BEYOND FAIRNESS METRICS: HUMAN EXPERIENCES AND DIFFERENTIAL OUTCOMES IN AI-MEDIATED WORK SETTINGS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kowe Kadoma

May 2026

BEYOND FAIRNESS METRICS: HUMAN EXPERIENCES AND DIFFERENTIAL OUTCOMES IN AI-MEDIATED WORK SETTINGS

Kowe Kadoma, Ph.D.

Cornell University 2026

AI systems are increasingly embedded into the everyday infrastructure of modern workplaces, from resume screening systems to AI features built directly into productivity tools. Many AI tools—such as automatic speech recognition (ASR) tools or writing assistants—do not make explicit decisions. Yet, these tools can meaningfully influence how people are evaluated by mediating how work is produced and communicated. Understanding these effects requires expanding beyond the narrow scope of much existing fairness research, which has focused primarily on reducing model errors, mitigating bias in training data, and optimizing performance metrics between groups. To build AI systems that are equitable and inclusive, it is equally essential to understand the human experience—how AI tools shape what people see, how they evaluate others, and how individuals feel seen within AI-mediated environments.

This dissertation investigates how AI tools can produce differential outcomes even when they are not explicitly making decisions. In the first study, we examine how the audience evaluates the speaker and their content when there are inaccurate ASR generated subtitles, demonstrating how technical errors can quickly become social judgments. The second study illustrates how the mere suspicion of using an AI tool—regardless of whether it was actually used—can negatively influence job performance and hiring outcomes. The third study demonstrates how the stylistic mismatches between an AI tool's suggestions and a user's voice can undermine

feelings of inclusion, control, and ownership.

Taken together, these studies show that AI systems function not merely as productivity aids but as sociotechnical infrastructure that shapes perception, evaluation, and lived experience while completing work tasks. By foregrounding human experiences, this dissertation identifies harms that remain invisible to model-centric evaluations and contributes a framework for supporting agency and fostering inclusion. In doing so, it charts a path toward more equitable, accountable, and human-centered measurements for AI tools.

BIOGRAPHICAL SKETCH

Kowe Kadoma is a doctoral candidate at Cornell University where she is a member of the Social Technologies Lab. She leverages quantitative methods to characterize novel harms in AI systems and proposes frameworks to create safe, trustworthy, and inclusive AI. Prior to Cornell, Kowe received an undergraduate degree in Computer Engineering from Florida A&M University. Her work is supported by a LinkedIn Fellowship, GEM Fellowship, and Digital Life Initiative (DLI) Fellowship.

ACKNOWLEDGEMENTS

There are so many people to thank, and this acknowledgments section is my best attempt to (briefly) express my gratitude toward everyone who has supported me. First and foremost, I would like to thank my committee, and in particular my advisor, Mor. Whenever I said "I don't know" (which was a lot, especially in the earlier years), he always reminded me that "if we had all the answers, there wouldn't be any research". I'm thankful for his patience when progress was slow and for his belief in my ideas, even when I didn't believe in them. I appreciate Emma's enthusiasm, cheer, and collaborative nature. From Emma, I've never underestimated the power of a cold email or simply showing appreciation for other researchers' work. A few "I really like your work" emails to researchers I admire have led to fruitful collaborations, job advice, and great friendships. I hope to one day be as brilliant as James, who can write so eloquently about both legal and technical matters. Lastly, it has been a delight to work with Danaé, who made my ideas better (and is always fun to chat with, even about non-research topics like the drama at pottery class).

I am deeply indebted to the friends I've met throughout the PhD journey. Jennifier Otiono, you always challenged me to be a more critical researcher and kinder to myself. Liz Ankrah, it is always a delight to see you at conferences, and I'm thankful for your encouragement in between. Carolina Ortega Peréz, I cannot imagine my time at Cornell Tech without you. Whether it was the conversations in the microkitchen or getting to the gym at 6:30 in the morning, each interaction with you was joyful. Sanketh Menda, I'm thankful for all the times you believed in me, even when I didn't believe in myself. Tobi Weinberg, your radical optimism and memes at obscure hours got me through many difficult moments. Ian Solano-Kamaiko, I am on the way to supporting your fiefdom. Amazia Thompson, I hope

to one day have the faith that you have. Your support, prayers, and humor made my time much more enjoyable. And to my cohort mates, Farhana Shahid, Federica Bologna, Fenyang Lin, Daye Kang, and Daniel Mwesigwa, I wish I could have spent more time with you, but I loved every second we spent together.

There are many people who made my time at Cornell Tech enjoyable. I owe special thanks to the Social Technologies Lab, Marianne Aubine Le Quéré, Yiqing Hua, Maurice Jakesch, Sterling Williams-Ceci, Ayana Monroe, Travis Lloyd, Isabel Corpus, Andy Zhao, Lior Zalmanson, and Yotam Liel, for their support and community. Without Matt Thomas and Tyler Wilcox from the Cornell Statistical Consulting Unit, my research would not be what it is. Outside of research, I also want to acknowledge the security team at Bloomberg and in particular Jacob, with whom I had many memorable conversations, and the staff at The House, who work tirelessly to make the space welcoming. After leaving Bloomberg for the day, I often looked forward to chatting with Ramon and Rob.

Beyond Cornell, I'm thankful for the spaces that allowed me to connect with other like-minded scholars. In the Black Women in HCI Slack and the Black Doc Discord, I met Robin Brewer, Ihudiya Finda Williams, Jaye Nias, Shamika LaShawn, Jay Foriest, Judith Uchidiuno, Natalie Araujo, Kirabo, Markia Smith, Troy Kearse, Earl Huff, Milah George, Krystal Maughan, Ro Encarnación, LaToya Hogg, and Dahlia Al-Haleem. I'm grateful our paths crossed.

I appreciate those who wholeheartedly supported me—even if they never quite knew what my PhD was about. Thank you to my friends from the UES small group, Darryl Romano, Steve and Stephanie Su, Loredanna Gasparotto, Rebekah Dawn, and Jiayi Hg. Through RIR, I met Chris Abarca Lopez, JiJi Mielke, Marisara Quiñones-Ortiz, and Lavender Zhang who reminded that if I can do a PhD then marathon training isn't too bad.

I'm also grateful for my childhood friends, Nitali Arora and Elizabeth Cho. I can't begin to describe how much you two mean to me. I'm thankful for the friends I met in college. Zion Haynes, who has been my day one and gave me the title of "doctor" long before I had earned it. Aaliyah Brown, who—funnily enough—moved to NYC just two weeks before I did; I hope to face adversity as gracefully and impeccably dressed as you do. Benji Charles and Shelby Green, thank you for your friendship and love throughout this journey. I work this hard to be Theo's rich auntie. DeCarlo Carr, I can't believe after all these years, that your pep talks are just as effective. To Darryl Brooks, Jeavon White, Ama Marfo, and Dani Forrestal, I would not have passed ECE courses without you.

Lastly, I want to thank those who have known me the longest. I would not be here without the support of my parents, siblings, aunts, uncles, and grandparents who were "too supportive" in all the ways that mattered. You listened to me, laughed with me, cried with me, and prayed for me every step of the way. I hope I've made you proud.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

On November 30, 2022, OpenAI released ChatGPT to the public. What began as a novelty tool quickly became infrastructure, and within months, AI tools supported everyday workplace tasks: summarizing meetings, drafting documents, and generating ideas. By 2025, AI use had become not only widespread but expected.

Yet this normalization came with a contradiction. While employees were encouraged—sometimes implicitly required—to use AI tools to work faster and more efficiently, some organizations increasingly prohibited AI use in evaluative contexts such as hiring. Anthropic, for example, barred job applicants from using AI tools during portions of the application process, even as its own employees developed and relied on similar tools internally. The AI ban was difficult to enforce for several reasons. For example, while tools to detect AI use exist, they are unreliable and disproportionately misclassify the writing of non-native English speakers as AI-generated [121, 67]. Therefore, in the absence of reliable detection methods, enforcement often rested on suspicion rather than evidence, and groups already subject to greater scrutiny may have been disproportionately accused even as AI use became increasingly necessary to remain competitive.

This tension between when to use AI tools and who is harmed from their (alleged) use reveals a limited understanding of the impacts of AI tools in workplace settings with regards to fairness. Much of the AI fairness literature has focused on automated decision-making systems where models make binary decisions about people. For example, models that screen resumes and classify applicants as advance or reject [206, 43] or models that rank candidates in hiring pipelines [177]. Fairness, in this framing, is assessed through measurable outcomes like error rates or allocation gaps across demographic groups [88, 174]. But most AI tools used

in contemporary workplaces do not make explicit decisions. Tools like writing assistants, code-suggestion systems, or automatic captions shape impressions, expectations, and interpersonal judgments by mediating how work is produced, how competence is signaled, and how people are evaluated by others [107, 94, 51]. In these settings, harms emerge not from model outputs alone but from how people interpret and respond to AI-mediated work. This shift—from decision-making systems to workplace mediation tools—requires moving beyond traditional fairness metrics to account for harms that cannot be evaluated by the system in isolation.

Thus, I turn to sociotechnical systems research for guidance on examining how technical systems and social processes jointly shape outcomes. Drawing from human-computer interaction (HCI), science and technology studies (STS), and machine learning (ML), this literature treats AI not as an isolated artifact but as a tool embedded within organizational norms, power relations, and human judgment [174, 173, 201]. From this literature, I use Shelby's taxonomy of harms in algorithmic systems [174] to understand the landscape of potential harms and to anticipate who is affected by them. Shelby identifies five categories of harm. *Representational harms* capture the ways models can reinforce stereotypes or depict groups in limiting ways (e.g., a large language model (LLM) generating racist speech). *Allocative harms* point to unequal access to resources or opportunities as a result of model-driven decisions (e.g., a resume screening model that recommends candidates from one demographic group at higher rates). *Quality-of-service harms* arise when systems simply work better for some people than others, creating uneven experiences across users (e.g., a speech-to-text model producing more error-prone text for certain speakers). *Interpersonal harms* highlight how AI can strain trust or communication between people and communities (e.g., privacy violations when models make predictive inference beyond what users openly disclose). Finally, *so-*

cial system harms draw attention to the broader institutional and structural effects that can deepen existing inequalities over time (e.g., the energy cost of developing a model and harms from intensive water and fuel usage). By not focusing on a particular domain or technology, this taxonomy provides a valuable foundation for analyzing harms; however, its generality leaves open questions about how to anticipate impacts that arise in specific settings. Therefore, I apply this taxonomy in a domain where the effects of AI deployment are especially pronounced.

