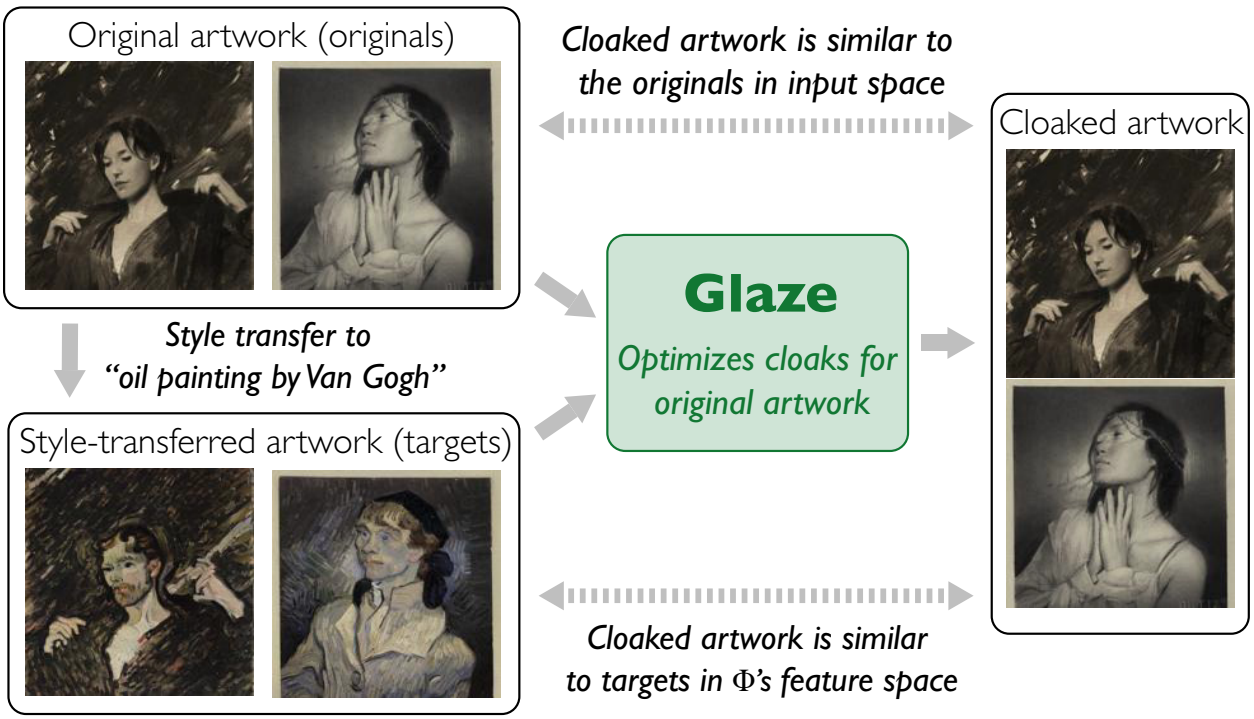


Figure 4.11: High level overview of how *Glaze* perturbs the style-specific features of the artwork. **a)** *Glaze* style transfers the original artwork to a different style, which changes its style but leaves other features unaltered. **b)** *Glaze* optimizes a cloak that makes the artwork's features representation match that of the style-transferred art, while constraining the amount of visible changes to the artwork.

understanding of artistic styles. When asked to generate art pieces in victim's style, the model will fail to mimic the style of the victim, and instead output art pieces in a recognizably different style.

Our key intuition is to identify and isolate *style-specific features* of an artist's original artwork, i.e., the set of image features that correspond to artistic style. Then *Glaze* computes cloaks while focusing the perturbation budget on these style-specific features to maximize impact on stylistic features.

As discussed, identifying and calculating style-specific features in model's feature space is difficult due to the poor interpretability of model features and how art style manifests differently across artworks. We overcome these two challenges by designing a style-dependent

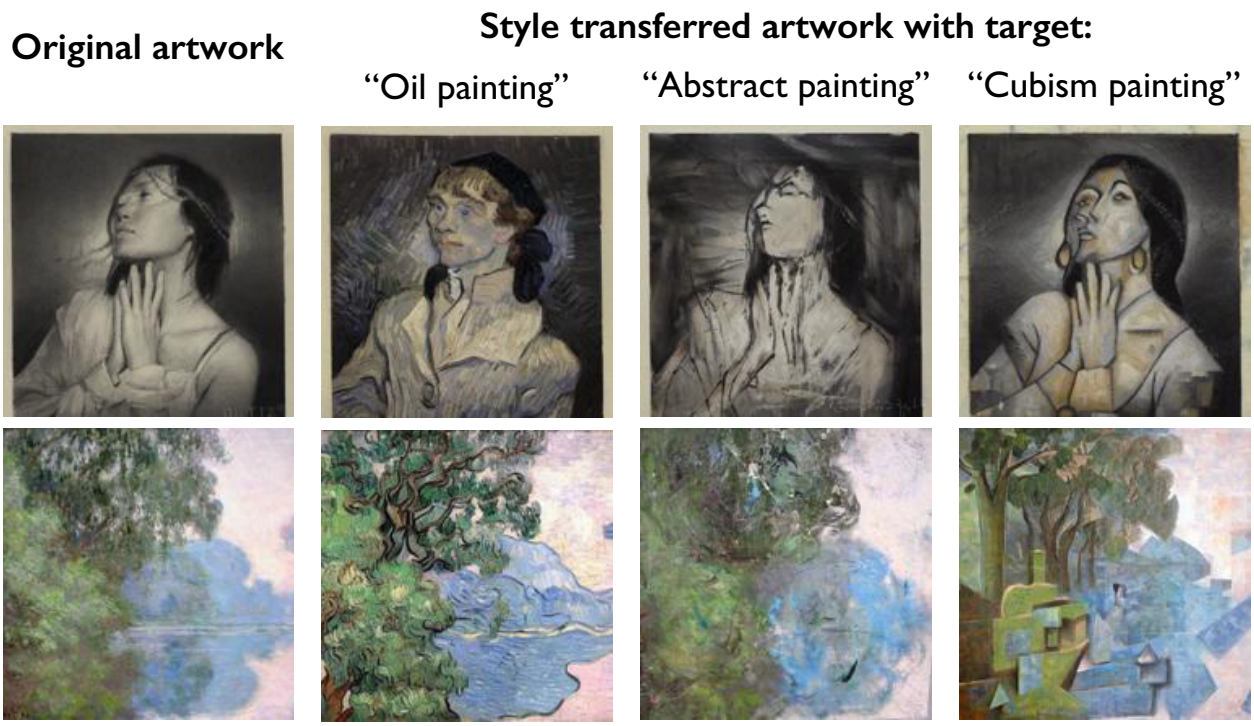


Figure 4.12: Example style-transferred artwork with different target styles.

and artwork-dependent method that operates at image space. Given an artwork, we leverage "style transfer," an end-to-end computer vision technique, to modify and isolate its style components. "Style transfer" transforms an image into a new image with a different style (e.g., from impressionist style to cubist style) while keeping other aspects of the image similar (e.g., subject matter and location).

We leverage style transfer in our protection technique as follows. Given an original artwork from the victim artist, we apply style-transfer to produce a similar piece of art with a different style, e.g., in style of "an oil painting by Van Gogh" (Figure 4.11 a). The new version has similar content to the original, but its style mirrors that of Van Gogh. We show more style-transfer examples with different target styles in Figure 4.12. Now, we can use the style-transferred artwork as projection target to guide the perturbation computation. This perturbs the original artwork's style-specific features towards that of the style-transferred version. We do this by optimizing a cloak that, when added to the original artwork, makes

Figure 4.13: Example *Glaze* protection results for three artists. **Columns 1-2**: artist's original artwork; **column 3**: mimicked artwork when artist does not use protection; **column 4**: style-transferred artwork (original artwork in column 1 is the source) used for cloak optimization and the name of target style; **column 5-6**: mimicked artwork when artist uses cloaking protection with perturbation budget $p = 0.05$ or $p = 0.1$ respectively. All mimicry examples here use SD-based models.

its feature representation similar to the style-transferred image. Since the content is identical between the pair of images, cloak optimization will focus its perturbation budget on style features.

4.4.3 Glaze*'s Protection Performance*

Style mimicry success when *Glaze* is not used. Mimicry attacks are very successful when the mimic has access to a victim's original (unmodified) artwork. Examples of mimicked artwork can be found in Figure 4.13. The leftmost two columns of Figure 4.13 show a victim artist's original artwork, while the third column depicts mimicked artwork generated by a style-specific model trained on victim's original artwork when *Glaze* is not used. In our user study, over $> 95\%$ of respondents rated the attack as successful. Table 4.14, row 1, gives the artist-rated and CLIP-based genre shift for mimicry attacks on unprotected

Generic model	Artist dataset	w/o *Glaze*		w/ *Glaze* (p=0.05)	
		Artist-rated PSR	CLIP-based genre shift	Artist-rated PSR	CLIP-based genre shift
SD	Current	$4.6 \pm 0.3\%$	$2.4 \pm 0.2\%$	$94.3 \pm 0.8\%$	$96.4 + 0.5\%$
	Historical	$4.2 \pm 0.2\%$	$1.3 \pm 0.2\%$	$93.3 + 0.6\%$	$96.0 + 0.3\%$
DALL·E-m	Current	$31.9 \pm 3.5\%$	$6.4 \pm 0.8\%$	$97.4 \pm 0.2\%$	$97.4 + 0.3\%$
	Historical	$29.8 \pm 2.4\%$	$5.8 \pm 0.6\%$	$96.8 \pm 0.3\%$	$97.1 + 0.2\%$

Figure 4.14: *Glaze* has a high protection success rate, as measured by artists and CLIP, against style mimicry attacks. We compare protection success when artists do not use *Glaze* vs. when they do (with perturbation budget 0.05).

art.

SD models produce stronger mimicry attacks than DALL·E-m models, according to our user study (see Table 4.14). This is unsurprising, as DALL·E-m models generally produce lower-quality generated images. CLIP-based genre shift does not reflect this phenomenon, as this metric does not assess image quality.

***Glaze*'s success at preventing style mimicry.** *Glaze* makes mimicry attacks markedly less successful, as shown in Figure 4.13. Columns 5 and 6 (from left) show mimicked artwork when the style-specific models are trained on artwork protected by *Glaze*. For reference, column 4 shows an example style-transferred artwork $\Omega(x, T)$ used to compute *Glaze* cloaks for the protected art pieces. Overall, *Glaze* achieves $> 93.3\%$ artist-rated PSR and $> 96.0\%$ CLIP-based genre shift (see Table 4.14). *Glaze*'s protection performance is slightly higher for current artists than for historical artists. This is likely because the historical artists' images are present in the training datasets of our generic models (SD, DALL·E-m), highlighting the additional challenge of protecting well-known artists whose style was already learned by the generic models.

How large of perturbations will artists tolerate? Increasing the *Glaze* perturbation budget enhances protection performance. We observe that both artist-rated and CLIP-based

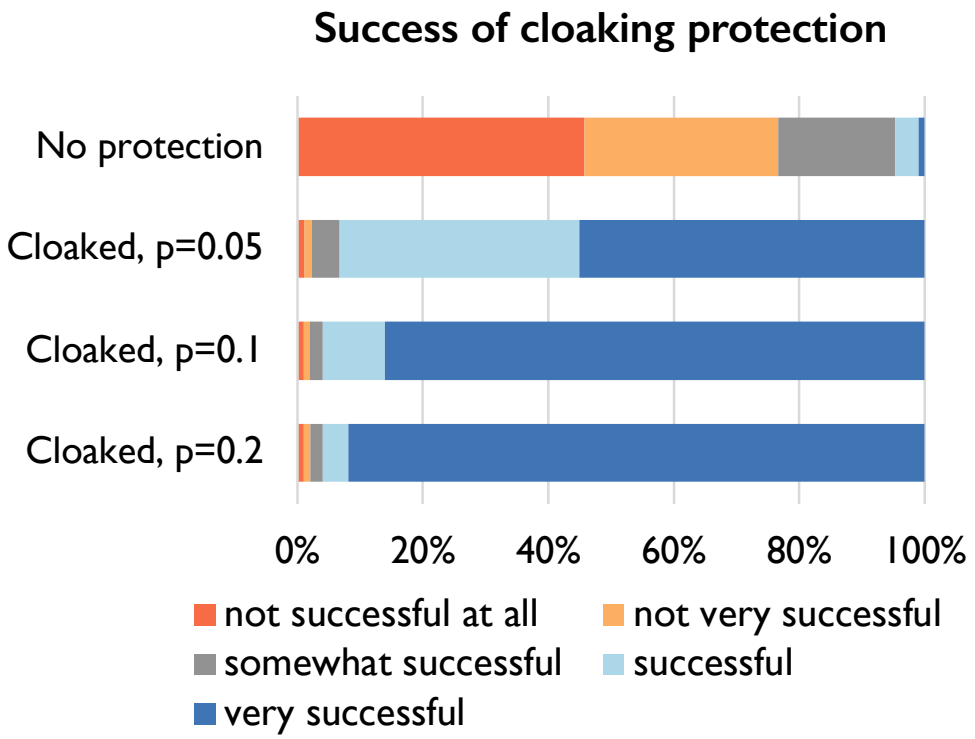


Figure 4.15: *Glaze*'s cloaking protection success increases as cloak perturbation budget increases. The top row of the figure shows baseline performance with the mimic trains on uncloaked images (p=0).

genre shift increase with perturbation budget (see Figure 4.15, Table 4.16, and Figure 4.19). Given this tradeoff between protection success and *Glaze* protection visibility on original artwork, we evaluate how perturbation size impacts artists' willingness to use *Glaze*.