Workplace settings combine high-stakes evaluations, structural pressure [100, 193], and widespread adoption [63, 89] in ways that make it a uniquely revealing site for studying AI harms. Decisions made in and about work directly shape access to income, stability, social status, and long-term opportunity [182]; consequently, even subtle shifts in how work is produced or evaluated can result in financial and career consequences which may not be distributed evenly. Decades of organizational research demonstrate that workplace settings are already structured by persistent patterns of discrimination often along the lines of race [151, 12], gender [75], sexual orientation [171], and disability [159]. In these settings, AI tools may intensify existing disparities with workers having a limited ability to opt out of use. Prior research demonstrates that workplace technologies are rarely optional in practice, even when nominally voluntary [45, 158]. As AI tools are framed as efficiency-enhancing or as organizational "best practice", non-use can itself become a signal of deficiency or lack of competence. At the same time, expectations around appropriate AI use remain ambiguous and inconsistently enforced. These risks are amplified by the scale at which AI has been integrated into contemporary work. A Gallup poll estimates that nearly 40% of American workers are already using AI in their workflows [101]. Similarly, nearly 40% of German employees are also using AI in their workflow [63]. In both of these countries and many others these

numbers only predicted to rise [29]. Motivated by these consequences, this dissertation investigates how AI systems interact with existing organizational norms and inequalities, and how these interactions produce forms of harm that remain largely invisible to traditional fairness metrics.

Throughout this work, I advocate for moving beyond traditional fairness metrics by incorporating human-centered evaluations to capture harms that are difficult to measure in AI systems. In doing so, this dissertation enriches and complements Shelby's taxonomy in three key ways. First, I extend the concept of quality-of-service harms. Quality-of-service harms are often conceptualized as arising from direct interaction with an algorithmic system that performs unevenly across identity characteristics, resulting in increased labor, service loss, or feelings of alienation for affected users [174]. My work, in contrast, shifts attention away from how *users experience* technological failure to how an *audience interprets* those failures and how those interpretations impact the user. In doing so, I demonstrate how quality-of-service harms do not end with the user's immediate interaction with the system and can be amplified by the audience's judgments. Second, I add to the taxonomy by identifying and characterizing a new type of harm—perceptual harms. Perceptual harms are neither situated in the AI model nor in the direct interaction with it. Instead, perceptual harms are grounded in the norms and expectations surrounding AI-mediated work and who engages with it. I show how perceptual harms differentially impact groups of people. Finally, I propose a framework for inclusive AI systems which can mitigate quality-of-service harms. The framework explores how design choices between personalized models, where users feel highly included, and generic models shape the user's sense inclusion and agency during human–AI collaboration.

My contributions are most valuable to practitioners and industry stakeholders

who are responsible for evaluating and deploying AI systems. Across three empirical chapters that examine distinct AI technologies and from various perspectives, I present actionable insights to mitigate harms. I begin by investigating automatic speech recognition (ASR) systems, a technology widely used in transcription and customer-facing applications. I show that although the word error rates—the standard technical measure of accuracy—are similar across various users, the *perceived quality* of the system differs. By incorporating the perceived quality alongside the word error rate, I offer a more holistic and realistic evaluation method that captures both system behavior and the social dynamics of using the system. In the following chapter, I then examine perceptual harms in large language models (LLMs). Perceptual harms occur when the perceived use, regardless of whether AI was actually used, results in differential outcomes between social groups. I show that these harms are model-agnostic, occur in a variety of domains, and arise from social norms rather than technical failures. I identify opportunities to mitigate these harms through policy interventions and interface design, highlighting how addressing the social dynamics around AI use is just as important as responsibly developing the models themselves. In the final chapter, I examine how people use AI writing assistants to complete realistic tasks and show that highly inclusive (or more personalized) assistants can strengthen marginalized users' sense of agency. In doing so, I highlight a core design trade-off: generic assistants may scale more easily and are better suited for the task, but personalized systems can produce meaningfully better experiences for users who are often underserved. I outline the implications of these trade-offs for practitioners, emphasizing how choices about personalization shape user effectiveness and equity in deployed systems. Next, I provide a high-level summary of each chapter.

In Chapter Two, I present my work on ASR tools. Unlike prior work that

examines the causes of ASR errors for various speakers or how speakers respond to these errors, I examine how the audience evaluates speakers with various accents based on the model's performance. The findings demonstrate that even when technical metrics indicate similar system performance, audience perception of the system's output has a greater influence on user experience.

In Chapter Three, I present a framework that explores the norms and expectations around AI use. The results show that certain groups of people are suspected of using AI at higher rates; however, all groups of people are equally penalized when suspected. These findings present initial evidence of a novel, emergent harm, one that is difficult to measure with technical assessments, and one that users may not be acutely aware of.

Unlike the previous chapters which focused on how the audience evaluates a person and their work because of AI, in Chapter Four, I focus on how the user evaluates their own work as a result of AI. The findings demonstrate that individuals who felt more included by the AI assistant also experienced greater agency and ownership during the co-writing process. This effect was more pronounced among participants of minoritized genders.

In the concluding section, I discuss the core themes that have emerged across the chapters. I argue that relying on technical accuracy alone obscures important disparities and demonstrate how integrating human-centered metrics reveals forms of inequity that traditional evaluations miss. Finally, I examine the rapid rise of agentic AI systems and outline the new fairness challenges they introduce.

CHAPTER 2

THE SOCIAL CONSEQUENCES OF AI ERRORS

In many real-world settings, the exchange between a speaker and an audience is filtered through an AI system which can introduce errors into the interaction. Oftentimes, these errors are not evenly distributed across different users. Therefore, this chapter focuses on how audiences evaluate speakers with different accents when AI-mediated communication breaks down, and how those evaluations ultimately affect the user. Through an online experiment with 207 participants, we found that error-prone subtitles lowered evaluations of both the speaker and their content, regardless of accent. On the surface, this suggests no disparate treatment between accent groups once the subtitle quality is held constant. Yet this obscures a deeper inequity: because AI systems are more error prone for certain accents, speakers with those accents are more frequently exposed to the negative judgments triggered by system errors. Our exploratory analysis further shows that perceived subtitle quality varies across accents even when word error rates appear similar, revealing how benchmarks can mask socially meaningful disparities.

2.1 Introduction

Automatic speech recognition (ASR) systems convert spoken language into text using signal processing and machine learning [131, 187]. These systems now power increasingly ubiquitous technologies, from voice assistants like Apple Siri and Amazon Alexa to transcription services and automatic subtitles on platforms like YouTube, Zoom, and Microsoft Teams. Automatic subtitles facilitate accessibility, enable cross-linguistic communication, and support users in noisy environments. Beyond convenience, they play a critical role in providing equitable access to information for individuals who are deaf or hard of hearing [112, 6, 42] as well as

non-native speakers [176].

Despite their promise, ASR systems are not error-free, and their accuracy is not distributed evenly across different users. ASR systems often exhibit biases, or systematic differences in performance, with respect to race [49, 72, 109], gender [84, 189, 85], and dialect [73, 189]. For example, researchers found that the embeddings (or word associations) in the pre-trained speech models of ASR systems often associated positive words with abled individuals over those with disabilities, White Americans over Black Americans, men over women, and US accents over non-US accents [179].

These performance differences have real consequences for the individuals whose speech is being analyzed [72, 203, 119, 104]. Wenzel et. al [203] found that when ASR errors occur, Black speakers experience significantly higher levels of self-consciousness and lower self-esteem compared to White speakers. In the context of automatic subtitles for Q&A sessions in conferences, Li et. al [119] demonstrated that non-native English speakers experience frustration and confusion when subtitle errors occur. Furthermore, non-native English speakers were concerned about the potential social stigma that comes with using ASR tools. Participants in the study expressed concern that using an ASR tool could lead the audience to perceive them as less competent, hinting at the importance of how the audience perceives the speaker [119]. If an ASR tool can affect how the audience perceives the speaker, then individuals whose accents produce more ASR errors may face a compounded disadvantage. Not only are their talks more difficult to understand and less enjoyable due to subtitle errors [27, 188], but their clarity and competence are also undermined [119], which could lead to further negative consequences in professional and social settings.

In this study, we examine the relationship between subtitle quality and evalua-

tions of the speaker and the content through a carefully designed online experiment. Our work builds on Kim et al. [104], who compared subtitle errors across native and non-native speakers. Their design, however, did not control for confounders like accent with speaker identity and appearance. In this work, we overcome this limitation by using generative AI to manipulate the accent while holding all other speaker characteristics constant. In our experiment, participants watched videos of broad-audience talks and evaluated the speaker and the content of each talk. Participants watched two videos each, which were randomly selected from a set of four videos. Our experimental manipulation could present each of these videos alongside accurate or error-prone subtitles (the first manipulated factor) with the speaker using a Standard American English (SAE) accent or a non-Standard American English (non-SAE) accent (the second factor). A non-SAE accent could refer to regional English varieties within the United States or non-standard or globally regional accents; however, in this work, we examine an English variety spoken in India.

Our main results show that subtitle errors negatively impact the evaluation of the speaker and the content. Given the widespread deployment of ASR systems across professional, educational, and social platforms, such errors may disadvantage a large number of speakers. The negative evaluations could be particularly concerning for individuals whose accents trigger higher error rates. Our exploratory results indicate the audience perceives subtitle quality differently depending on the speaker's accent, even through the underlying error rates are similar. In addition to our empirical findings, our work offers a key methodological contribution in the form of a novel experimental design that controls for the speakers' appearance. In total, we provide more robust evidence of the downstream effects of subtitle errors, highlighting the need to consider the potential adverse impacts of these

technologies.

2.2 Background and Hypotheses

Subtitles are widely used in asynchronous settings, such as movies, TV, and online platforms like YouTube, as well as in live, synchronous settings, such as video conferencing software like Zoom. While subtitles were originally introduced in film and television to support deaf and hard-of-hearing audiences [60], research has shown that hearing viewers also benefit from subtitles through improved comprehension and engagement [27, 190, 111]. These benefits are often explained by the dual coding theory, which posits that people remember information more effectively when it is presented through both audio and visual channels [33]. The effectiveness of subtitles, however, hinges on their quality. Many factors impact subtitle quality, including accuracy, display size, and speed [27, 113, 163]. Researchers, user associations, and regulators have extensively studied subtitle quality, specifically subtitle accuracy, finding that inaccurate subtitles reduce the audience's comprehension and engagement [5, 47, 197]. Inaccurate subtitles do more than limit audience understanding—they can potentially negatively impact the speaker. ASR-generated subtitles are often less accurate than human-generated subtitles [165, 164, 27], meaning that speakers in video lectures, interviews, or live calls may be unfairly penalized by errors outside their control.