We find that artists are willing to add fairly large *Glaze* perturbations to their artwork in exchange for protection against mimicry. To measure this, we show 3 randomly chosen pairs of original/cloaked artwork to each of the 1,156 artists in our first study. For each art pair, we ask the artist whether they would be willing to post the cloaked artwork (instead of the original, unmodified version) on their personal website. More than 92% of artists select "willing" or "very willing" when $p = 0.05$. This number only slightly increases to 94.3% when $p = 0.03$. Figure 4.17 details artists' preferences as perturbation budget increases. (see Figure 4.18 for examples of cloaked artwork with increasing p). Based on these results,

Perturbation budget	Artist-rated PSR	CLIP-based genre shift
0 (no cloak)	$4.6 \pm 1.4\%$	$2.4 \pm 0.8\%$
0.05	$93.3 \pm 0.6\%$	$96.0 \pm 0.3\%$
0.1	$95.9 \pm 0.4\%$	$98.2 \pm 0.1\%$
0.2	$96.1 \pm 0.3\%$	$98.5 \pm 0.1\%$

Figure 4.16: Performance of our system (artist-rated protection success rate and CLIP-based genre shift rate) increases as the perturbation budget increases. (SD model, averaged over all victim artists).

we use perturbation budget $p = 0.05$ for all our experiments, since most artists are willing to tolerate this perturbation size.

Surprisingly, over 32.8% artists are willing to use cloaks with $p = 0.2$, which is clearly visible to human eye (see Figure 4.18). While we are surprised by this high perturbation tolerance, in our follow-up free response artists noted that they would be willing to tolerate large perturbations because of the devastating consequence if their styles are stolen. One participant stated that "I am willing to sacrifice a bit image quality for protection." Many artists ($> 80\%$) also noted that they have already used traditional, more visually disruptive techniques to protect their artwork online when posting online, i.e., adding watermark or reducing image resolution. One participant stated that "I already use low to medium resolution images only for online posting, thus this would not impact my quality control too much."

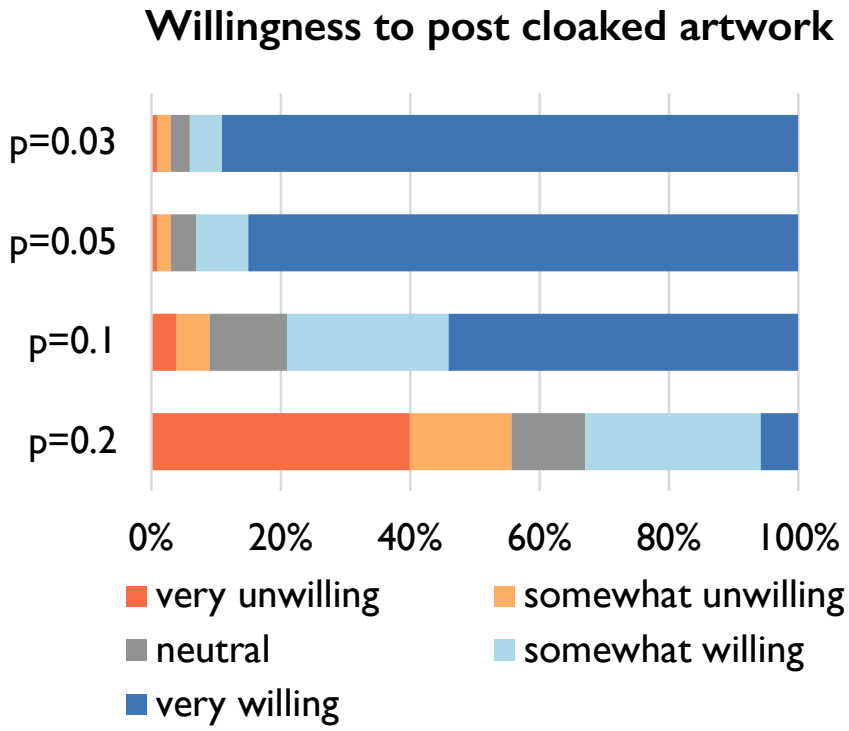


Figure 4.17: Artists' willingness to post cloaked artwork in place of the original decreases as perturbation budget of the cloaks increases.

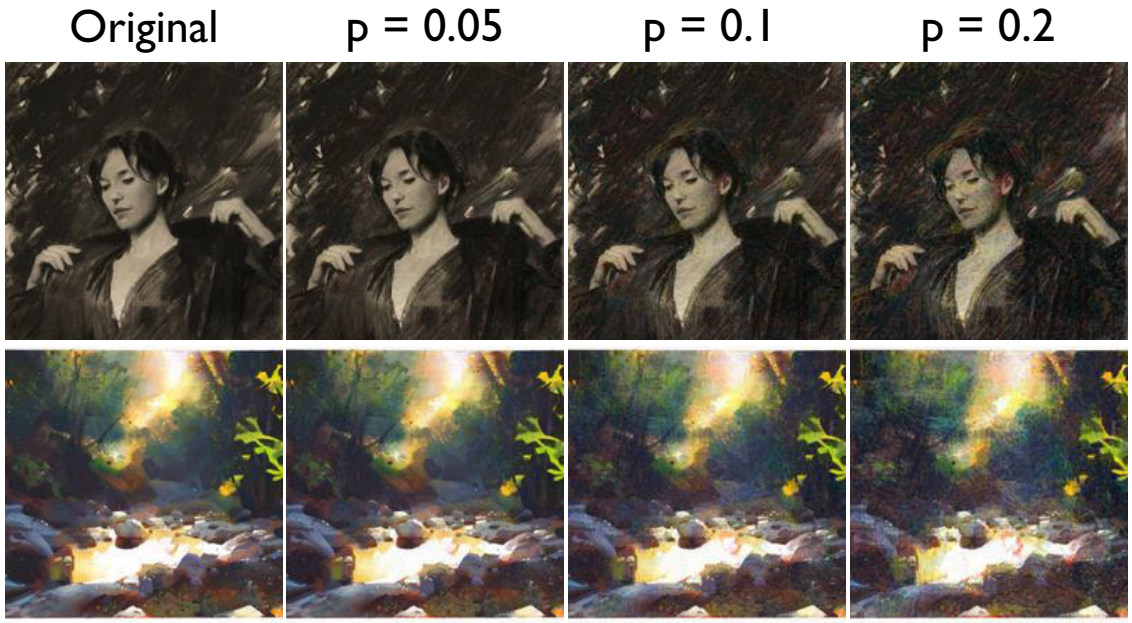


Figure 4.18: Original artwork and cloaked artwork computed using three different cloak perturbation budgets.

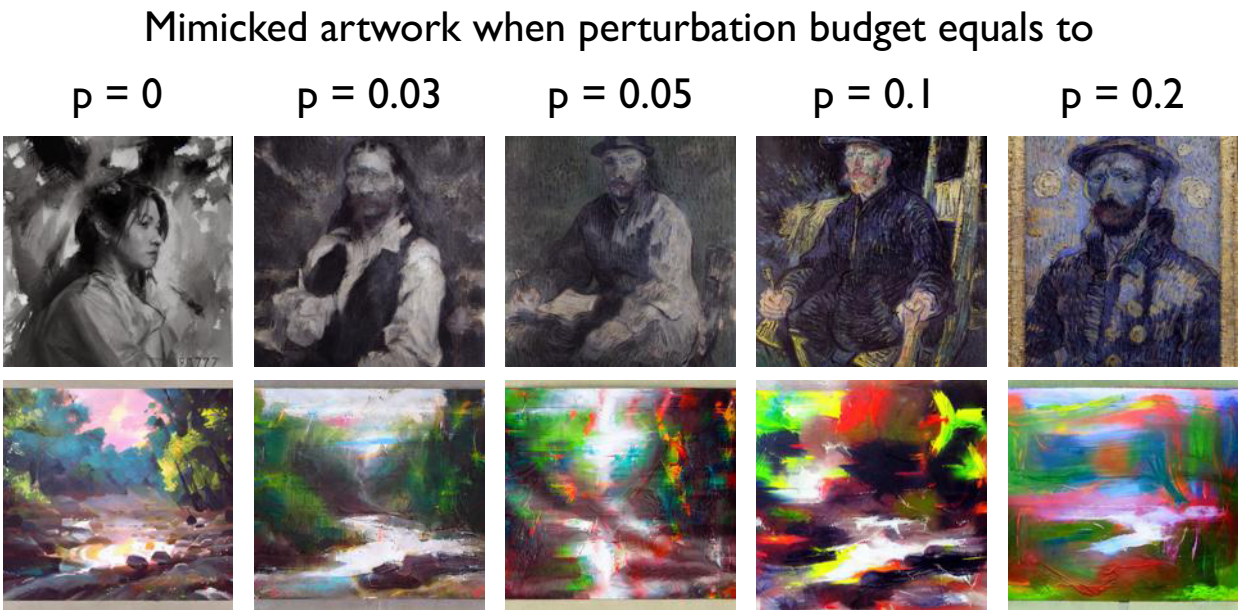


Figure 4.19: Mimicked artwork when artist uses an increasingly high perturbation budget to protect their original art.

Figure 4.20: *Glaze* remains successful under two challenging scenarios. Left: when artist and mimic use different feature extractors. Right: when artists can only cloak a portion of their artwork in mimic's dataset. Bottom of the figure shows artist-rated PSR and CLIP-based genre shift for the corresponding setting.

CHAPTER 5

DISCUSSION

This dissertation seeks to understand and improve ways of measuring and mitigating bias by evaluating the efficacy of existing data and methods, and how they represent real-world perceptions. By improving methods to measure gendered language in text, I demonstrate the importance of re-examining and updating both the data and methodology to *capture modern perspectives in context* that may perpetuate gender biased stereotypes. To *prevent* biased preconceptions, I demonstrate how emotion voice conversion models can be leveraged to reduced emotion expressions, and improve perceptions of the speaker.

Bias in Text. Understanding how we use gendered language in text allows us to understand how people are being represented, and additionally how they may be perceived based on those representations. Correct or not, these perspectives reflect and perpetuate gender biased stereotypes. Through an examination of traditional and novel methods to measure gendered language in text, I find that gendered language persists over time, with evolving vocabulary and context. While traditional gender lexicons and associated methods maintain merit, this work shows how leveraging large scale data improves representation and recognition of modern language use. However, even with additional data, lexicons still struggle to understand language in context. To remedy the shortcomings of traditional methods, I trained a deep-learning language model to recognize broader sentiments expressed beyond single words or short phrases. By training on entire paragraphs or articles of labeled text, the end-to-end method captures the gendered language of larger bodies of text. The gender score generated by the end-to-end model provides a better representation of the overall intent of the language compared to the lexicon approaches.

By re-evaluating the type of data we could gather, and how it could be applied to state-of-the-art methods, we can improve our ability to measure gendered language in text across

various medias. Overall, simply measuring gendered language does not inherently reduce bias. It may, however, encourage people to be more aware of the wording they use when describing people, and the impact it can have on the subject and readers of the content. Further, though it remains costly in time and effort to update data and methods, this work shows how the resulting improvements are significant and necessary. By gaining a better understanding of how language may reflect and perpetuate biased perspectives, perhaps next we can enact changes that prevent the biased perceptions.

Bias in Voice. With the use of voice conversion models, this work shows the potential for preventing biased perceptions by changing the tone of a speaker's voice. While emotion voice conversion models continue to be developed and improved, this work presents a real-world use case for emotion voice conversion models that can result in benefits for both the speaker and listener. After evaluating the datasets these models are originally trained and evaluating on, it's clear that current dataset labels remain very limited in their ability to capture an adequate spectrum of emotion expressions. Nonetheless, by using them to *reduce* emotion expressions (of any emotion), I find speakers are more likely to appear more competent, trustworthy, and less anxious. In certain scenarios, such as talking to someone at a call center, these changes can greatly improve the experience for both parties. The expressed disinterest and skepticism about voice manipulation highlights the need for end-users to be considered when voice conversion models are created and used in the real world. However, the expressed understanding of how certain conversational experiences could benefit from someone by altering their tone of voice suggests there may be scenarios where such a tool could be useful. Overall, while emotion voice conversion models show the ability to improve conversation experiences, researchers should maintain awareness of people's perspectives, and use caution when implementing such tools in the real world.

Artificial Content in the Real-World Deep learning presents great potential, while also posing a great threat to people and society at large. As these models continue to become more sophisticated, the content created becomes more and more difficult for humans and systems to distinguish as artificially generated. As presented in this thesis, this poses security threats, such as in the case of voice authentication. Additionally, as shown with Glaze, artificial content can displace the very humans whose work is used to train the models. It remains up to developers to be cognizant of the models they create and release to the public, and to those who can work towards defenses on behalf of the general public.

REFERENCES

[1] The biases we hold against the way people speak. https://www.nytimes.com/2020/07/21/books/review/how-you-say-it-katherine-kinzler.html. Accessed: 2022-11-30.

[2] *VideoProv: Verifiable Provenance for Videos from Mobile Devices*, 2022.

[3] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. *Mining text data*, pages 163–222, 2012.

[4] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. Void: A fast and light voice liveness detection system. In *Proc. of USENIX*, 2020.

[5] Feras Al Taha, Pascal E Fortin, Antoine Weill-Duflos, and Jeremy Cooperstock. Reversing voice-related biases through haptic reinforcement. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, pages 60–62, 2018.

[6] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. Detecting ai-synthesized speech using bispectral analysis. In *CVPR Workshops*, 2019.

[7] Nalini Ambady, Debi LaPlante, Thai Nguyen, Robert Rosenthal, Nigel Chaumeton, and Wendy Levinson. Surgeons' tone of voice: a clue to malpractice history. *Surgery*, 132(1):5–9, 2002.

[8] Rindy C Anderson and Casey A Klofstad. Preference for leaders with masculine voices holds in the case of feminine leadership roles. *PloS one*, 7(12):e51216, 2012.

[9] Pablo Arias, Laura Rachman, Marco Liuni, and Jean-Julien Aucouturier. Beyond correlation: acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1):12–24, 2021.

[10] Claire E Ashton-James, Joshua M Tybur, Verena Grießer, and Daniel Costa. Stereotypes about surgeon warmth and competence: the role of surgeon gender. *PLoS One*, 14(2):e0211890, 2019.

[11] Inger Askehave and Karen K Zethsen. Gendered constructions of leadership in danish job advertisements. *Gender, Work & Organization*, 21(6):531–545, 2014.

[12] Leanne E Atwater, Joan F Brett, David Waldman, Lesley DiMare, and MaryVirginia Hayden. Men's and women's perceptions of the gender typing of management subroles. *Sex Roles*, 50(3):191–199, 2004.

[13] Jean-Julien Aucouturier, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4):948–953, 2016.

[14] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[15] Mahzarin R Banaji and Curtis D Hardin. Automatic stereotyping. *Psychological science*, 7(3):136–141, 1996.

[16] Fang Bao, Michael Neumann, and Ngoc Thang Vu. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *INTERSPEECH*, pages 2828–2832, 2019.

[17] Sandra L Bem. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2):155–162, 1974.

[18] Sandra L Bem and Daryl J Bem. Does sex-biased job advertising "aid and abet" sex discrimination? *Journal of Applied Social Psychology*, 3(1):6–18, 1973.

[19] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991–1013, 2004.

[20] Francine D Blau and Lawrence M Kahn. The gender wage gap: Extent, trends, and explanations. Technical report, National Bureau of Economic Research, 2016.

[21] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347, 2001.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

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

[24] Barbara Borkowska and Boguslaw Pawlowski. Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1):55–59, 2011.

[25] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. Key challenges in defending against malicious socialbots. In *Proc. of LEET Workshop*, 2012.

[26] Constantine Boussalis, Travis G Coan, Mirya R Holman, and Stefan Müller. Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, 115(4):1242–1257, 2021.

[27] Blake Brittain. AI-created images lose U.S. copyrights in test for new technology. Reuters, February 2023.

[28] Anna Brown and Eileen Patten. The narrowing, but persistent, gender gap in pay. `http://www.pewresearch.org/fact-tank/2017/04/03/gender-pay-gap-facts/`, 2017.

[29] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[30] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(14):183–186, April 2017.

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

[32] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[33] José Carlos Castillo, Álvaro Castro-González, Fernándo Alonso-Martín, Antonio Fernández-Caballero, and Miguel Ángel Salichs. Emotion detection and regulation from personal assistant robot in smart environment. In *Personal assistants: Emerging computational technologies*, pages 179–195. Springer, 2018.

[34] Stephen J Ceci, Donna K Ginther, Shulamit Kahn, and Wendy M Williams. Women in academic science a changing landscape. *Psychological Science in the Public Interest*, 15(3):75–141, 2014.

[35] Mary A Cejka and Alice H Eagly. Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, 25(4):413–423, 1999.

[36] Henry S Cheang and Marc D Pell. The sound of sarcasm. *Speech communication*, 50(5):366–381, 2008.

[37] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real bob? adversarial attacks on speaker recognition systems. *arXiv preprint arXiv:1911.01840*, 2019.

[38] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018.

[39] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Proc. of ICDS*, 2017.

[40] Jacob Clark Blickenstaff*. Women and science careers: leaky pipeline or gender filter? *Gender and education*, 17(4):369–386, 2005.

[41] Jean Costa, Malte F Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[42] Karen L Cropsey, Saba W Masho, Rita Shiang, Veronica Sikka, Susan G Kornstein, and Carol L Hampton. Why do faculty leave? reasons for attrition of women and minority faculty from a medical school: four-year results. *Journal of Women's Health*, 17(7):1111–1118, 2008.

[43] Amy JC Cuddy, Susan T Fiske, and Peter Glick. When professionals become mothers, warmth doesn't cut the ice. *Journal of Social issues*, 60(4):701–718, 2004.

[44] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proc. of CVPR*, 2020.

[45] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[46] M Lee Dean and Charlotte Chucky Tate. Extending the legacy of sandra bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science. *Sex Roles*, 76(11-12):643–654, 2017.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[48] Kristin Donnelly and Jean M Twenge. Masculine and feminine traits on the bem sex-role inventory, 1993–2012: a cross-temporal meta-analysis. *Sex Roles*, pages 1–10, 2016.

[49] Mathew Dryhurst. AI art and the problem of consent. ArtReview, January 2023.

[50] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Expressive voice conversion: A joint framework for speaker identity and emotional style transfer. *arXiv preprint arXiv:2107.03748*, 2021.

[51] Alice H Eagly, Wendy Wood, and Amanda B Diekman. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender*, 12:174, 2000.

[52] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[53] `http://www.fakenewschallenge.org`, 2017.

[54] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5279–5283. IEEE, 2018.

[55] Ethan Fast, Tina Vachovsky, and Michael S Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*, 2016.

[56] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Exploiting burstiness in reviews for review spammer detection. 2013.

[57] Elizabeth Flux. What does the rise of AI mean for the future of art? Sydney Morning Herald, Dec. 2022.

[58] Martha Foschi, Larissa Lai, and Kirsten Sigerson. Gender and double standards in the assessment of job applicants. *Social Psychology Quarterly*, 57(4):326–339, 1994.

[59] Vicki Fox. AI art & the ethical concerns of artists. Beautiful Bizarre, March 2023.

[60] Paul E Gabriel and Susanne Schmitz. Gender differences in occupational distributions among workers. *Monthly Labor Review*, 130:19, 2007.

[61] Haichang Gao, Honggang Liu, Dan Yao, Xiyang Liu, and Uwe Aickelin. An audio captcha to distinguish humans from computers. In *2010 Third International Symposium on Electronic Commerce and Security*, pages 265–269. IEEE, 2010.

[62] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109, 2011.

[63] Peter Glick and Susan T Fiske. The ambivalent sexism inventory: differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512, 1996.

[64] Oana Goga, Giridhari Venkatadri, and Krishna P. Gummadi. The doppelgänger bot attack: Exploring identity impersonation in online social networks. In *Proc. of IMC*, 2015.

[65] Linda S Gottfredson. Circumscription and compromise: A developmental theory of occupational aspirations. *Journal of Counseling psychology*, 28(6):545, 1981.

[66] Louise Goupil, Emmanuel Ponsot, Daniel Richardson, Gabriel Reyes, and Jean-Julien Aucouturier. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, 12(1):861, 2021.

[67] The World Bank Group. Many societies gradually moving to dismantle gender discrimination, yet more can be done, says world bank group president jim yong kim. `http://www.worldbank.org/en/news/press-release/2013/09/24/societies-dismantle-gender-discrimination-world-bank-group-president-jim-yong-kim`, 2017.

[68] Matt Growcoot. Midjourney founder admits to using a 'hundred million' images without consent. PetaPixel, Dec 2022.

[69] Nadia Guerouaou, Guillaume Vaiva, and Jean-Julien Aucouturier. The shallow of your smile: the ethics of expressive vocal deep-fakes. *Philosophical Transactions of the Royal Society B*, 377(1841):20210083, 2022.

[70] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[71] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: evidence from taskrabbit and fiverr. In *Proc. of CSCW*, 2017.

[72] Marvin A Hecht and Marianne LaFrance. How (fast) can i help you? tone of voice and telephone operator efficiency in interactions 1. *Journal of Applied Social Psychology*, 25(23):2086–2098, 1995.

[73] Melissa Heikkila. This artist is dominating AI-generated art. and he's not happy about it. MIT Technology Review, Sept 2022.

[74] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proc. of ECCV*, pages 793–811, 2018.

[75] Karen Holtzblatt, Aruna Balakrishnan, Troy Effner, Emily Rhodes, and Tina Tuan. Beyond the pipeline: addressing diversity in high tech. In *Proc. of CHI Extended Abstract*, 2016.

[76] Lisa K Horvath and Sabine Sczesny. Reducing women's lack of fit with leadership positions? effects of the wording of job advertisements. *European Journal of Work and Organizational Psychology*, 25(2):316–328, 2016.

[77] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. Defending your voice: Adversarial attack on voice conversion. *Proc. of IEEE SLT Workshop*, 2021.

[78] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proc. of CVPR*, 2020.

[79] Nitin Jindal and Bing Liu. Review spam detection. In *Proc. of WWW*, 2007.

[80] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proc. of WSDM*, 2008.

[81] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *International Conference on Machine Learning*, pages 526–534, 2016.

[82] Benedict C Jones, David R Feinberg, Lisa M DeBruine, Anthony C Little, and Jovana Vukovic. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, 79(1):57–62, 2010.

[83] Kshiti D Joshi and Nancy L Schmidt. Is the information systems profession gendered?: characterization of is professionals and is career. *ACM SIGMIS Database*, 37(4):26–41, 2006.

[84] Roza G Kamiloğlu, Agneta H Fischer, and Disa A Sauter. Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic bulletin & review*, 27(2):237–265, 2020.

[85] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. of CVPR*, 2019.

[86] Casey A Klofstad, Rindy C Anderson, and Susan Peters. Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738):2698–2704, 2012.

[87] M. W. Kraus, B. Torrez, J. W. L. Park, and F. Ghayebi. Evidence for the reproduction of social class in brief speech. *Proceedings of the National Academy of Sciences*, 116(46):22998–23003, 2019.

[88] Michael W Kraus, Brittany Torrez, Jun Won Park, and Fariba Ghayebi. Evidence for the reproduction of social class in brief speech. *Proceedings of the National Academy of Sciences*, 116(46):22998–23003, 2019.

[89] Amanda Krause. People are editing photos of celebrities to give them Instagram-inspired faces. experts say it could be harmful, 2020. `https://www.insider.com/why-edited-photos-of-celebrities-can-be-harmful-2020-9`.

[90] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *Proc. of ICASSP*, 2018.

[91] Arie W Kruglanski and Icek Ajzen. Bias and error in human judgment. *European Journal of Social Psychology*, 13(1):1–44, 1983.

[92] Siddique Latif, Muhammad Asim, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W Schuller. Augmenting generative adversarial networks for speech emotion recognition. *arXiv preprint arXiv:2005.08447*, 2020.

[93] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proc. of INTERSPEECH*, 2017.

[94] Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. Emotion and decision making. *Annual review of psychology*, 66(1), 2015.

[95] Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review spam. In *Proc. of IJCAI*, 2011.

[96] Zhuohang Li, Cong Shi, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Practical adversarial attacks against speaker recognition systems. In *Proc. of HotMobile*, 2020.

[97] A Lieto, D Moro, F Devoti, C Parera, Vincenzo Lipari, Paolo Bestagini, and Stefano Tubaro. "hello? who am i talking to?" a shallow cnn approach for human vs. bot speech classification. In *Proc. of ICASSP*, 2019.

[98] Tze Wei Liew, Su-Mae Tan, and Chin Lay Gan. Interacting with motivational virtual agent: The effects of message framing and regulatory fit in an e-learning environment. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 136–141. IEEE, 2018.

[99] Songxiang Liu, Yuewen Cao, and Helen Meng. Emotional voice conversion with cycle-consistent adversarial network. *arXiv preprint arXiv:2004.03781*, 2020.

[100] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of ICCV*, 2015.

[101] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[102] Lela London. How beauty filters are making us 'look better' but feel worse. `https://www.forbes.com/sites/lelalondon/2020/03/23/in-self-isolation-filter-dysmorphia-and-beauty-filters-will-threaten-our-mental-he?sh=370b0c903831`, 2020.

[103] Anne Maass and Luciano Arcuri. Language and stereotyping. *Stereotypes and stereotyping*, pages 193–226, 1996.

[104] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

[105] Josh Mandell. Spotify patents a voice assistant that can read your emotions, 2020. https://www.forbes.com/sites/joshmandell/2020/03/12/spotify-patents-a-voice-assistant--that-can-read-your-emotions/?sh=5e9de2f038d5.

[106] Huanru Henry Mao, Taylor Shin, and Garrison Cottrell. Deepj: Style-specific music generation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 377–382. IEEE, 2018.

[107] Cynthia M Marlowe, Sandra L Schneider, and Carnot E Nelson. Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology*, 81(1):11, 1996.