ASR systems are often less accurate due to biases that affect the system's performance [8, 55]. Biases occur when the model produces different outcomes for individuals belonging to different subgroups (e.g., gender or race) despite having the same ground-truth (or human-verified scores). Many biases are seen as byproducts of the models' construction and, in particular, the training data used [174, 202]. It is important to note that while training data is often seen as a source of bias,

it is not the only source of bias in the machine learning life cycle. For example, bias could emerge when the methods used to compute features used by the ML models are themselves biased or when the features subtly encode protected features like gender or race [186, 17]. Several studies have examined biases in ASR systems across several demographic groups like race [109, 49, 152], gender [84], and dialect [73, 123, 189, 84]. In one of the most notable of these studies, Koenecke et. al [109] showed that commercial ASR systems have a significantly lower word error rate (WER) for White speakers than Black speakers, even when the speakers uttered identical phrases.

When ASR systems produce errors as a result of biases, users from marginalized backgrounds can disproportionately experience harms [174, 39, 204]. For example, such users may need to increase their effort to make an ASR system like a voice assistant work for them. Harrington et al. [72] highlighted the difficulties older Black speakers encountered in accessing health-related information via Google Home, requiring them to engage in a form of "cultural code switching." Likewise, Cunningham et al. [39] noted that African American speakers often engage in a form of "invisible labor" by adjusting their natural speech patterns to make voice assistants work. In addition to increased labor, users of marginalized backgrounds may experience emotional harm in the form of deep feelings of frustration, self-consciousness, and shame when ASR systems do not perform well [203, 136].

Most research on harms in ASR systems focuses on voice assistants, where failures typically occur in private settings. By contrast, errors in automatic captioning are public, introducing an audience whose judgments can amplify emotional harms to the speaker. As Li et. al [119] notes, non-native speakers, in particular, worry that relying on ASR captioning tools in settings like academic talks may reflect poorly on them. Without the added burden of an erroneous captioning

tool, speakers with non-standard accents face other social pressures, as prior work shows that they are seen as less credible compared to speakers with standard accents [117, 57, 130]. Taken together, these findings suggest that subtitle errors can disproportionately disadvantage speakers with certain accents by exacerbating social biases and negatively impacting evaluations. In this study, therefore, we examine the relationship between subtitle quality and the evaluations of speakers and content with the following questions:

RQ1 How do subtitle errors impact the evaluation of the speaker and the content?

RQ2 How does the speaker's accent impact their evaluation when subtitle errors occur?

We hypothesize that:

- H1: Speakers and their content will be negatively impacted when subtitle errors occur.
- H2: Speakers with non-Standard American English (non-SAE) accents will receive more negative evaluations than speakers with Standard American English (SAE) accents.

2.3 Methods

We conducted a mixed-factorial experiment to investigate how the subtitle errors impact the perceptions of the speakers and their content (RQ1) and how the speakers' accent influences these evaluations when subtitle errors occur (RQ2). An overview of the experimental design is illustrated in Figure 2.1. As seen in the figure, participants were randomly assigned to one of two accent conditions (SAE or non-SAE) and watched two videos featuring speakers with the same accent. Each

participant viewed one video with accurate subtitles and one with error-prone subtitles in a randomized order. After watching each video, participants evaluated the subtitle quality, the speaker and their delivery, and the content of the talk. Following the experiment, we collected demographic information.

The following sections detail our experimental materials and procedure. We first describe how we selected the videos (Section 2.3.1) and created the subtitles (Section 2.3.2), before describing our measurements (Section 4.3.3) and participant recruitment (Section 3.3.5). The research design was approved by Cornell IRB and preregistered on OSF[1].

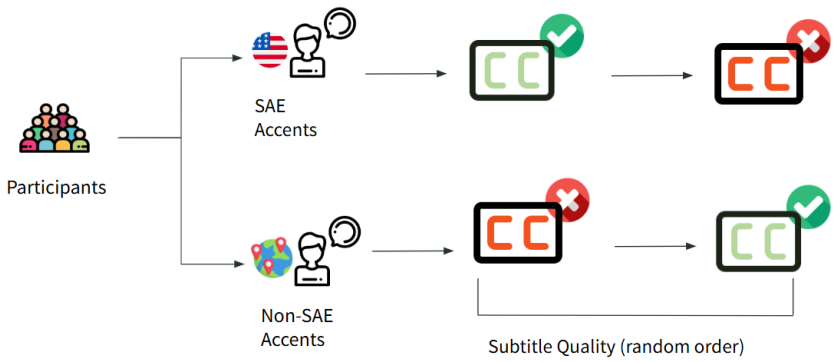


Figure 2.1: An overview of the experimental procedure.

2.3.1 Video & Audio Components

We selected publicly available TED (Technology, Entertainment, Design) Talks for our video stimuli since TED speakers address a broad audience, ensuring that the presentations are easy to understand. For the validity of our accent manipulation, it is essential that the speaker is perceived as someone who could plausibly speak with either an SAE accent or a non-SAE accent. South Asian men were an appropriate choice for this purpose, as their representation in U.S. tech and business

[1]https://osf.io/9p6sb/?view$_o nly = e773ccfebdb34b60b0b7d3c327682f3b$

occupations is higher than their share of the U.S. workforce [195, 53, 35]. Additionally, English is one of India's official languages and widely used in educational and business settings [28, 181]. As a result, without knowing where the speaker grew up, the speaker could plausibly be perceived as having either an accent heavily influenced by one of the regional languages of India or a more Americanized accent.

To further ensure ecological validity and avoid prior exposure, we selected regional TED Talks from India with fewer than 200,000 views and at least 5 years old. We identified four talks in which the speakers (all men, likely in their 20s or 30s, and of South Asian appearance) discussed their entrepreneurial journeys. The videos ranged from 13 to 21 minutes, and we extracted a 1-minute snippet from the middle of each talk. This duration aligns with prior research showing that trait impressions are formed within milliseconds of hearing a voice [142], and observations beyond one minute yield diminishing returns for forming accurate behavioral impressions [38, 147].

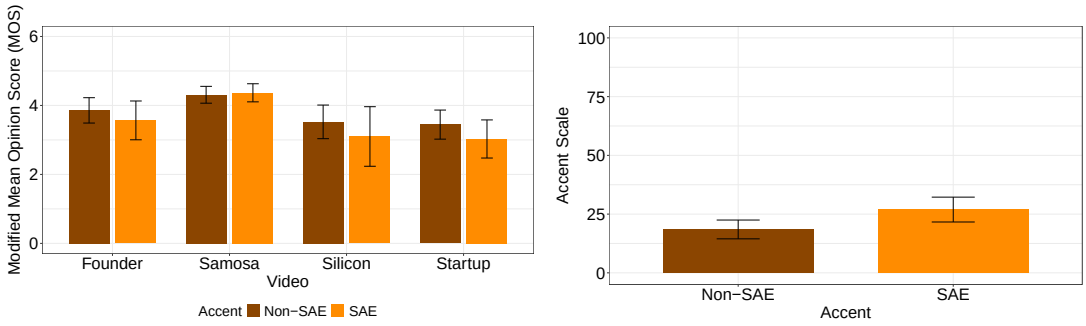
Because we are interested in how the accent affects the perception of the speaker, we needed to have two versions of each video—one with an SAE accent and one with a non-SAE accent. To ensure similar audio quality between the two accents, the audio in all of the videos was AI-generated using ElevenLabs, an AI voice-generation and manipulation service. Prior work has shown that compared to the voices from Speechify, another popular AI voice service, the voices from ElevenLabs are seen as more realistic with characteristics like stutters, mistakes, and breaths [139].

We used ElevenLabs' English V2 model to produce two new audio tracks with different speaker accents for each video, one for each accent type. The model produced an audio speech track that closely matched the original audio timing,

ensuring that the new audio could be aligned with the original video without discrepancy. We processed each video using ElevenLab's voice changer, using middle-aged, male voices, in a conversational and narrative/story style. For the non-SAE condition, while the original voice already had the target accent, we still created a new voice to ensure that the video would have a high-quality recording without background noise or echoes while retaining an accent similar to the original. We used the voices of ElevenLabs' AI "Hindi speakers" with the default stability, similarity, and style exaggeration settings for the voice changer, meaning that the new voice closely resembles the original voice and retains or somewhat amplifies its characteristics. For the SAE condition, we repeated a similar procedure with middle-aged, male "American English" AI voices. To ensure the American accent would be dominant in the generated voice, we set the similarity and style exaggeration to zero. We then created two versions of each of our videos, replacing the original audio with the newly generated audio tracks

We confirmed the audio realism and accent differences with two, pre-test experiments. We recruited 30 participants from Prolific to listen to a series of audio snippets and label the quality of the voice using a modified Mean Opinion Score (MOS) measure [139]. Participants were randomly placed into the SAE or non-SAE condition. After listening to each snippet, we asked participants to rate how natural the voice sounded, the flow of the audio, and whether the voice sounded robotic with a 5-point Likert scale. Figure 2.2a shows the mean and 95% confidence intervals for the audio quality (Y-axis) for all videos (X-axis). Based on prior work, a score of 4.3-4.5 indicates excellent quality [139]. The MOS for all of our videos is above 3, indicating a fair quality and greater. There were no significant differences in audio quality between the non-SAE audio and SAE audio ($\beta = -0.290$, SE $= 0.32$, $z = -0.898$, $p = 0.37, n.s.$, 95% CI $[-0.924, 0.343]$).

We also confirm the difference between the accent groups with a between subjects pre-test experiment. Participants watched a 1-minute video clip of a speaker with a non-SAE or SAE accent. They were then asked to rank the accent on a scale from 0 (non-SAE) to 100 (SAE). As seen in Figure 2.2b, the non-SAE accent ($M = 18.47, SD = 17.02$) was rated as less American than the SAE accent ($M = 26.93, SD = 22.50$). A t-test confirmed the result to be significant ($t(30) = 2.54, p = 0.012$).