[108] Catherine C. Marshall and Frank M. Shipman. Who owns the social web? *Commun. ACM*, 60(5):52–61, apr 2017.

[109] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proc. of NAACL*, 2019.

[110] Lisa A McLoughlin. Spotlighting: Emergent gender bias in undergraduate engineering education. *Journal of Engineering Education*, 94(4):373, 2005.

[111] Michela Menegatti and Monica Rubini. Gender bias and sexism in language, 2017.

[112] Midjourney. Community Showcase, 2022.

[113] Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proc. of ACL-IJCNLP*, 2009.

[114] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[115] David I Miller and Jonathan Wai. The bachelor's to ph. d. stem pipeline no longer leaks more women than men: a 30-year analysis. *Frontiers in psychology*, 6:37, 2015.

[116] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[117] Harriet Minter. Why so few women in tech? your answer depends on your gender. https://www.theguardian.com/careers/2016/dec/12/why-so-few-women-in-tech-your-answer-depends-on-your-gender, 2017.

[118] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, 2018.

[119] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012.

[120] Asuka Moritani, Ryo Ozaki, Shoki Sakamoto, Hirokazu Kameoka, and Tadahiro Taniguchi. Stargan-based emotional voice conversion for japanese phrases. *arXiv preprint arXiv:2104.01807*, 2021.

[121] Lisa M Moynihan, Mark V Roehling, Marcie A LePine, and Wendy R Boswell. A longitudinal study of the relationships among job search self-efficacy, job interviews, and employment outcomes. *Journal of Business and Psychology*, 18(2):207–233, 2003.

[122] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Spotting opinion spammers using behavioral footprints. In *Proc. of KDD*, 2013.

[123] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. All your voices are belong to us: Stealing voices to fool humans and machines. In *Proc. of ESORICS*, 2015.

[124] David Neumark, Roy J Bank, and Kyle D Van Nort. Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics*, 111(3):915–941, 1996.

[125] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236, 2008.

[126] U.S. Bureau of Labor Statistics. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. `https://www.bls.gov/cps/aa1998/CPSAAT39.PDF`, 1998.

[127] U.S. Bureau of Labor Statistics. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex. `https://www.bls.gov/cps/cpsaat39.htm`, 2017.

[128] Astrid Paeschke and Walter F Sendlmeier. Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *Isca tutorial and research workshop (itrw) on speech and emotion*, 2000.

[129] Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Copypaste: An augmentation method for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6324–6328. IEEE, 2021.

[130] Pratik Parikh, Ketaki Velhal, Sanika Potdar, Aayushi Sikligar, and Ruhina Karani. English language accent classification and conversion using machine learning. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*, 2020.

[131] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*, pages 2799–2804, 2018.

[132] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet::similarity: Measuring the relatedness of concepts. In *Proc. of HLT-NAACL (Demonstration Paper)*, 2004.

[133] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[134] Daniel Oliveira Peres, Dominic Watt, and Waldemar Ferreira Netto. Emotional thin-slicing: A proposal for a short-and long-term division of emotional speech. In *INTERSPEECH*, pages 591–595, 2017.

[135] Cyril R Pernet and Pascal Belin. The role of pitch and timbre in voice gender categorization. *Frontiers in psychology*, 3:23, 2012.

[136] Luke Plunkett. AI creating 'art' is an ethical and copyright nightmare. Kotaku, August 2022.

[137] Deborah M Powell, Joshua S Bourdage, and Silvia Bonaccio. Shake and fake: The role of interview anxiety in deceptive impression management. *Journal of business and psychology*, 36(5):829–840, 2021.

[138] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[139] Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. David: An open-source platform for real-time emotional speech transformation: With 25 applications in the behavioral sciences. *bioRxiv*, page 038133, 2016.

[140] Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. Linguistic analysis of differences in portrayal of movie characters. In *Proc. of ACL*, volume 1, pages 1669–1678, 2017.

[141] `http://www.formstack.com/forms/?1653778-I3QqcHV4xC`, 2017.

[142] Peter A Riach and Judith Rich. An experimental investigation of sexual discrimination in hiring in the english labor market. *The B.E. Journal of Economic Analysis & Policy*, 5(2):1–22, 2006.

[143] Cecilia L Ridgeway. Interaction and the conservation of gender inequality: Considering employment. *American Sociological Review*, 62(2):218–235, 1997.

[144] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3502–3506. IEEE, 2020.

[145] Sonja Rohrmann, Myriam N Bechtoldt, Henrik Hopp, Volker Hodapp, and Dieter Zapf. Psychophysiological effects of emotional display rules and the moderating role of trait anger in a simulated call center. *Anxiety, Stress & Coping*, 24(4):421–438, 2011.

[146] Kevin Roose. An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. The New York Times, Sept. 2022.

[147] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32(3):287, 1968.

[148] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proc. of ICCV*, 2019.

[149] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proc. of ICCV workshops*, 2015.

[150] As fake news spreads lies, more readers shrug at the truth. `https://www.nytimes.com/2016/12/06/us/fake-news-partisan-republican-democrat.html`, 2016.

[151] Myra Sadker and David Sadker. *Failing at fairness: How America's schools cheat girls*. Simon and Schuster, 2010.

[152] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. Vesus: A crowd-annotated database to study emotion production and perception in spoken english. In *INTERSPEECH*, pages 316–320, 2019.

[153] Rob Salkowitz. AI is coming for commercial art jobs. can it be stopped? Forbes, Sept 2022.

[154] Klaus R Scherer. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In *INTERSPEECH*, volume 4, pages 379–382, 2000.

[155] Klaus R Scherer and James S Oshinsky. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1:331–346, 1977.

[156] Annett Schirmer, Man Hey Chiu, Clive Lo, Yen-Ju Feng, and Trevor B Penney. Angry, old, male–and trustworthy? how expressive and person voice characteristics shape listener trust. *Plos one*, 15(5):e0232431, 2020.

[157] Toni Schmader, Jessica Whitehead, and Vicki H Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, 57(7-8):509–514, 2007.

[158] Bastian Schnell, Goeric Huybrechts, Bartek Perz, Thomas Drugman, and Jaime Lorenzo-Trueba. Emocat: Language-agnostic emotional voice conversion. *arXiv preprint arXiv:2101.05695*, 2021.

[159] Stephen A Schullo and Burton L Alperson. Interpersonal phenomenology as a function of sexual orientation, sex, sentiment, and trait categories in long-term dyadic relationships. *Journal of Personality and Social Psychology*, 47(5):983, 1984.

[160] Sabine Sczesny, Magda Formanowicz, and Franziska Moser. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7:25, 2016.

[161] Maliheh Shirvanian, Manar Mohammed, Nitesh Saxena, and S Abhishek Anand. Voicefox: Leveraging inbuilt transcription to enhance the security of machine-human speaker verification against voice synthesis attacks. In *Annual Computer Security Applications Conference*, pages 870–883, 2020.

[162] Steven Davidoff Solomon. Why so few women reach the executive rank. `https://dealbook.nytimes.com/2013/04/02/why-so-few-women-reach-the-executive-rank/`, 2017.

[163] Janet T Spence, Robert Helmreich, and Joy Stapp. Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 32(1):29, 1975.

[164] Janet T. Spence, Robert L. Helmreich, and Joy Stapp. The personal attributes questionnaire: A measure of sex role stereotypes and masculinity-femininity. *JSAS Catalog of selected documents in psychology*, 4(43), 1974.

[165] Rhea E Steinpreis, Katie A Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7):509–528, 1999.

[166] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[167] Yaşar Suveren. Unconscious bias: Definition and significance. *Psikiyatride Guncel Yaklasimlar*, 14(3):414–426, 2022.

[168] Latany Sweeney. Discrimination in online ad delivery. In *arXiv:1301.6822*, 2013.

[169] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[170] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. Gender bias in the job market: A longitudinal analysis. In *Proc. of CSCW*, 2017.

[171] `http://qwone.com/~jason/20Newsgroups/`, 2008.

[172] Erik R Thomas. Sociophonetic applications of speech perception experiments. *American speech*, 77(2):115–147, 2002.

[173] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. The voice conversion challenge 2016. In *Interspeech*, pages 1632–1636, 2016.

[174] In praise of fake reviews. `https://thenewinquiry.com/essays/in-praise-of-fake-reviews`, 2014.

[175] Donald Tomaskovic-Devey. *Gender & racial inequality at work: The sources and consequences of job segregation*. Number 27. Cornell University Press, 1993.

[176] Eileen M Trauth, Curtis C Cain, KD Joshi, Lynette Kvasny, and Kayla M Booth. The influence of gender-ethnic intersectionality on gender stereotypes about it skills and knowledge. *ACM SIGMIS Database*, 47(3):9–39, 2016.

[177] Jean M Twenge. Changes in masculine and feminine traits over time: A meta-analysis. *Sex roles*, 36(5-6):305–325, 1997.

[178] Mikel deVelasco Vázquez, Raquel Justo, Asier López Zorrilla, and Maria Inés Torres. Can spontaneous emotions be detected from speech on tv political debates? In *10th IEEE International Conference on Cognitive Infocommunications*, page 289, 2019.

[179] Bimal Viswanath, Muhammad A. Bashir, Muhammad B. Zafar, Simon Bouget, Saikat Guha, Krishna P. Gummadi, Aniket Kate, and Alan Mislove. Strength in numbers: Robust tamper detection in crowd computations. In *Proc. of COSN*, 2015.

[180] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proc. of ACL*, 2014.

[181] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*, 2015.

[182] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Review graph based online store review spammer detection. In *Proc. of ICDM*, 2011.

[183] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *Proc. of INFOCOM*, 2019.

[184] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. *arXiv preprint arXiv:2005.13770*, 2020.

[185] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. of ACL*, 2017.

[186] Samuel F Way, Daniel B Larremore, and Aaron Clauset. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proc. of WWW*, 2016.

[187] Antoine Weill-Duflos, Feras Al Taha, Pascal E Fortin, and Jeremy R Cooperstock. Barrywhaptics: Towards countering social biases using real-time haptic enhancement of voice. In *2019 IEEE World Haptics Conference (WHC)*, pages 365–370. IEEE, 2019.

[188] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y. Zhao. "hello, it's me": Deep learning-based speech synthesis attacks in the real world. In *Proc. of CCS*, November 2021.

[189] Mingke Xu, Fan Zhang, Xiaodong Cui, and Wei Zhang. Speech emotion recognition with multiscale area attention and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, 2021.

[190] Junichi Yamagishi, Christophe Veaux, and Kirsten. MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

[191] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proc. of CCS*, 2019.

[192] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, page 35, 1997.

[193] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[194] Romi Zäske and Stefan R Schweinberger. You are only as old as you sound: Auditory aftereffects in vocal age perception. *Hearing Research*, 282(1-2):283–288, 2011.

[195] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proc. of CCS*, 2017.

[196] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proc. of CVPR*, 2017.

[197] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proc. of NAACL*, volume 2, 2018.

[198] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proc. of EMNLP*, pages 4847–4853, 2018.

[199] Kun Zhou, Berrak Sisman, and Haizhou Li. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv preprint arXiv:2002.00198*, 2020.

[200] Kun Zhou, Berrak Sisman, and Haizhou Li. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. *arXiv preprint arXiv:2103.16809*, 2021.

[201] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *arXiv preprint arXiv:2105.14762*, 2021.

[202] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2021.

[203] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv preprint arXiv:2005.07025*, 2020.

APPENDIX A

SURVEYS FOR MEASURING GENDER STEREOTYPES

A.1 Measuring Word and Gender Associations

A.1.1 Task 1

On the following pages, you will be shown lists of adjectives. We are trying to determine **the extent to which people think each adjective is associated with men or with women.**

Your task is to evaluate the extent to which you associate each adjective shown with a typical man or woman.

In other words, if you think the adjective it is typically used to characterize a man (or a woman), then it is considered to be associated with a man (or a woman).

You will evaluate each adjective for its association with a man or woman, but not both.

Just to make sure you understand the task, here is an example: If the word shown is "square" and you strongly disagree that this word is typically associated with a man, then you would answer 1 on the scale below.

- Strongly agree
- Somewhat agree
- Slightly agree
- Neutral
- Slightly disagree
- Somewhat disagree
- Strongly disagree

I feel that _____ is commonly associated with the characterization of a typical MAN.

Show a list of 50 adjectives, rate agreement in 7 point Likert Scale

I feel that _____ is commonly associated with the characterization of a typical WOMAN.

Show a list of 50 adjectives, rate agreement in 7 point Likert Scale

A.1.2 Task 2

Your task is to evaluate **the extent to which you associate each VERB shown with a typical man or woman.** This is similar to Task 1.

In other words, if you think the verb is typically used to characterize a man (or a woman), then it is considered to be associated with a man (or a woman)

Please evaluate the verb as though a man (or a woman) is the **subject** (i.e., the person doing the action). Just to make sure you understand the task, here is an example: If the verb shown is "ran", imagine the following sentence: "The man ran away from the fox." We would like you to evaluate how strongly you associate the verb with respect to the **man**, not the fox.

You will evaluate each verb for its association with a man or woman, but not both.

I feel that _____ is commonly associated with the characterization of a typical MAN.

Show a list of 50 verbs, rate agreement in 7 point Likert Scale

I feel that _____ is commonly associated with the characterization of a typical WOMAN.