(a) All of the audio sounds fairly natural.

(b) The non-SAE audio sounded less American than the SAE audio.

Figure 2.2: **Audio Analysis.** All of the voices sounded natural, and there were subtle differences between the accent groups.

Finally, we assessed audio–video synchronization of our video snippets with a separate between-subjects study. After watching a single video, we asked participants to rate the synchronization between the speaker's lip movements and the audio on a 5-point scale, based on the International Telecommunication Union's best practices for video quality [194]. Figure 2.3 shows the mean and the 95% confidence intervals for the synchronization. The means ranged from 3.75-4.6, indicating that there was a slight mismatch that did not interfere with the video watching experience. There were no significant differences in audio-video synchronization between the videos with non-SAE accents and SAE accents ($F(1, 71) = 0.43, p = 0.51, n.s.$).

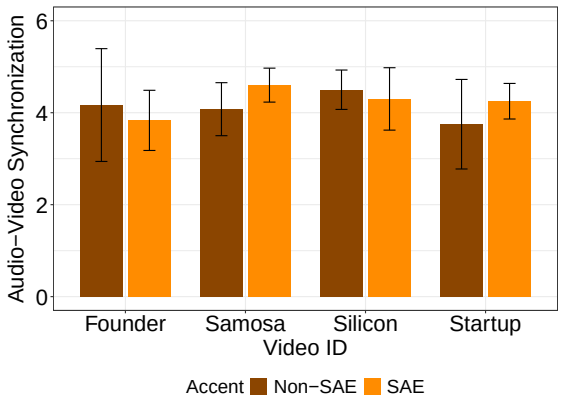


Figure 2.3: **Audio Video Synchronization.** All of the videos had a medium synchronization.

2.3.2 Subtitles

We created accurate and error-prone subtitles for each video. We used a human annotator (one of the authors) to create accurate (ground truth) subtitles. To obtain realistic yet significant error-prone subtitles, we compared transcriptions of our audio-processed videos across several popular platforms: Whisper, Google Meet, Zoom, YouTube, Otter, Apple, and Adobe. To this end, we uploaded the videos with both SAE and non-SAE audio tracks to these services and evaluated each service's performance on each audio file using the word error rate (WER). WER is a standard measure of discrepancy between the human transcription (our ground truth) and the machine transcription. WER is defined as the sum of substitutions, deletions, and insertions between the machine and ground truth transcriptions, divided by the number of words in the ground truth. A higher WER means a more error-prone transcription [109].

In our test, Google Meet was the most error-prone service with an average WER score of 0.31. For each video, we chose the Google Meet transcription with the higher WER between the two accent conditions. As a result, our subtitles were a realistic example of errors that *could* be created for a similar video by an

Table 2.1: **ASR Performance.** Word Error Rates (WER) across transcription systems for different accents. Lower values indicate better performance.

Non-Standard American English Accents							
Video ID	Whisper	Google Meet	Zoom	Youtube	Otter	Apple	Adobe
Samosa	0.2988	0.3402	0.1909	0.1909	0.2448	0.2324	0.2365
Founder	0.1761	0.3920	0.2557	0.2386	0.2216	0.3409	0.1307
Silicon	0.1312	0.3000	0.2563	0.1977	0.1813	0.2625	0.1250
Start Up	0.1173	0.1954	0.2123	0.1285	0.1397	0.1788	0.1844
Mean	0.1809	**0.3070**	0.2287	0.1879	0.1968	0.2534	0.1691

Standard American English Accents							
Video ID	Whisper	Google Meet	Zoom	Youtube	Otter	Apple	Adobe
Samosa	0.1452	0.3859	0.2033	0.1618	0.1992	0.2863	0.2158
Founder	0.1875	0.3864	0.3011	0.2784	0.2443	0.3977	0.1705
Silicon	0.2375	0.2750	0.2063	0.1813	0.1938	0.2438	0.1313
Start Up	0.1341	0.2061	0.2234	0.1117	0.1508	0.2011	0.1955
Mean	0.1761	**0.3135**	0.2335	0.1833	0.1970	0.2823	0.1783

actual system. Since our input to these models was relatively clear audio without background noise or other environment challenges, we believe the error rate could be well within the bounds of automatically generated transcriptions in the wild. We provide more details on the WER and examples of the types of errors in Tables 2.1 and 2.2.

2.3.3 Measures

Our direct measures included variables related to the quality of the subtitles (as a control and manipulation check), the speaker's delivery, and the content of the talk.

Subtitle Quality: We draw on speech and translation literature to develop comprehensive subtitle quality measures. Subtitle quality assessment requires examining multiple dimensions beyond accuracy, like timing and user experience [205, 6]. The measures we used derive from established frameworks that assess subtitle

Table 2.2: **Transcription Errors**. We provide a sample of the ground-truth and the error-prone transcriptions for a sample of our videos. Noticeable differences in the sentences are highlighted in blue.

Ground Truth	**Transcription**
Cool year for us. But every single one of these guys asked me, Munaf, what's your vision with the Borhi Kitchen? And TBK pretty much is a journey...	Coolio for us, but every single one of these guys asked me Munaf, what's your vision with the body kitchen? And TV came pretty much is a journey...
I went into a wealth management role, nice cushy job. I was enjoying my life.	I went into [omitted 'a'] wealth management role. Nice khushi job. I was enjoying my life.
And he just told me what calls you? And he said Manav just be available and something will call you, and this called me. So I left my corporate job	And he just told me what calls you [omitted '?'] and he said, Man of just be available. And something will call you and this called me. So I met my pocket job.

effectiveness across functional, technical, and readability parameters [156]. We focus on technical performance with the statements *the subtitles were accurate*, and *the subtitles were timely & synchronized*. We also explore the impact of subtitles on cognitive processing, as prior work has shown that subtitle presence can influence cognitive load [120, 111]. We capture both the potential cognitive benefits and the costs of subtitle processing with the following statements: *the subtitles improved my understanding of what was said*, and *the subtitles were distracting*. The response options for all the statements ranged from *strongly disagree* (1) to *strongly agree* (5). For our analysis, we perform a row-wise average across the responses to create the *subtitle index* measure, which captures the overall subtitle quality.

Speaker's Delivery: We draw from communication and behavioral science literature, specifically public speaking and audience engagement, to capture how the audience perceives the speaker. We used survey items from Liu et. al [125] and Curtis et. al [40] which included: *the speaker articulated and pronounced clearly* and *speaker seemed knowledgeable about the topic*. Similar to our subtitle quality

measure, the response options ranged from *strongly disagree* (1) to *strongly agree* (5). For our analysis, we calculate a row-wise average of these measures to create the *speaker index* for our overall speaker evaluation measure.

Content Quality: While the speaker measures capture *how* the speaker conveyed ideas, the content quality measures capture the evaluation of the content of the presentation. We modified the statements from Liu et. al [125] to omit unnecessary details about the speech opening or conclusion (e.g., the thesis statement clearly presented). We measure content with the following statements: *the main points were clear* and *the message felt genuine*. The response options were the same as those for the content questions. Similarly, we create a *content index* by performing a row-wise average across the three statements to be used in our analysis.

2.3.4 Participant Recruitment and Checks

We recruited 250 participants for our experiment using a standard US sample of English-speaking adults on the crowdsourcing service Prolific. To determine an appropriate sample size, we hypothesized a medium effect size using an ANOVA repeated measures, within-between interaction, aiming for 80% power based on a pilot experiment.

To ensure that participants watched the videos with the audio on, we added an attention check question after each video. One attention check asked whether participants heard a beep during the video, whereas the other asked about background noise during the talk. To ensure high-quality responses for our analysis, we excluded participants who failed the attention checks (19 participants) or correctly guessed the purpose of the experiment (17 participants). We removed an additional seven participants due to a bug that prevented them from viewing one

of the videos. In total, we had usable data from 207 participants. Overall, 33% of participants were between 45-54 years old, 77% identified as White, and 41% received a bachelor's degree. Table 2.3 provides the full details on the participants in our final dataset.

Table 2.3: **Participant Demographics**. An overview of participants' demographic backgrounds.

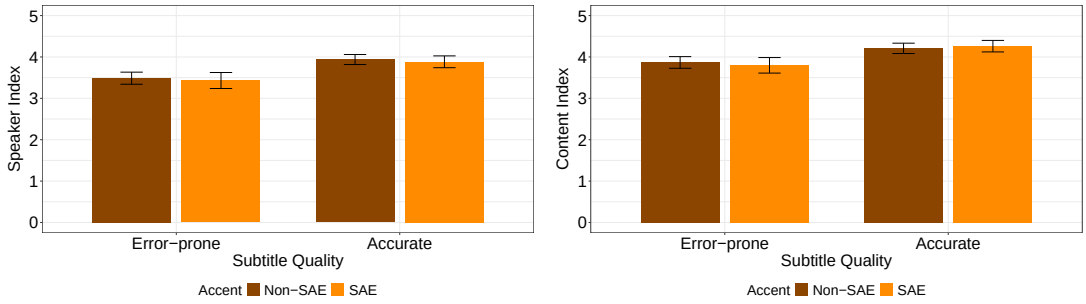
Category	Percentage
Age	
18–24 years old	5.31%
25–34 years old	14.49%
35–44 years old	24.64%
45–54 years old	33.33%
55–64 years old	18.84%
65+ years old	3.38%
Gender	
Female	53.14%
Male	45.89%
Non-binary / third gender	0.97%
Education	
Less than high school degree	0.97%
High school graduate (or equivalent)	11.59%
Some college but no degree	15.94%
Associate degree (2-year)	10.63%
Bachelor's degree (4-year)	40.58%
Master's degree	16.43%
Professional degree (JD, MD)	1.45%
Doctoral degree	2.42%
Race	
Asian	3.38%
Black / African American	15.94%
Hispanic	2.42%
Prefer to self-describe	1.45%
White / Caucasian	76.81%

2.4 Results

Our results show the impact of ASR-generated subtitle quality on speaker and content evaluations and hint at potential differences between speakers with different accents. Our primary, preregistered analysis shows that subtitle errors negatively impact both the speaker and the content evaluations (RQ1). At the same time, the

analysis shows that the speaker's accent does not impact how they are perceived when subtitle errors occur (RQ2). However, a follow-up, exploratory analysis tentatively suggests that the speaker's accent may still result in disparate impact of error-prone AI-generated subtitles when considering the *perceived quality* of the subtitles. In doing so, our work hints at the limitations of benchmarks alone and shows how considering the human perspective leads to a more robust evaluation.