Show a list of 50 verbs, rate agreement in 7 point Likert Scale

A.2 Crowdsourcing Gendered Text from Online Sources

The goal of this study is to understand **how people understand gender stereotypes**. Your task is to find articles on the Internet **that describe people who exemplify or contradict common gender stereotypes of US society**. Here, *gender stereotypes* refer

to general impressions that society places on men and women, **not necessarily stereotypes that you hold yourself** for men and women.

For example, if you think men are *stereotypically* viewed in society as powerful and rich, and you are asked to find an article about stereotypical men, you might want to look for articles that describe a male Wall Street trader.

The article can come from news webpages, Wikipedia pages, online social media, obituaries, self-description in public resumes, etc. You will need to choose 4 different articles, and we will pay you $0.75 for each article. You will be given instructions for each article that we want you to find. Please read these instructions carefully each time because they may vary for each article.

Please search the web and choose an article so that:

1) The article contains description of a ***man*** *(or* ***woman****).* It could be related to his (or her) appearance, characteristics, habit, or a story about the person.

2) His *(or Her)* characteristics and behaviors are ***consistent*** *with (or* ***contradictory to****)* a typical man (or woman), according to stereotypes in society.

- Please copy and paste the paragraphs that are related to the person (if only a few paragraphs are related to the person, please only copy and paste these paragraphs).
- Please also paste the link to the webpage where you found the article.
- Please briefly state why the person in the paragraphs is consistent or contradicts common gender stereotypes.

APPENDIX B

EMOTION TONE USER STUDY QUESTIONNAIRES

B.1 Measuring Feasibility of Altering Voices

You will be presented with a series of short audio clips, and asked to rate each clip on a series of characteristics along a scale. You may replay each clip as many times as you would like. For each clip, only consider the **tone of voice**, not the content of the sentence.

1. Please rate the overall quality of the audio:
    - 1: very good quality (imperceptible distortion)
    - 2: good quality (perceptible but not annoying distortion)
    - 3: decent quality (perceptible and slightly annoying distortion)
    - 4: poor quality (annoying but not objectionable distortion)
    - 5: very poor quality (very annoying and objectionable distortion)
2. Please indicate which emotion(s) you perceive from the speech. Mark all that apply.
    - neutral
    - angry
    - happy
    - sad
    - fear
    - surprise
    - other _____

For each selection made in the previous question, we asked:

3. Please indicate the extent you perceive the emotion _____

    - definitely _____
    - probably _____
    - somewhat _____
    - neutral

Now we will present several pairs of audio clips, and ask several questions about the pairs. You may replay the clips as many times as you like. Remember to **focus only on the tone** and ignore the content of the speech.

For each question below, the participants were presented the following choices:

- Definitely #1
- Somewhat #1
- Neutral
- Somewhat #2
- Definitely #2

- Please indicate which voice sounds more **trustworthy**
- Please indicate which voice sounds more **competent**
- Please indicate which voice sounds more **warm**
- Please indicate which voice sounds more **anxious**
- Please indicate which voice sounds more **confident**
- Please indicate which voice sounds more **polite**
- Please indicate which voice sounds more **positive**

- Please indicate which voice sounds more **negative**
- Suppose you were talking on the phone to a telephone operator (e.g., bank, airline, cable company). Which voice would you prefer to speak with on the other end of the line?
- Suppose you were watching a political debate. Which voice would you prefer to vote for?
- Suppose you were interviewing personal assistants for yourself. Which voice would you prefer to hire as your assistant?
- Suppose you were going to have surgery. Which voice would you prefer to be your surgeon?
- Suppose you were going to have a job interview. Which voice would you prefer to interview you?

B.2 Measuring Desirability of Altering Voices

Suppose there is a software tool that can change one's tone of voice. For instance, imagine speaking to someone at a call center (e.g., a support member for your bank or Internet provider), and this software could make either your own voice or the other person's voice sound less angry.

Alternately, suppose you are giving a presentation online (i.e., Zoom) and that you have the opportunity to filter your voice with a software to make you sound less anxious (more confident).

- Would you use such a tool to alter your voice?
  - Definitely, very often
  - Sometimes
  - Rarely
  - Definitely not, ever

- Under what circumstances would you consider it **acceptable or desirable** to use such a tool? _____
- What circumstances would you consider it **unacceptable or undesirable** to use such a tool? _____
- Would you consider it morally acceptable to use this tool to alter your voice **if others were unaware** that you were using it, if it improved the effectiveness of the speaker?
  - Totally acceptable
  - Sometimes acceptable
  - Occasionally acceptable
  - Rarely acceptable
  - Totally unacceptable
  - Under what types of circumstances would this be acceptable or unacceptable? Please explain your reasoning. _____
- Would you consider it morally acceptable if this tool was used to alter someone's voice **without the speaker's knowledge**, if it improved the effectiveness of the speaker?
  - Totally acceptable
  - Sometimes acceptable
  - Occasionally acceptable
  - Rarely acceptable
  - Totally unacceptable
  - Under what types of circumstances would this be acceptable or unacceptable? Please explain your reasoning. _____

- Suppose the tool was used to **reduce anger** of the speaker, such as an angry customer speaking to an operator at a call center. Would you find the tool to be morally acceptable?
  - Yes, definitely
  - Only if the speaker was aware of the tool being used
  - Only if the audience was aware of the tool being used
  - Only if all parties were aware the tool being used
  - Definitely not acceptable under any circumstances
  - Please explain your reasoning. _____
- Suppose the tool was used to **reduce anxiety** of the speaker, such as helping a presenter sound more confident. Would you find the tool to be morally acceptable?
  - Yes, definitely
  - Only if the speaker was aware of the tool being used
  - Only if the audience was aware of the tool being used
  - Only if all parties were aware the tool being used
  - Definitely not acceptable under any circumstances
  - Please explain your reasoning. _____
- Suppose the tool was used to **increase positivity** of the speaker, such as making someone sound like they're smiling. Would you find the tool to be morally acceptable?
  - Yes, definitely
  - Only if the speaker was aware of the tool being used
  - Only if the audience was aware of the tool being used

– Only if all parties were aware the tool being used

– Definitely not acceptable under any circumstances

– Please explain your reasoning. _____

APPENDIX C

GENERATED FAKE REVIEWS

We present examples of generated reviews targeting a specific restaurant (description shown below). The samples are generated under different temperatures and from the models trained using five-star, three-star and one-star Yelp training review corpus.

Target Restaurant Description: Experience MKT, a market-to-table San Francisco restaurant and bar. Taste your way through ultra-creative handcrafted cocktails, small plates for sharing, and steaks from the finest California ranches. Meet friends and relax in our retro urban-chic atmosphere. In the heart of SoMa - with panoramic fifth-floor views of Market Street and downtown - this is one of San Francisco's best restaurant venues for social drinks and dining.

C.0.1 Machine-generated One-Star Reviews

Generated with *temperature* 0.1

I was so excited to try this place out for the first time and the food was awful. I ordered the chicken sandwich and it was so salty that I could not eat it. I was so disappointed that I was so disappointed in the food. I was so disappointed that I was so disappointed with the service.

Generated with *temperature* 0.3

I had a drink at this restaurant. I was there for a couple of months. The restaurant was clean but the food was not good. The service was very slow and the service is so bad. I would not recommend this place to anyone.

Generated with *temperature* 0.5

The worst place I've ever been too. I waited for a while and the place was not available and the service was absolutely terrible. The server was rude and not apologetic and unprofessional. I will not be going back to this location.

Generated with *temperature* 0.7

I have been here twice, very expensive for what you get. We stayed here before at our table and were seated in a line for 15 minutes and the restaurant was full and the cashier recommended the drinks and they forgot the cookie cover charge and we were told it was not an issue. I would never recommend this place to anyone.

Generated with *temperature* 1.0

DO NOT WASTE YOUR TIME AND MONEY! The absolute worst service I have ever experienced. This place is a joke. The waitress was rude and said she would put the manager to come out but never happened. I wish I could give zero star.

C.0.2 Machine-generated Three-Star Reviews

Generated with *temperature* 0.1

The food was good but not great. I had the lasagne and the sauce was a little too sweet for my taste. The salad was good but the sauce was a little too sweet for my taste. The sauce was a little too salty for my tastes. The sauce was a little too sweet for my taste.

Generated with *temperature* 0.3

I've had the fried chicken and the chicken breast with chicken soup with sweet potato fries. I don't know what a lot of the meat was on the salty side as well. The sauce was a bit salty and didn't have much flavor. The salad was tasty and well presented but the bread was super tasty.

Generated with *temperature* 0.5

I was here for a weekend brunch and the food was ok. I love the pizza that is a chain restaurant. I think the service is excellent. I had the spaghetti and they were very good and the hot dog was good. I got the red velvet chocolate cake special which was very good but the service was a little slow. The food was good, but not up to par with other places nearby.

Generated with *temperature* 0.7

The food wasn't bad. The cupcakes are okay and the service is excellent but the prices are a bit high. I do like the fresh made salad and drink specials. I would recommend this place for a place to grab a bite for a couple of times.

Generated with *temperature* 1.0

Came here for lunch today and the place was pretty empty. The steak was good but the chicken they had a little less oily and overcooked. I would recommend this place if you are looking for a cheap place to stop by.

C.0.3 Machine-generated Five-Star Reviews

Generated with *temperature* 0.1

I have been going to this place for a few years now and I have never had a bad experience. The service is great! They are always so friendly and helpful. I will definitely be back and I will be back for sure!

Generated with *temperature* 0.3

This place is amazing! The bartenders are absolutely amazing. The pasta is delicious and I love their pastries and it is amazing. I love the breakfast, friendly staff and the price is very reasonable. I have never had a bad experience here. I will be back for sure!

Generated with *temperature* 0.5

I love this place. I went with my brother and we had the vegetarian pasta and it was delicious. The beer was good and the service was amazing. I would definitely recommend this place to anyone looking for a great place to go for a great breakfast and a small spot with a great deal.

Generated with *temperature* 0.7

I have been a customer for about a year and a half and I have nothing but great things to say about this place. I always get the pizza, but the Italian beef was also good and I was impressed. The service was outstanding. The best service I have ever had. Highly recommended.

Generated with *temperature* 1.0

The food here is freaking amazing, the portions are giant. The cheese bagel was cooked to perfection and well prepared, fresh & delicious! The service was fast. Our favorite spot for sure! We will be back!

C.1 User Study Surveys

C.1.1 Fake/Real Review Detection

Figure C.1 shows a screenshot of the survey designed for examining human performance on machine-generated review detection.

C.1.2 Review Helpfulness Rating

Figure C.2 shows another screenshot of the second round survey designed for collecting the helpfulness score of the machine-generated reviews.

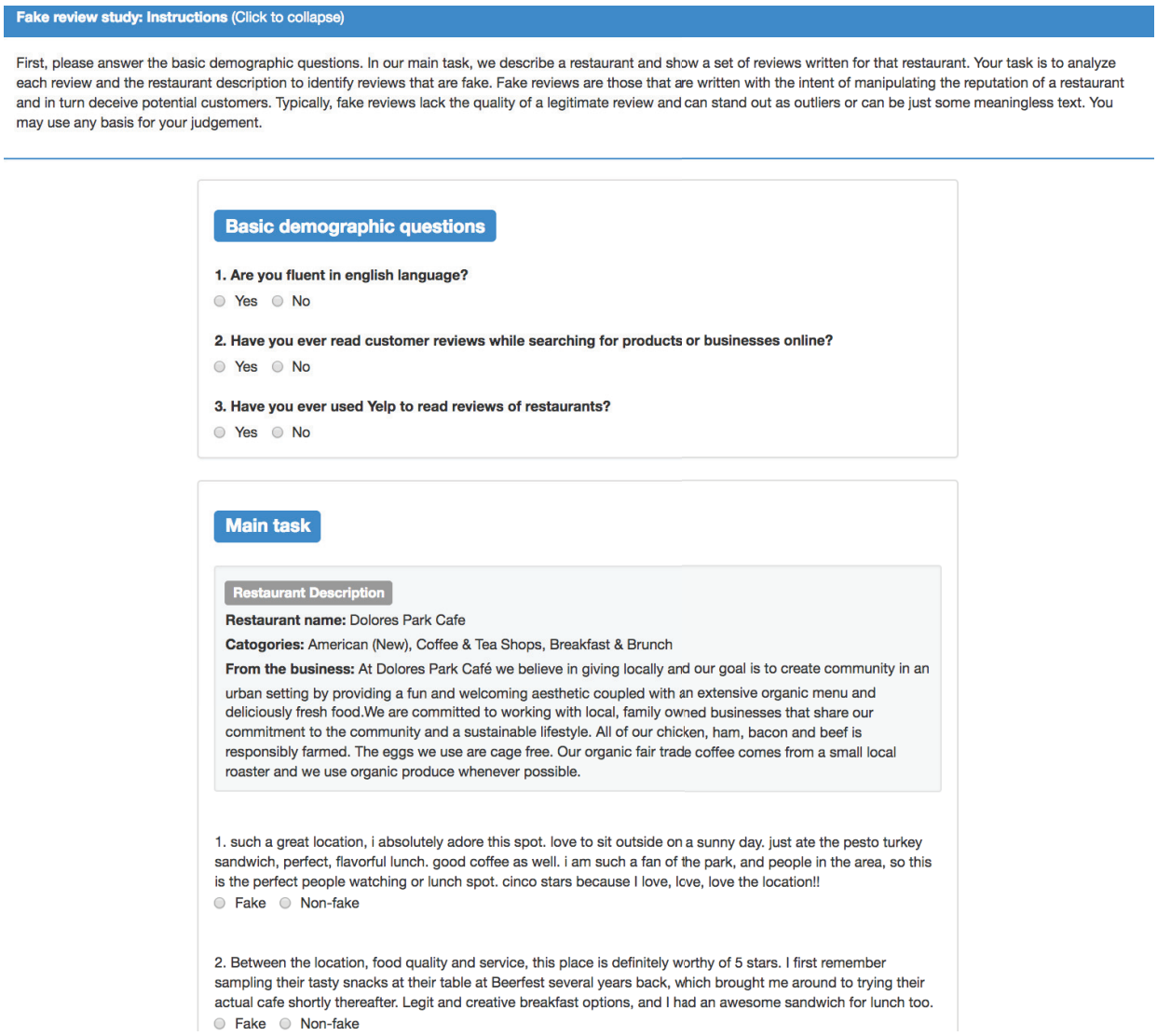


Figure C.1: Examining human performance on machine-generated review detection.

C.2 Detailed Description of Our Face Datasets

Evaluation of *Aletheia* requires a dataset covering a wide range of face edit types and tools. As no existing datasets provides edit type labels, we built our own dataset by altering real face images with editing tools using automatic scripts.

- *Original faces* – By combining several public datasets (see Table C.1), we built a dataset containing 821,358 face images of >30,461 identities. This combined set ensures

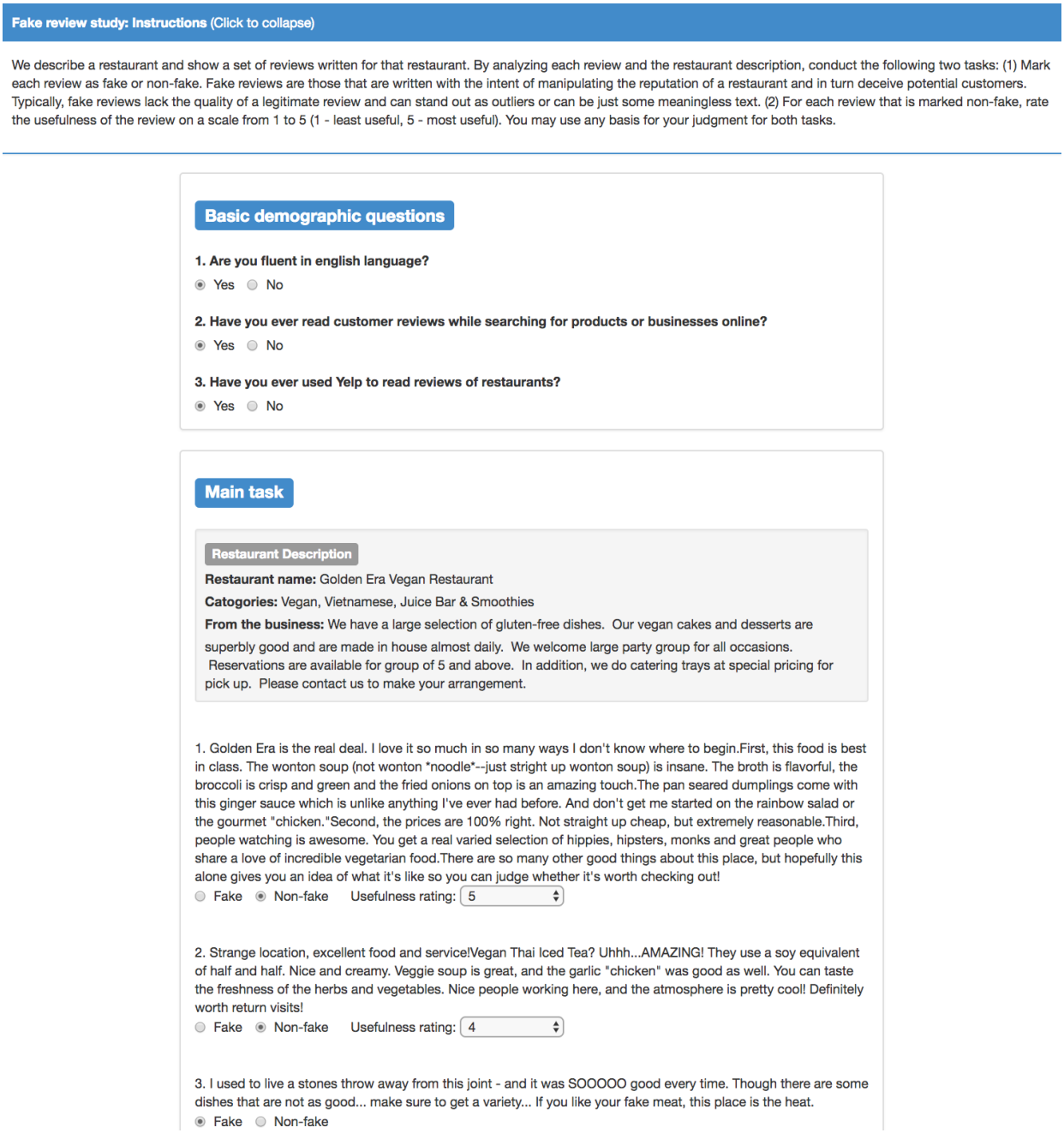


Figure C.2: Collecting helpfulness rating of the machine-generated reviews.

diversity and includes a wide variety of images from celebrities and normal people. For consistency, each image only contains a single face.

- *Edited faces* – We built and ran scripts to generate edited images from 1000 original face images (randomly sampled), producing 42.5K edited images labeled by the edits.

Sub dataset	# of identities	# of images	Type (Source)
CelebA [100]	10,177	202,599	celebrities (Internet)
FFHQ [85]	unknown	70,000	normal people (Flickr)
DeeperForensics [78]	unknown	1,000	faces (YouTube video, using the first frame)
FF++ [148]	unknown	1,000	faces (YouTube video, using the first frame)
IMDB-WIKI [149]	20,284	523,051	actors (IMDb, Wikipedia)
UTKFace [196]	unknown	23,708	faces (Internet)
Total	**>30,461**	**821,358**	**normal people & celebrities**

Table C.1: Our **original face** dataset includes 820K+ face photos from both normal people and celebrities.

Category	Edit type	# of images	Edit tools
Global processing	Add filter	3,941	FaceApp, PortraitPro
	Change brightness	5,936	PhotoShop, PortraitPro
Modify facial attributes	Change age	4,059	FaceApp, StarGAN, HRFAE
	Change gender appearance	2,000	AttGAN, StarGAN
	Change face shape	2,954	PhotoShop, PortraitPro
	Change skin tone	2,376	AttGAN, PortraitPro
	Change hair color	2,486	FaceApp, StarGAN
	Resize eye/nose/mouth	7,914	PhotoShop, FaceAPP, PortraitPro
	Add makeup	4,925	FaceApp, PortraitPro
Change facial expression		1,967	FaceAPP, GANimation
Add/Remove Eyeglasses		1,989	FaceApp, OpenCV code
Change face identity (facewap)		2,000	FF++, DeeperForensics
12 Edit Types		**42,547**	**10 Edit Tools**

Table C.2: **Edited faces:** we generated and labeled more than 42K edited images, covering 12 popular face editing types and 10 popular edit tools (3 commercial and 7 open-source tools).

As detailed in Table C.2, our dataset covers 12 edit types and 10 editing tools, including both commercial software/apps (PhotoShop, PortraitPro, FaceApp) and open-source models (StarGAN, AttGAN, GANimation, HRFAE, OpenCV sticker code, FF++, DeeperForensics 1.0). Each image contains a single type of edit. For each edit type, we generate edited images using at least two different tools. Due to variations in both the number of available tools and their edit options, our edited face dataset is not balanced across edit types. To avoid bias, we up-sampled under-represented types when reporting results that aggregate over edit types.

APPENDIX D

PERCEPTIONS OF FACE EDITS USER STUDY

D.1 Personalization User Study Details

Context Establishment. Suppose you're sharing a photo online to a platform, similar to Facebook or Instagram. Similar to when you post pictures to these online platforms, other people who can view the picture may edit your photo and upload it to the same platform. Some people may do this for fun (e.g., add fun stickers). Other people may do this maliciously (e.g., cyberbullying).

This platform detects if an image has been edited and re-uploaded. When you upload an image, you can specify a set of preferences associated with the image. Each preference setting either allows or disallows a particular type of editing. After an image is uploaded, the platform can detect and remove any of your pictures that have been edited in a way that violates your current settings.

D.1.1 Edit Type Preference

Q1. Here are some examples of **brightness change**. Suppose the brightness can be adjusted along a spectrum, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.1).

Q2. Here are an example of **adding a filter**. Suppose the filters may be different colors, and can be adjusted along a spectrum. Indicate to what extent you would typically allow this type of editing (see Figure D.2).

Q3. Here are some examples of **adding stickers**. Indicate which of these edits you would typically allow (see Figure D.3).

Q4. Here are some examples of **changing facial attributes (face shape, eyes/nose/mouth**

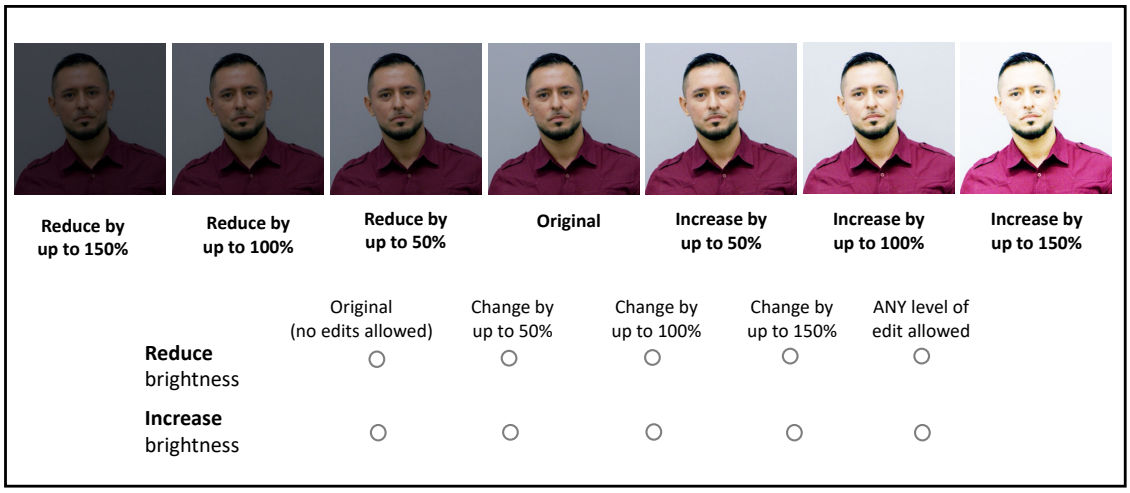


Figure D.1: Question about preferences for **brightness change**.

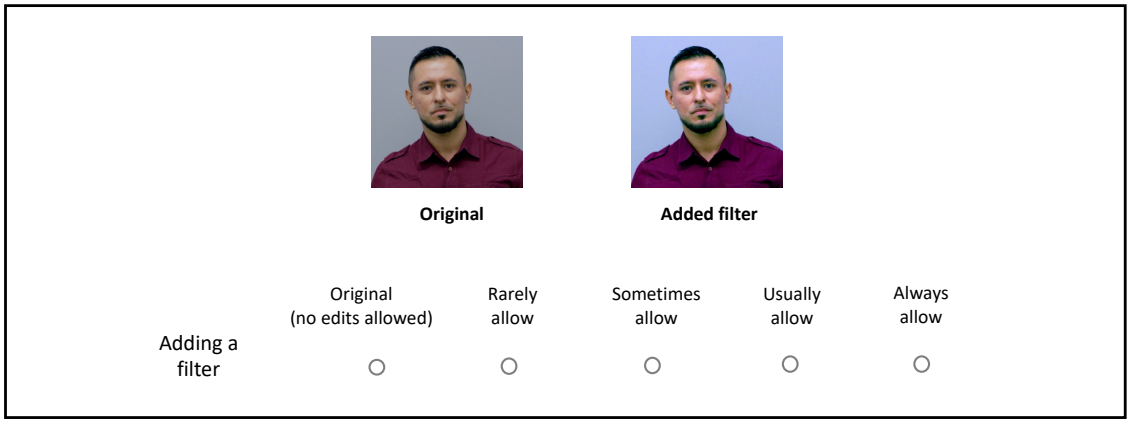


Figure D.2: Question about preferences for **adding a filter**.

size). Suppose these attributes can be adjusted along a spectrum, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.4).

Q5. Here are some examples of **changing hair color/style**. Suppose the edits may be different styles or colors, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.5).

Q6. Here is an example of **adding makeup**. Suppose the edits may be different styles or colors, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.6).