2.4.1 The Effects of Subtitles & Accents



(a) Accurate subtitles positively influence speaker evaluations.

(b) Accurate subtitles positively influence content evaluations.

Figure 2.4: The impact of subtitle quality on evaluations of speaker and content.

Figure 2.4 shows the mean, with a 95% confidence interval, for our overall speaker evaluation measure (Fig. 2.4a) and the overall content evaluation measure (Fig. 2.4b) based on the 2x2 experimental manipulations of subtitle quality (x-axis) and speaker accent, with the SAE accent (in orange) and the non-SAE accent (in brown). We first examine the impact of the subtitles on the speaker evaluations, as seen in Figure 2.4a. The two columns on the left of the figure compare the evaluation scores given to speakers with error-prone subtitles (M_{SAE}=3.48, $M_{non\text{-}SAE}$=3.43). These scores are lower than those of the same speakers with accurate subtitles on the right of the figure ($M_{non\text{-}SAE}$=3.98, M_{SAE}=3.88). From the

figure, it appears as if there were no differences between speakers with SAE accents and speakers with non-SAE accents for a given subtitle quality. To statistically test whether the observations in the figure are significant, we used a linear mixed model with subtitle quality and accent as fixed effects; we included participants as a random effect to account for the repeated measures design and individual variability. Our analysis confirmed that subtitle quality has a significant main effect on speaker evaluations ($\beta = 0.451$, SE = 0.082, $z = 5.491$, $p < 0.001$, 95% CI [0.290, 0.612]), but accent did not have a significant main effect ($\beta = -0.056$, SE = 1.08, $z = -0.519$, $p = 0.604, n.s.$, 95% CI [−0.267, 0.155]). There were no interaction effects ($\beta = 0.001$, SE = 0.122, $z = 0.007$, $p = 0.995, n.s.$, 95% CI [−0.238, 0.240]). These results support H1, showing that the speakers receive better evaluations when there are accurate subtitles.

While the figure uses the averaged speaker measures, we performed a similar analysis and observe a similar trend for each of the individual speaker measures. The subtitle quality had a significant main effect on the speaker clarity ($\beta = 0.602$, SE = 0.124, z= 4.864, $p < 0.001$, 95% CI [0.359, 0.844]) and how knowledgeable the speaker appeared ($\beta = 0.324$, SE = 0.088, $z = 3.701$, $p < 0.001$, 95% CI [0.153, 0.496]), with the speaker seen as more clear and knowledgeable in the accurate subtitles condition. The accent did not affect how clear ($\beta = -0.004$, SE = 0.153, $z = -0.023$, $p = 0.981, n.s.$, 95% CI [−0.303, 0.296]) or knowledgeable ($\beta = -0.090$, SE = 0.113, $z = -0.796$, $p = 0.426, n.s.$, 95% CI [−0.311, 0.131]) the speaker appeared. Additionally, there were no interaction effects between subtitle quality and accent with regards to speaker clarity ($\beta = -0.123$, SE = 0.184, $z = -0.670$, $p = 0.503, n.s.$, 95% CI [−0.483, 0.237]) and how knowledgeable the speaker appeared ($\beta = 0.093$, SE = 0.131, $z = 0.714$, $p = 0.475, n.s.$, 95% CI [−0.163, 0.349]).

As Figure 2.4b shows, the content is also rated higher when there are accurate subtitles ($M_{\text{non-SAE}}$=4.21, M_{SAE}=4.26) than when there are error-prone subtitles ($M_{\text{non-SAE}}$=3.87, M_{SAE}=3.80). In line with our previous analysis, we employed a mixed-effects model, treating subtitle quality and accent as fixed effects and participant as a random effect. The model confirmed that subtitles have a significant effect on how the content is perceived ($\beta = 0.341$, SE = 0.078, $z = 4.371$, $p < 0.001$, 95% CI [0.188, 0.493]), but the accent does not ($\beta = -0.069$, SE = 0.106, $z = -0.656$, $p = 0.512, n.s.$, 95% CI [−0.277, 0.138]). There were no significant interaction effects between subtitle quality and accent ($\beta = 0.122$, SE = 0.116, $z = 1.055$, $p = 0.291, n.s.$, 95% CI [−0.105, 0.349]). We see the same trend for our individual content measures asking whether the content is clear and whether it is genuine. The subtitle quality has a significant main effect on content clarity ($\beta = 0.487$, SE = 0.099, $z = 4.935$, $p < 0.001$, 95% CI [0.293, 0.680]) and content genuineness ($\beta = 0.195$, SE = 0.079, $z = 2.469$, $p = 0.014$, 95% CI [0.040, 0.349]), with accurate subtitles improving the clarity and genuineness. The accent did not have a significant main effect on the content's clarity ($\beta = -0.084$, SE = 0.126, $z = -0.665$, $p = 0.506, n.s.$, 95% CI [−0.331, 0.163]) or genuineness ($\beta = -0.055$, SE = 0.107, $z = -0.514$, $p = 0.607, n.s.$, 95% CI [−0.264, 0.154]). There were no interaction effects between subtitle quality and accent for clarity ($\beta = 0.173$, SE = 0.146, $z = 1.181$, $p = 0.238, n.s.$, 95% CI [−0.114, 0.460]) or genuineness ($\beta = 0.071$, SE = 0.117, $z = 0.609$, $p = 0.543, n.s.$, 95% CI [−0.158, 0.301]).

2.4.2 Exploratory Analysis: The Effects of Perceived Subtitle Quality

Viewers may differ in how sensitive they are to subtitle errors: while some may find error-prone subtitles highly disruptive, others may not register them as strongly. These differences in subtitle evaluation can shape evaluations of both the speaker and the content, making perceived subtitle quality an important construct to examine. Beyond serving as a useful manipulation check, perceived quality also provides insight into how the audience experience mediates the effects of subtitle errors. To better capture this relationship, we conducted an exploratory analysis that incorporated participants' overall assessments of subtitle quality—that is, their perceived subtitle quality—in addition to the given subtitle condition.

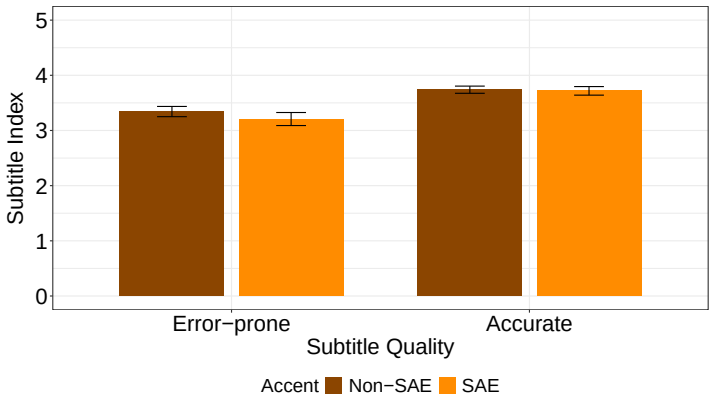


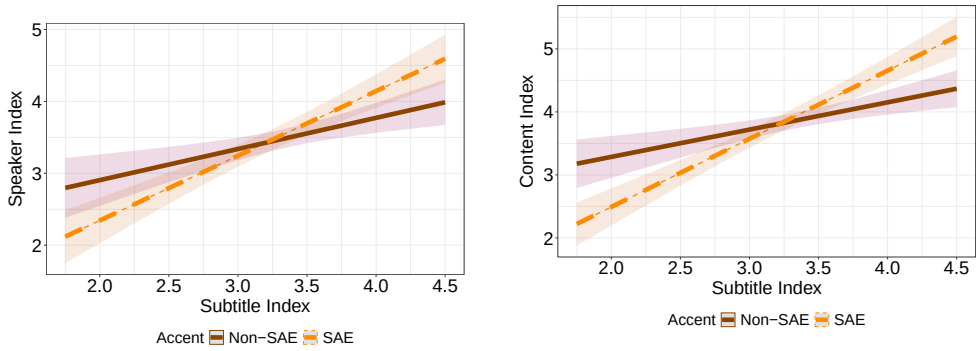
Figure 2.5: Accurate subtitles are more well-received than error-prone subtitles.

We first examine the effects of our independent measures on perceived subtitle quality in Figure 2.5. As expected, the accurate subtitles to the right ($M_{\text{non-SAE}}$=3.74, M_{SAE}=3.72) are perceived more positively than the error-prone subtitles on the left ($M_{\text{non-SAE}}$= 3.34, M_{SAE}=3.21). Among the error-prone subtitles, those with an SAE accent (orange) have a lower mean than the non-SAE accent (brown). To statistically evaluate this difference, we used a linear mixed model with the

given subtitle quality and accent as fixed effects and included participant as a random effect. The model confirmed that the given subtitle quality ($\beta = 0.396$, SE $= 0.055$, $z = 7.174$, $p < 0.001$, 95% CI [0.288, 0.504]) and the accent ($\beta = -0.135$, SE $= 0.064$, $z = -2.123$, $p = 0.034$, 95% CI [-0.261, -0.010]) have a significant main effect on how the subtitles are perceived. There were no significant interaction effects between the accent and the given subtitle quality ($\beta = 0.115$, SE $= 0.082$, $z = 1.399$, $p = 0.162, n.s.$, 95% CI [-0.046, 0.275]). While the accent and given subtitle quality manipulation were significant, there are similar means which indicates there was not a large effect size. This finding suggests that while our subtitle manipulation was successful, the differences were not acute – and that a stronger manipulation (i.e., higher error rate) would possibly lead to even stronger outcomes.

With a direct measure of how participants perceived the subtitles, we can explore whether these perceptions affect the evaluations of the speaker and the content. It is important to note that the perceived subtitle quality was not experimentally manipulated; as such, causal interpretations are limited, and these findings should be treated as exploratory. We investigated the impact of perceived subtitle quality on the speaker evaluations by running a linear model. This model included the perceived subtitle quality, the given subtitle quality, and accent as fixed effects, while participant was treated as a random effect. The model confirmed that accent ($\beta = -1.495$, SE $= 0.583$, $z = -2.562$, $p = 0.010$, 95% CI [-2.638, -0.351]) and perceived subtitle quality ($\beta = 0.433$, SE $= 0.126$, $z = 3.423$, $p < 0.001$, 95% CI [0.185, 0.681]) have a significant main effect on speaker evaluation. Specifically, speakers with SAE accents received lower evaluations than speakers with non-SAE accents, and higher perceived subtitle quality was associated with better speaker evaluations than lower perceived subtitle quality. Surprisingly, given subtitle qual-

ity did not have a significant main effect ($\beta = 0.541$, SE $= 0.767$, $z = 0.705$, $p = 0.481, n.s.$, 95% CI $[-0.963, 2.045]$). Furthermore, we observed a significant interaction between perceived subtitle quality and accent ($\beta = 0.467$, SE $= 0.175$, $z = 2.663$, $p = 0.008$, 95% CI $[0.123, 0.811]$). As shown in Figure 2.6a, speakers with non-SAE accents received higher evaluations (y-axis) than speakers with SAE accents when the perceived subtitle quality (x-axis) was poor. When the subtitles were perceived as higher quality, speakers with SAE accents were evaluated more favorably than speakers with non-SAE accents. There were no significant three-way interactions, nor were there any significant interactions between perceived subtitle quality and given subtitle quality, or between the given subtitle quality and accent.



(a) Non-SAE speakers are given the benefit of the doubt when the subtitles are perceived poorly.

(b) The content of non-SAE speakers presentations are more favorable than SAE speakers when the subtitles are perceived poorly.

Figure 2.6: The perception of subtitles and speaker on the content of the talk.

We ran a linear model with content perceptions as our dependent variable. Our analysis revealed that accent had a significant main effect ($\beta = -2.092$, SE $= 0.538$, $z = -3.890$, $p < 0.001$, 95% CI $[-3.146, -1.038]$), with Standard American English accents negatively influencing content perception. Moreover, the perceived subtitle quality also exhibited a significant main effect ($\beta = 0.433$, SE $= 0.117$, $z = 3.703$,

$p < 0.001$, 95% CI [0.204, 0.662]). Surprisingly, the given subtitle quality did not have a significant main effect ($\beta = 0.446$, SE = 0.710, $z = 0.628$, $p = 0.530, n.s.$, 95% CI [−0.946, 1.838]). There was a significant interaction between the perceived subtitle quality and accent ($\beta = 0.649$, SE = 0.161, $z = 4.018$, $p < 0.001$, 95% CI [0.332, 0.965]). As illustrated in Figure 2.6b, when subtitles were perceived as low quality, the content was better received when presented by speakers with non-SAE accents compared to speakers with SAE accents. When subtitles were perceived as high quality, the content was more positively received when delivered by speakers with SAE accents than by those with non-SAE accents.

2.5 Discussion

Taken together, our results provide some context on the impact of ASR-generated subtitle errors on speaker evaluations and who is most likely to be affected by them. Our study builds on Kim [104] by manipulating the accent within the same video stimuli. By doing so, we eliminate confounds inherent in comparing different speakers (e.g., facial features and gestures). In this discussion, we address the question of why the ASR-generated errors reflect poorly on the speaker and content, and under what circumstances the speaker's accent may impact their evaluation. Lastly, we discuss design implications and highlight the challenges and the importance of building more inclusive ASR systems.

We found that the speakers and the content of the talk were negatively impacted by subtitle errors (supporting Hypothesis 1), in contrast to earlier work showing that errors in ASR-based subtitles do not impact the perception of the speaker or their content [104]. Whereas the earlier study controlled for what the speakers said, we controlled for the speakers' appearance and delivery style. By minimizing variability in speaker-related attributes, our design made subtitle qual-

ity the primary source of difference between conditions which is perhaps why our experiment was able to detect this effect. The impact of subtitle errors on the speaker and content evaluations could be due to the fact that when subtitle errors occur, it can be more cognitively demanding for the audience and harm their comprehension and engagement [27, 111, 190].

Although we found that speakers were negatively evaluated when subtitle errors occurred, our analysis revealed that speakers with non-SAE accents and speakers with SAE accents were equally penalized (leading us to reject Hypothesis 2). This finding is somewhat surprising, given prior work suggesting that social biases often lead to lower evaluations of speakers with non-standard accents, deeming what they say as less credible [117, 57, 130]. Furthermore, prior work in ASR and accent evaluation did expose differences between non-standard (in that case, "non-native") and standard ("native") accents [104]. One possible explanation for the lack of an effect in our experiment lies in our accent manipulation. To hold constant the speaker's image and the talk's content, we relied on an AI-generated American voice overlaid on audio originally produced with an Indian accent. As a result, the final voice may not have sounded fully American.

At the same time, when we incorporate *perceived* subtitle quality into the analysis, as we did in our exploratory analysis, accent appears to matter. While metrics such as WER provide a convenient, quantitative summary of ASR performance, they often fail to align with how viewers perceive the quality of subtitles [103, 104, 98]. Numerical accuracy alone does not capture whether the transcript preserves meaning, maintains clarity, or supports the viewer's comprehension of the talk. Prior work has shown that viewers evaluate subtitles holistically, relying on perceived coherence and meaningfulness rather than on word-level fidelity [144, 103]. We found that perceived subtitle quality overpowers given subtitle

quality, and in line with Kim [104], the speaker's accent has a significant effect on how the subtitles (or the system's performance) are perceived. We find that when the subtitles are perceived as low quality, speakers with SAE accents receive lower evaluations than speakers with non-SAE accents. Although most work on harms indicates burdens against historically marginalized groups, our exploratory work demonstrates that, in this case, harms can also affect socially dominant groups.

One possible explanation for the differential speaker evaluations is that they reflect underlying expectations about who technology is designed to serve and perform well for. Empirical evidence across domains—from facial recognition and speech recognition to healthcare algorithms—shows that many technologies systematically underperform for particular demographic groups [20, 155]. The underperformance may set an implicit expectation that errors for these groups may be attributable to system limitations rather than individual shortcomings. When technology fails for groups who are presumed to be the "default" or primary users, such failures could be interpreted as reflecting personal deficits rather than systemic bias.

The implications of our findings may be substantial, with ASR-generated subtitles used in a variety of consequential domains like education, law, business organizations, and live broadcasting. For example, many organizations are increasingly using ASR-based subtitling tools during video interviews [77]. As a part of the hiring process, candidates will record themselves answering predefined interview questions without an interviewer present. The video will then be uploaded to a review system where the ASR system will create a transcript and subtitles for the video which will later be evaluated by the hiring manager. Prior work found that the ASR systems for video interviews produced more error-prone subtitles and transcriptions for speakers with non-standard accents [77]; taken with our finding

that subtitle errors harm speaker evaluations overall, job applicants with non-standard accents could encounter additional challenges when finding a job. Even within the organization, these speakers could still encounter challenges. Speakers with non-standard accents already tend to receive lower evaluations in workplace settings, even when their performance matches that of speakers with standard accents [180, 65]. When video-conferencing platforms such as Zoom or Microsoft Teams rely on automatic speech recognition (ASR) for subtitles or transcripts, errors in those outputs can intensify existing biases. Audiences may attribute transcription mistakes to the speaker rather than the system, further disadvantaging individuals with non-standard accents [104].

Despite the results conflicting with previous work [104], both of the studies suggest a path where ASR-generated subtitle errors harm both speaker and content evaluations and imply that the harms may disproportionately lie with speakers with non-standard accents. We encourage researchers and designers to explore strategies that mitigate these cascading harms. Improving model accuracy—such as through better training data from diverse dialects and speakers to reduce representational harms [39]—is an important step. However, technical fixes alone are not enough. Addressing the biases that lead to unequal error rates will not fully eliminate mistakes, since ASR systems, like most machine-learning technologies, cannot attain complete accuracy, and external factors (environment noise, low-quality audio) can make the problem arbitrarily difficult. Kim et. al [104] demonstrate that identical ASR performance can be perceived differently depending on the speaker's accent, with subtitles judged as more helpful for speakers with non-standard accents than for those with standard accents. These findings point to an important, yet underexplored, dimension of fairness in ASR: *perceived* subtitle quality. While standard evaluation metrics such as word error rate treat all errors equally, users

often weigh them differently (e.g., mis-recognition of names or numbers may be more damaging than omission of minor function words) [144]. Perceived subtitle quality may therefore shape how audiences evaluate both the speakers and the content in ways that raw accuracy scores cannot capture. The HCI community can help investigate how subtitle design and presentation (e.g., font size, placement, chunking, or timing) influence perceived quality and how adaptive or dynamic subtitle interfaces, those which change to maintain the same perceived subtitle quality based on the speaker, might mitigate harms caused by recognition errors. By integrating user-centered measures of perceived quality alongside traditional accuracy metrics, researchers and practitioners can help foster more equitable ASR systems and improve user experiences across diverse linguistic backgrounds.

Finally, we note that this work has some limitations that should be considered. First, our experiment relied on crowd workers who are Western, Industrialized, Educated, Rich, and Democratic (WEIRD) and whose views may differ from the US general public. Second, while we manipulated the accent within a single demographic group (South Asian men), we have no reason to expect the results we observed are unique to this group. Prior research consistently shows that ASR systems across a wide range of languages exhibit systematically higher error rates for certain regional accents and dialects [16, 66]. These multilingual findings suggest that accent-related ASR disparities are a general phenomenon rather than one confined to a particular demographic group or language. Third, while we show the errors negatively impact the speaker and content evaluations, we did not examine the mechanism that causes the lower evaluations or examine whether the speaker or the system is blamed for the errors. Lastly, our experimental design relied on a between subjects accent comparison which reduced the statistical power to detect subtle accent effects. We deliberately selected a between-subjects design to avoid

revealing the study purpose and contaminate participants' judgments.

2.6 Conclusion

Our findings show that subtitle quality significantly impacts speaker and content evaluations. Although the speaker's accent did not have an effect on the speaker or the content evaluations in our main study, previous research demonstrates that ASR systems perform worse for certain accents. As a result, speakers with non-standard accents are more likely to have error-prone subtitles and will disproportionately receive lower evaluations compared to speakers with standard accents. We encourage the research community to further explore subtitle quality and how it impacts speaker evaluations, as our exploratory results hint that perceived subtitle quality may be a more robust metric than error rates alone. Future work could examine what defines a good subtitle. While including additional information, such as coughs or filler words like "umm" and "ahh", may be factually correct, will it improve viewer perceptions? Moreover, which errors are significant? Audiences might be more forgiving of certain subtitle mistakes, while specific errors could directly harm a speaker's evaluation.