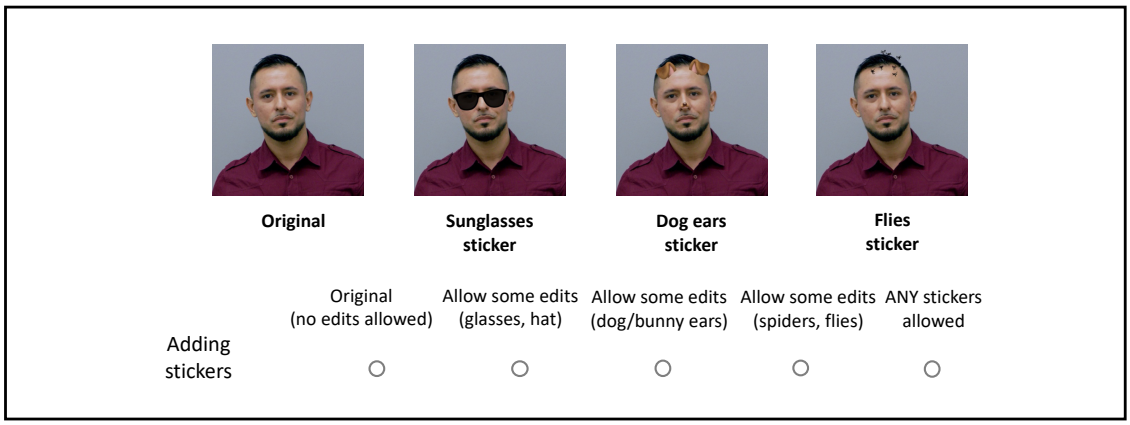


Figure D.3: Question about preferences for **adding stickers**.

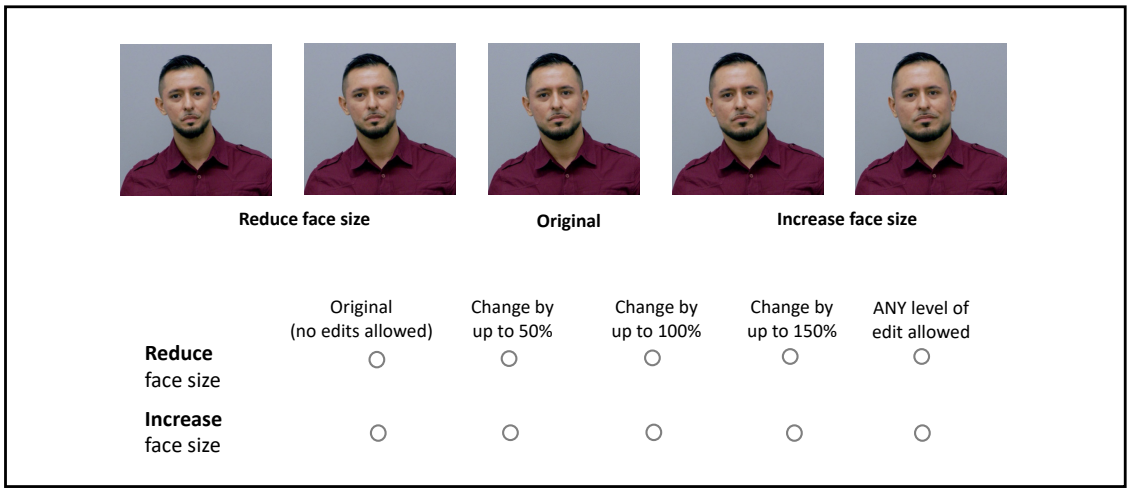


Figure D.4: Question about preferences for **changing facial attributes**.

Q7. Here are some examples of **face swapping**. Indicate to what extent you would typically allow this type of editing (see Figure D.7).

Q8. Here are some examples of **changing facial expression (enhance/reduce smile/frown)**. Suppose these expressions can be adjusted along a spectrum, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.8).

Q9. Here is an example of **gender appearance change**. Indicate to what extent you would typically allow this type of editing (see Figure D.9).

Q10. Here are some examples of **age change**. Suppose the age can be adjusted along a

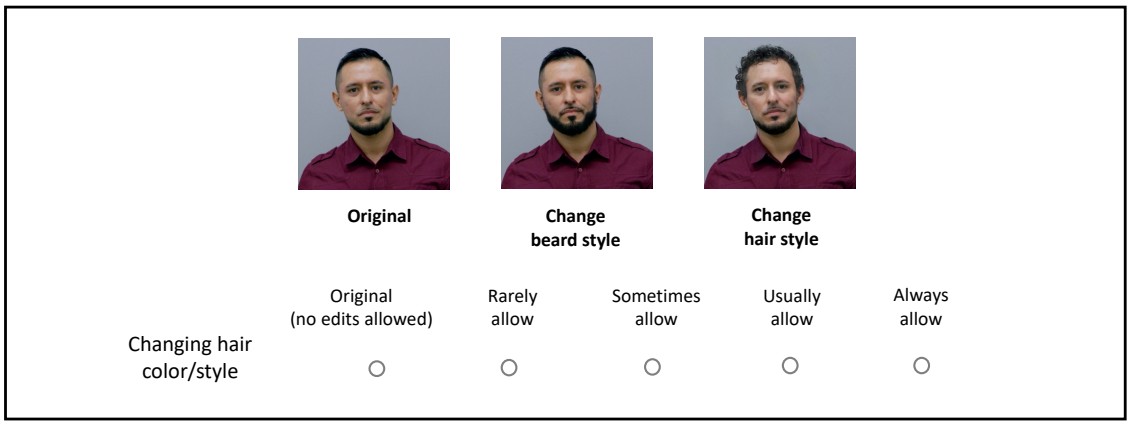


Figure D.5: Question about preferences for **changing hair color/style**.

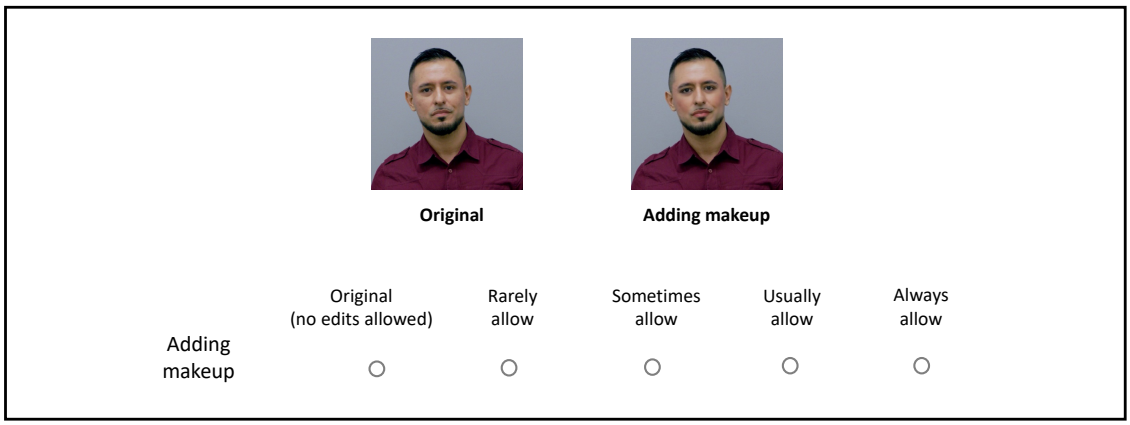


Figure D.6: Question about preferences for **adding makeup**.

spectrum, similar to the example below. Indicate to what extent you would typically allow this type of editing (see Figure D.10).

D.1.2 Acceptable Image Selection

Below is an original image, followed by various edited images. Which of the following edited images would you allow (see Figure D.11)? Select all images that apply.

D.1.3 Privacy Setting Preference

Q1. Now you have seen both scenarios, would you change any of your responses from scenario 1 (viewable to only friends and family) after considering scenario 2 (viewable to the

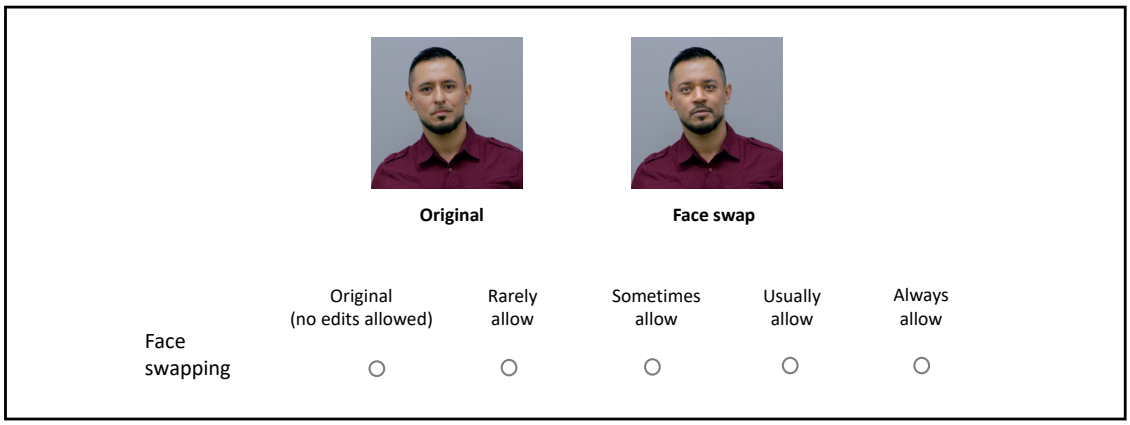


Figure D.7: Question about preferences for **face swapping**.

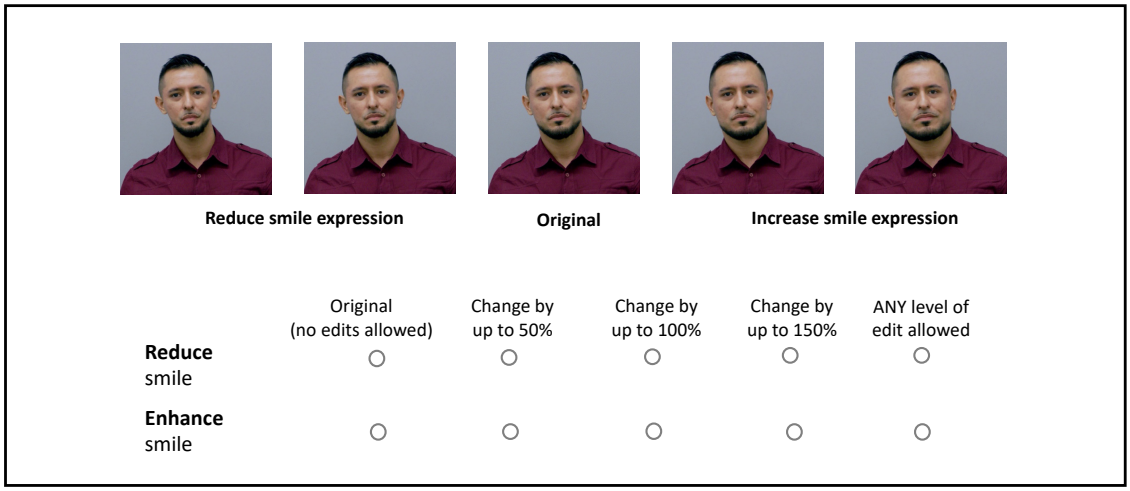


Figure D.8: Question about preferences for **changing facial expression**.

public)?

- Yes, I would accept MORE edits in the scenario viewable only to friends and family
- Yes, I would accept LESS edits in the scenario viewable to friends and family
- No, I would NOT change my responses

Q2. Please explain your reasoning. _____

As described earlier, suppose we are building such an online social platform with face-editing detection. When you upload an image, you can specify a set of preferences associated with the image. Each preference setting either allows or disallows a particular type of editing,

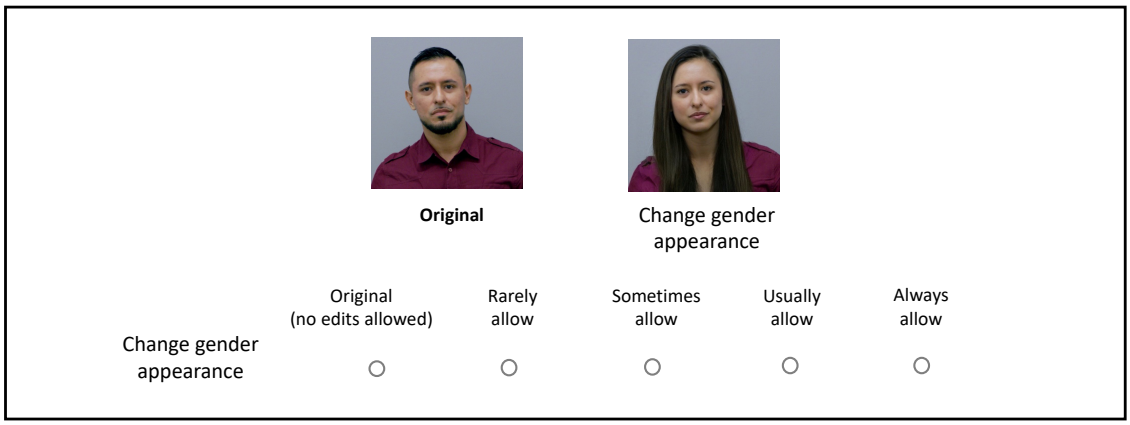


Figure D.9: Question about preferences for **changing gender appearance**.

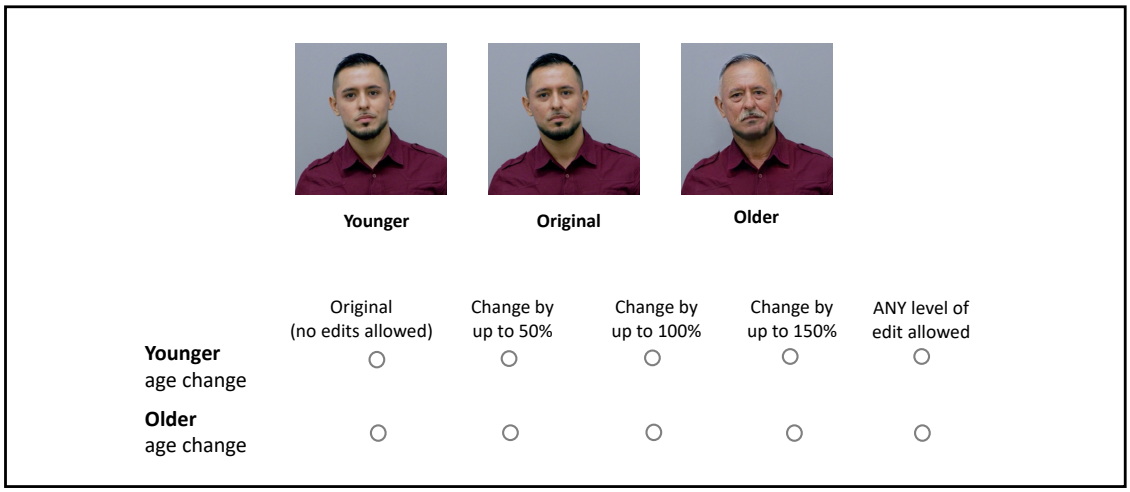


Figure D.10: Question about preferences for **changing age**.

like those you just saw. After the image is uploaded, the platform will detect and remove any edited images (based on your image) that violate your current settings.

Q3. Assume you can choose to either make your image editing settings private or public. Choosing a private policy means no one will know the details of your image editing settings. Choosing a public policy means everyone can know your image editing settings when they view the image. Which policy would you most likely choose?

- 0 - Definitely private
- ...

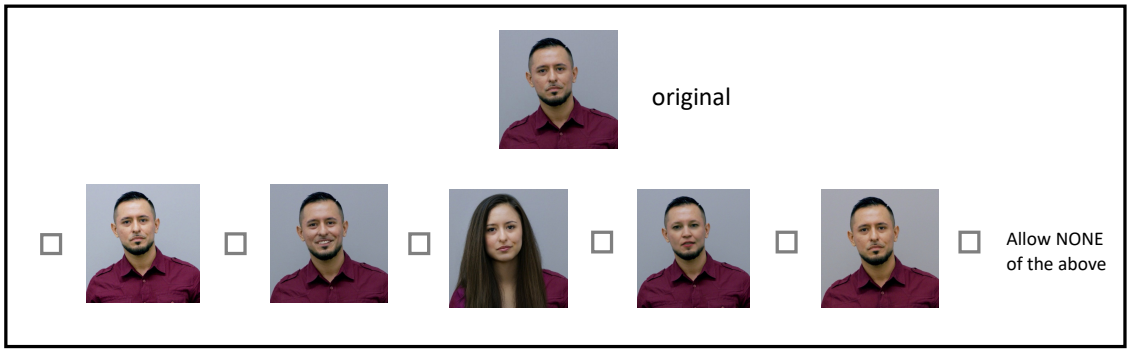


Figure D.11: We showed participants new examples of edited images at once and ask which they would (dis)allow.

- 5 - Neutral
- ...
- 10 - Definitely public

Q4. Under which of the following scenarios would you want to change your image editing settings? Select all that apply.

□ When you know that your friend(s) changed to a different policy

□ When you see an editing method is reported negatively by the news

□ When new editing methods are invented or popularized

□ When you see unwanted edited image not covered by your current policy

□ Other _____

Q5. If the platform you shared your image on detects an edited image of yours that violates your policy, what would you want the platform to do?

- The platform should directly block the image
- The platform should send you a notification, then let you decide

- The platform should block the image, then send me a notification
- Other _____

Q6. Which of the following matters more to you?

- The platform can identify as many disallowed edits as possible, but may falsely label some normal edits as disallowed
- The platform is highly accurate when it flags a disallowed edit, but may miss some disallowed edits

Q7. How frequently **do you edit** other people's face(s) in images/videos posted online?

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely or never
- N/A (I'm not active on online social platforms)

Q8. How frequently **do you upload images/videos** showing face(s) of people to online social platforms?

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- Never (I don't post images/videos online)

Q9. How frequently **do you observe such editing** in images/videos posted online?

- Very often (at least once a day)

- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- N/A (I'm not active on online social platforms)

Q7-9. How frequently do you _____?

- **Q7.** edit other people's face(s) in images/videos posted online
- **Q8.** upload images/videos showing face(s) of people to online social platforms
- **Q9.** observe such editing in images/videos posted online

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- N/A (I'm not active on online social platforms)

Q10. To what extent **are you concerned that other people may edit** the face in an image/video you posted and then repost it themselves?

- Very concerned
- Somewhat concerned
- Not concerned
- N/A (I don't post images/videos online)

Q11. How would you generally describe your habits online?

- 0 - I don't feel concerned about online privacy

- ...
- 5 - Neutral
- ...
- 10 - I feel very concerned about online privacy

D.2 Perceptions of Aletheia User Study

Context Establishment. Suppose you're sharing a photo online to a platform, similar to Facebook or Instagram. Similar to when you post pictures to these online platforms, other people who can view the picture may edit your photo and upload it to the same platform. Some people may do this for fun (e.g., add fun stickers). Other people may do this maliciously (e.g., cyberbullying).

This platform has the ability to detect if an image has been edited and re-uploaded. When you upload an image, you can specify a set of preferences associated with the image. Each preference setting either allows or disallows a particular type of editing. The types of edits this system can detect include:

- Change brightness
- Add filters
- Add stickers
- Change face shape/size
- Change hair style (including beard, hair color, etc)
- Adding makeup
- Face swap

– Change facial expression (i.e., smile/frown)

– Change gender appearance

– Change age

– Lighten/darken skin

After an image is uploaded, the platform can detect and remove any of your pictures that have been edited in a way that violates your current settings.

To define your own policy, the system would show you some examples of edited images for each type of edit and ask you which amount of editing you would allow, as shown (in Figure D.12):

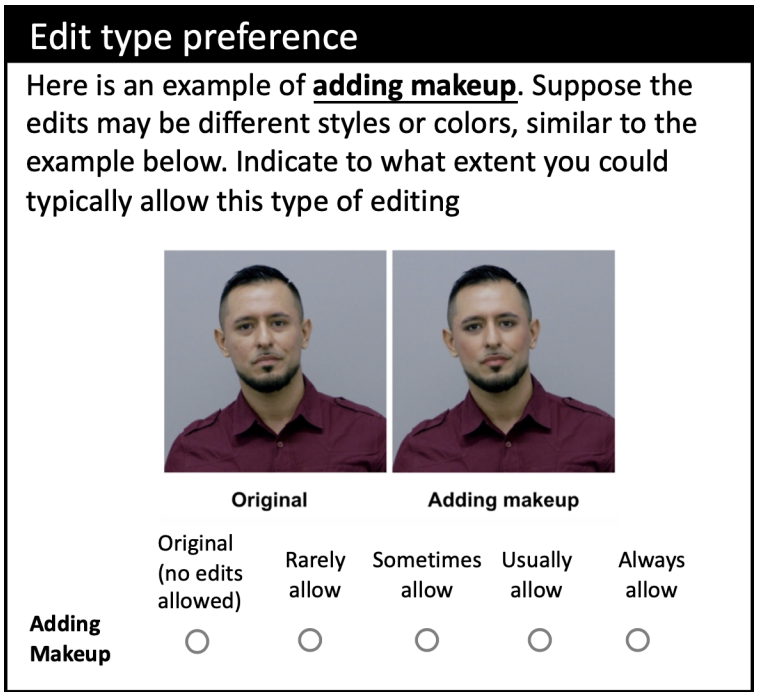


Figure D.12: Example question shown to participants to illustrate how users of Aletheia specify their edit preferences.

Now we would like to ask you some questions about your opinions on this platform. For these questions, we are assuming the images contain one or more faces. Suppose this system has been deployed for use on social media platforms.

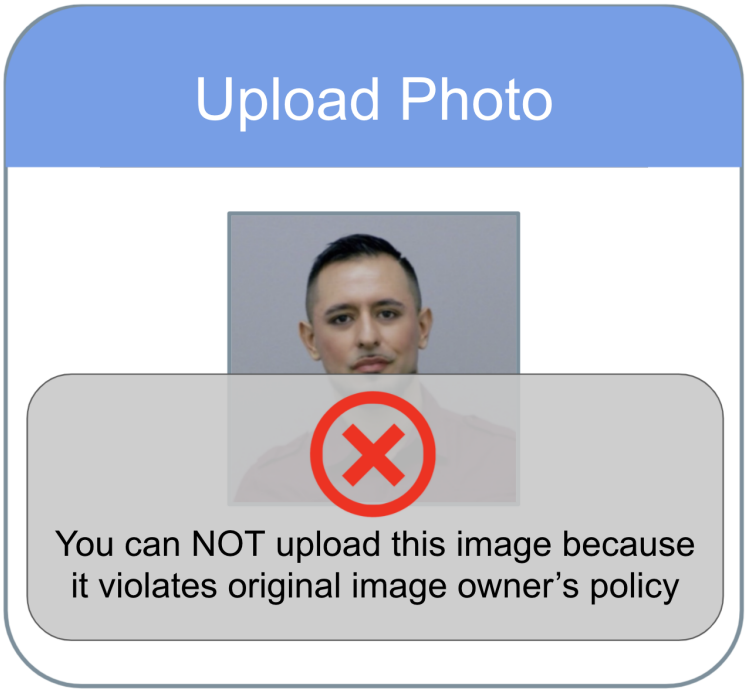


Figure D.13: Alethia blocks the image upload because it violates the original image's policy.

Q1. To what extent would you feel your images are protected by this system when posting images online?

- Definitely not protected
- Probably not protected
- Neutral
- Probably protected
- Definitely protected
- Please briefly explain your reasoning _____

Q2. How would you feel about setting your own policy, with regards to the tradeoff for time spent setting your policy compared to the protection offered?

- Definitely not interested (too time consuming)

- Probably not interested
- Neutral
- Probably interested
- Definitely interested (worth the protection)
- Please briefly explain your reasoning _____

Q3. Would you prefer to set a general policy for all images or specify your policy for each image?

- I would prefer to define different policies for each image
- I would prefer a general policy for all images
- Please briefly explain your reasoning _____

Q4. Many online image sharing sites allow for different levels of privacy for who can view an image (i.e., viewable to the public, or viewable to a private group). Would you prefer different policies for different privacy groups?

- Definitely not (same policy for all groups)
- Probably not
- Neutral
- Probably yes
- Definitely yes (different policies for different groups)
- Please briefly explain your reasoning _____

Q5. Which would be more important when posting images?

- Protect all faces in an image
- Protect my face in an image
- Please briefly explain your reasoning _____

Q6. Under which of the following scenarios would you want to change your image editing settings? Select all that apply.

- □ When you know that your friend(s) changed to a different policy
- □ When you see an editing method is reported negatively by the news
- □ When new editing methods are invented or popularized
- □ When you see unwanted edited image not covered by your current policy
- □ Other _____

Q7. If the platform you shared your image on detects an edited image of yours that violates your policy, what would you want the platform to do?

- The platform should directly block the image
- The platform should send you a notification, then let you decide
- The platform should block the image, then send me a notification
- Other _____

Q8. Which of the following matters more to you?

- The platform can identify as many disallowed edits as possible, but may falsely label some normal edits as disallowed

- The platform is highly accurate when it flags a disallowed edit, but may miss some disallowed edits

Q9. Would you use this system when posting images on social media sites?

- Definitely not
- Probably not
- Neutral
- Probably yes
- Definitely yes
- Please briefly explain your reasoning _____

Q10. If you were to design your own system for determining (un)allowable edits to your images posted online, what changes would you make to the system previously described?

- _____

Q11. How frequently **do you upload images/videos** showing face(s) of people to online social platforms?

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- Never (I don't post images/videos online)

Q12. How frequently **do you observe such editing** in images/videos posted online?

- Very often (at least once a day)

- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- N/A (I'm not active on online social platforms)

Q13. How frequently **do you edit** other people's face(s) in images/videos posted online?

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely or never
- N/A (I'm not active on online social platforms)

Q11-13. How frequently do you _____?

  - **Q11.** upload images/videos showing face(s) of people to online social platforms
  - **Q12.** observe such editing in images/videos posted online
  - **Q13.** edit other people's face(s) in images/videos posted online

- Very often (at least once a day)
- Somewhat often (a few times a week)
- Occasionally (a few times a month)
- Rarely
- N/A (I'm not active on online social platforms)

Q14. To what extent **are you concerned that other people may edit** the face in an image/video you posted and then repost it themselves?

- Very concerned

- Somewhat concerned
- Not concerned
- N/A (I don't post images/videos online)

Q15. How would you generally describe your habits online?

- 1 - I feel very concerned about online privacy
- 2
- 3 - I usually rely on default settings
- 4
- 5 - I don't feel concerned about online privacy