Beyond the design questions raised above, our findings point to a more foundational issue about how AI systems enter into social judgment. If errors can shape evaluation, what happens when AI is merely suspected? Must an AI system misperform—or even be used at all—for someone's work to be viewed differently? The next chapter extends this inquiry by examining how the perception of AI involvement, independent of actual system performance, shapes the user's evaluation and opportunity. In doing so, it introduces perceptual harms as a distinct form of harm that emerges not from technical failure, but from human perception of systems.

CHAPTER 3

AI SUSPICION SHAPES EVALUATION AND OPPORTUNITY

As AI tools are increasingly ubiquitous, it appears as if most communication is filtered through an AI lens. Prior work shows that *known* AI use can alter how recipients perceive and evaluate senders [79]. However, in an environment where it is often impossible to distinguish between AI-written and human-written text, the consequences of *suspected* AI use remain unclear. In particular, the risk of social backlash may not be evenly distributed: suspicions of AI use may trigger disproportionate penalties for certain social groups, especially those already subject to heightened scrutiny. In this chapter, we introduce and evaluate perceptual harms, a term for the harms caused to users when others perceive or suspect them of using AI. We examined perceptual harms in three online experiments, each of which examined different demographic groups such as gender, race, and nationality. We found some support for perceptual harms against certain demographic groups. At the same time, perceptions of AI use negatively impacted writing evaluations and hiring outcomes across the board. While the experiment focused on a specific type of model and context, we discuss how perceptual harms are not limited are not bound by technology or domain.

3.1 Introduction

The public release of generative artificial intelligence (AI) technologies like ChatGPT, Copilot, Claude, Stable Diffusion, and Midjourney has led to widespread adoption across various sectors. People are using generative AI for a variety of tasks, from inspiring ideas [178, 61, 210] and revising text [37] to creating images [48, 211] and popular music [31, 41, 54]. Recent scholarship in a variety of domains has shown several benefits to using generative AI tools. For example, in

the workplace, generative AI can increase workers' productivity and reduce the skill gap between employees [154, 74]. In education, it has been argued that generative AI can support engaging, personalized learning [7, 26].

Despite these possible benefits, generative AI technologies also have the potential to cause harm. Researchers have identified several types of harms in generative AI systems, many of which disproportionately burden historically marginalized groups [202, 174]. Many of the identified harms stem from biases that arise during the model's development when decisions about training data have downstream impacts on the model's outputs [135, 95]. Furthermore, several of these identified harms directly impact people who are interfacing with AI systems. For example, users may encounter stereotypical text in language models [1] or stereotypical images in text-to-image models [59, 62, 185] when prompting the model, often due to biases in the training data. However, there are some harms that are not caused by model outputs.

In this work, we define a new category of potential AI harms: perceptual harms. Perceptual harms, we argue, occur when the appearance (or perception) of AI use—regardless of whether AI was actually used—results in differential treatment between social groups. Because perceptual harms are not caused by the model's outputs, they can be categorized as a potential societal harm [174].

Perceptual harms are likely to occur as AI-generated content continues to proliferate in various domains, given the fact that AI disclosure statements are not common or practical and that people cannot reliably identify AI-generated content [46, 93]. Previous research in AI-mediated communication [70] has demonstrated, in multiple settings, that the suspicion that a communication partner used AI to create content results in reduced evaluations of that partner [92, 140]. Furthermore, when people use AI, they are seen as less intentional [80]. All these

results from prior work can be seen as examples of perceptual harms, but these papers did not reflect on any differential treatment among groups. In this context, we offer a framework to understand the various dimensions of perceptual harms and how to study the disparate impact the perception of AI use may have on different groups.

We measure perceptual harms across three dimensions of increasing consequence for the person being evaluated: suspicion of AI use, evaluation of content quality, and potential outcomes (e.g., hiring). These dimensions were chosen based on the literature highlighting people's inability to distinguish between AI-generated and human-generated content (inducing suspicion), which can influence how people assess the quality of said content [108, 93]. The suspicion and reduced quality evaluations potentially have an impact on outcomes for the people whose content is being evaluated. Furthermore, it is possible that perceptual harms may function independently across these three dimensions. For example, men might be more readily suspected of using AI than women, but women's content could still be judged more negatively than men's when AI use is suspected. Investigating perceptual harms across these three dimensions, this study asks the following questions:

RQ1 Are different groups of people suspected of using AI at different rates?

RQ2 How does the suspicion impact the evaluation of the work across different groups?

RQ3 How does the suspicion impact the outcomes for people of different groups?

We present three experiments, each of which addresses all of the research questions, with each experiment focusing on a different identity category: Experiment 1 focuses on gender, Experiment 2 on race, and Experiment 3 on nationality. We chose these three identity categories because they are legally protected categories

along the lines of which individuals in the U.S. context experience marginalization and thus are likely to display perceptual harms.

Specifically, we investigate perceptual harms in the context of online freelance marketplaces, which both provide a plausible environment and context for an online experiment of perceptual harms in LLMs and serve as an important context of study. Online freelance marketplaces are popular platforms where customers can hire workers for a variety of tasks [71]. These marketplaces have a significant economic impact by providing additional income for many workers, with many relying on them as their primary source of income [71]. While LLMs could help freelancers complete tasks more efficiently (and potentially increase earnings), these technologies have also increased competition within the marketplace as traditional customers are opting to use the LLM as opposed to human services [124, 83]. In an already competitive marketplace, perceptual harms could add to the disenfranchisement of specific groups, as evident, for example, by prior work showing that women and Black workers face discrimination in online platforms [71].

This study addresses the research questions above using a series of carefully designed online experiments to test the differences in outcomes for various demographic groups in settings that are as realistic as possible. In the experiments, participants evaluated the profiles of freelance professionals for a purported hiring decision. We used a within-subjects experimental design, with participants exposed to the (fictional) profiles from all demographic groups (e.g., women and men in Experiment 1) in a randomized order. After reading the written content from each freelancer, we asked participants whether they thought the freelancer had used AI-assistance in their writing sample and also asked them to evaluate the quality, content, and structure of the writing. We also asked participants about the likelihood of hiring the freelancer for a task similar to the writing sample.

The results are mixed. There is some evidence that different social groups may be suspected of using AI more than others. We found that socially dominant groups (men in Experiment 1) and non-dominant groups (foreign nationals in Experiment 3) may be impacted differentially on the basis of AI suspicion. However, we do not see evidence of differential evaluations or differential outcomes between groups. As expected, we find evidence that people associate AI writing with lower quality and that lower-quality work negatively impacts future job opportunities. Aside from these empirical results, this work also offers a framework for considering and evaluating perceptual harms in Generative AI and LLM applications and outlines a new way to think about the adverse impacts of these technologies.

3.2 Background and Hypotheses

We situate this study in two main areas of related literature: harms in sociotechnical systems, specifically in the context of AI, and AI-mediated communication. Building on both these areas of prior work, we present and explain the hypotheses driving our study.

3.2.1 Harms in Sociotechnical Systems

Sociotechnical systems—systems that are a combination of technical and social components—have been extensively studied for harms and biases, especially in the realm of machine learning and artificial intelligence [173, 174]. Many potential harms caused by AI systems are seen as byproducts of the models' construction and, in particular, the training data used [135, 174, 202, 95]. When socially constructed beliefs and unjust hierarchies in the data are reflected as model outputs, it is a particular type of harm known as representational harms [174, 202]. Rep-

resentational harms may occur in language models when stereotypical text or demeaning language is generated [174, 202]. Oftentimes, this harmful language is directed at historically marginalized groups like women [132, 110] and racial and ethnic minorities [1, 148]. With text-to-image models, representational harms may occur when models produce stereotypical images or fail to recognize particular identities [59, 62, 185].

Aside from the technical components of these systems, social aspects that impact users' experiences can also cause harms. For example, users of marginalized backgrounds may require an increased effort for the tool to work as well for them, a type of harm known as quality-of-service harms [174]. When these tools do not perform as well for users of marginalized backgrounds, it can cause deep feelings of frustration, self-consciousness, and shame, which users from non-marginalized backgrounds may not experience [203, 136]. Additionally, users of marginalized backgrounds may also not feel included by the tool, which can undermine their sense of agency [97]. Prior work has even found that visual cues in UIs can signal belongingness (or lack thereof) to users of different gender groups [138].

In this study, we propose a new category of potential harms, one that is situated neither in these technologies themselves nor in user interactions with them. Unlike model and user interaction-related biases, we propose that public opinion about these tools, combined with existing social stereotypes about groups or individuals, can lead to negative judgments and perceptual harms when someone is perceived or assumed to have used an AI system (whether or not they actually did so).

3.2.2 AI-Mediated Communication

AI-mediated communication (AI-MC) [70] represents a growing body of work examining the effects of introducing AI into human-to-human exchanges. Researchers

have provided robust evidence for some of the adverse outcomes when AI modifies, augments, or generates messages. One notable consequence of AI involvement in communication is the *Replicant Effect*, or rather the decrease in trust between people when someone suspects AI was involved [92]. The Replicant Effect has been shown in several settings such as chats [79] and online dating [207], and even in email, where recipients' trust decreased when they were told AI was involved in the writing process [127]. Most notably, for the context of our experiment, prior work has shown that the perception a self-presentation profile was written by AI can negatively impact the perceived trustworthiness of the profile writer's [92].

Additional research in AI-MC shows that most people cannot reliably differentiate between human-written text and AI-generated text [93, 87, 32, 44]. Oftentimes, people rely on false heuristics such as verbose language and grammatical errors [93, 56] to try to differentiate between human-written text and AI-generated text. While we know that AI involvement leads to decreased trust and that people can't differentiate between AI-generated text and human text, we do not know if these effects have differential impacts on people of different social groups.

3.2.3 Hypotheses

While most harms impact historically marginalized groups, in the context of AI use, we hypothesize that the dominant group will be most often suspected. White men are overrepresented in the technology sector [36, 115, 81] and have a more positive view toward technology [105]. Thus, while counterintuitive, we hypothesize that the perception of these trends in society will cause participants to more readily think White men are likely to use AI over groups like White women or Black men. At the same time, we hypothesize that a national identity suggesting non-US and presumed English-as-second-language speakers will be suspected of using AI more

than the baseline comparison group (white, presumed native English speakers). While nationality is a political construct, it is deeply intertwined with culture and language [209]. Although there is no official language in the United States, English remains the dominant medium of communication [68]. Prior work has shown that non-native English speakers are less likely to produce English without mistakes and could, therefore, be seen as needing more help in the form of AI assistance [86, 64, 21]. In this case, we believe the evaluators' assumptions about nationality and language will overpower the effect we expect in Experiment 1 and 2.

Our first hypotheses is thus:

- H1: In race and gender contexts, freelancers from the dominant group will be more suspected of using AI; however, contributors presented as non-English nationalities will be more suspected of using AI.

In the quality evaluation measurements, while all evaluated people could be affected by AI suspicion, we hypothesize that people from the non-dominant social groups will receive lower evaluations. Prior work has shown that while controlling for ability, age, and experience, there are still significant differences in performance evaluations between people of different genders and races. For example, in healthcare, female physicians likely unduly received lower teaching evaluations than their male counterparts [145]. Regarding race, it has been shown, for example, that in the US labor force, across a variety of occupations, Black workers received lower evaluations than White workers (more than likely due to biases) [198].

- H2: Controlling for suspicion, freelancers from the non-dominant group will receive lower quality evaluation scores.

Our hypothesis for the outcomes measurement is similar to the evaluation measurement—controlling for whether they were suspected of using AI, people

from the non-dominant social groups will be less likely to be hired. Correspondence audits in sociology have shown that while controlling for education and experience, there are still significant differences in hiring likelihood between people of different genders and races. For example, it has been shown that women with children received fewer callbacks, whereas men with children were not disadvantaged during the hiring process [34]. Additionally, for race, Bertrand and Mullainathan showed that Black job seekers are significantly less likely to be hired than White job seekers [12]. In both of these examples, women and Black job seekers had worse outcomes, likely due to biases in hiring.

- H3: Controlling for suspicion, freelancers from the the non-dominant group will be less likely to be recommended for hiring.

3.3 Methods

We addressed our research questions using a series of within-subjects online experiments, aiming to mimic a realistic scenario where crowd workers are asked to help review the profiles of freelance marketing contributors, including whether these freelancers used generative AI to create their content. We first describe how we created the profiles (Section 3.3.1) and the content associated with the freelancers (Section 3.3.2), before describing our measurements (Section 4.3.3) and experimental procedure (Section 3.3.4) in more detail. We then describe participant recruitment in Section 3.3.5.

3.3.1 Profile Creation

We manipulated the demographic information of the "freelancers" by changing the freelancers' stated names and the photos in their profiles shown to participants.

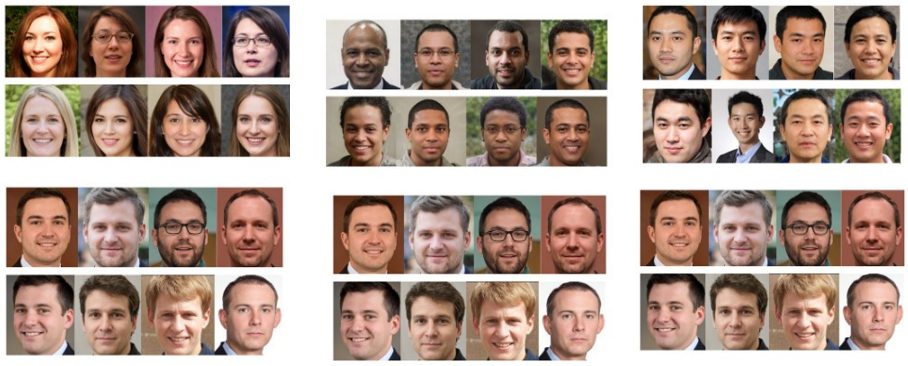
Our presentation parallels the design of popular freelancing sites like Fiverr or TaskRabbit, which do not explicitly state workers' gender, race, or nationality but feature names and photos on each profile.

For Experiment 1 (gender) and Experiment 2 (race), we created the displayed names from a list of common first and last names complied by Gaddis [58]. Gaddis used the list to test the racial perception of names used in online correspondence audits while controlling for socioeconomic status (SES) [58]. Gaddis selected 160 first names from the New York State birth record data categorized by gender, race, and SES. Gaddis also provides a set of 20 common surnames from the 2000 US Census Bureau that are associated with a skewed racial composition, i.e., significantly likely to be White or Black (e.g., Of people with the surname "Nielson", 95.6% of them are White).

For Experiment 1, we randomly selected two masculine-associated first names (i.e., Brett and Graham) and feminine-associated first names (i.e., Emily and Meredith) that are of the same race (White) and SES (high) from the initial list of first names provided by Gaddis [58]. We also selected four surnames that have a high occurrence for Whites in the US Census, like Walsh or Meyer. We randomly paired the set of first names with the set of common surnames to create the full names of our freelancers as either White men or White women (e.g., "Meredith Walsh").

For Experiment 2, where we focused on race, we reused the White masculine-associated names from Experiment 1 and created a set of Black names. We used a procedure similar to Experiment 1. From Gaddis, we randomly selected two Black masculine-associated first names that are of a high SES and paired the first names with two surnames that have a high occurrence for Blacks in the US Census.

In Experiment 3 (nationality), we used a list of names compiled by Hogan [78].



(a) Experiment 1 (gender) (b) Experiment 2 (race) (c) Experiment 3 (nationality)

Figure 3.1: The different sets of profile photos used in each experiment, with two demographic groups in each.

While the list does not have nationality metadata, the names were chosen to reflect the demographics of Toronto, Canada—a large, multi-ethnic city with many foreign-born, non-native English speakers. In our experiment, we randomly chose two distinctly East and Southeast Asian masculine-associated names from this list and used them along with the White masculine-associated names used in Experiment 1 and 2. The final set of names used in each experiment is shown in Table 3.1.

Table 3.1: **Profile Names**. Each column contains the profile names used in each experiment.

Experiment 1	Experiment 2	Experiment 3
Meredith Walsh	Andre Booker	Jun Liu
Emily Becker	Darius Washington	Fai Zhang
Brett Larsen	Brett Larsen	Brett Larsen
Graham Meyer	Graham Meyer	Graham Meyer

To create the visual representations of our freelancers, we used the AI-based image generator ThisPersonDoesNotExist [170]. The image generator created photorealistic images of fictional people based on the specified gender, age range, and ethnicity. We created a distinct set of eight images in a professional headshot

style for each demographic group in our studies. The photos featured (fake) individuals in their mid-to late-30s, which is the age of most workers on freelance platforms [172]. To create the profiles shown in our experiment interface, we randomly matched, within each demographic group, one of the two names to one of the eight photos. Figure 3.1 shows our final set of photos for each experiment.

The interface screen shown to participants was modeled after the profile pages of workers on Guru, a popular freelance site. Figure 4.1 shows an example profile as it was shown in the experiment interface, where each page features the name, photo, and location in the top banner, with the content below.

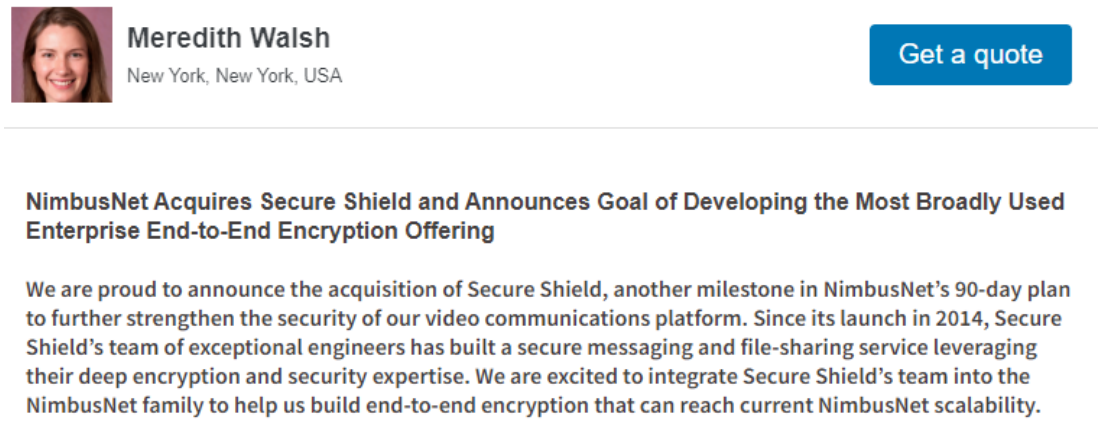


Figure 3.2: **Screenshot of the Profile Interface.** The top panel contains the profile photo, name, and location of the (fake) freelancing marketing professional. The sample press release, supposedly written by this freelancer, is below.

3.3.2 Content Creation

We asked participants to evaluate our freelancers based on the provided writing sample within the marketing and digital advertising domain. We selected this domain for our writing sample since marketing and digital advertising services are highly requested on freelance marketplaces [172, 191]. Within marketing, we chose press releases for our evaluation content, rather than other writing tasks, because

they are intended to be read by a general audience. Additionally, press releases have elements of creativity in self-promotion while also being informative [23], which ensures content variety.

We created a set of sixteen press releases for our freelancer profiles. To expand the generalizability of the experiment, we created four types of press releases—product launch, event announcement, acquisition or partnership, and new hire. We started with press release templates from the public relations websites Prowly and PR Lab so that our writing samples would mirror the visual and written structure of varied, real-life press releases. We then simplified the structure of the existing templates by removing datelines, subject lines, logos, and company contact information. The resulting set of templates for each press release type can be found in Table 3.2. After creating our templates, we identified existing press releases on Prowly, PR Lab, and Business Wire websites and modified them to match our template designs by replacing all people, products, and companies with fictional counterparts. All of our press releases were between 300-400 words long.

Table 3.2: Press release templates used in the experiments and the structure of each press release type.

Product Launch	**Event**	**Acquisition**	**New Hire**
• Title • Product overview • Uniqueness of the product	• Title • Event challenges • Future projection	• Title • Company overview • CEO's message	• Title • Overview of previous leadership • Announcement of new leadership

The content in all of the press releases was human-written; however, to ensure we arouse participants' suspicion of AI use, we manually modified half of the press releases across all types to sound more AI-like. Prior work has shown that people cannot distinguish between human-written content and AI-generated content, often relying on false heuristics to identify AI-generated content [93]. We directly used

these heuristics, namely, verbose language and rare or long words [56, 93], to modify the chosen press releases. We opted to partially modify the human-written press releases (as opposed to modifying the whole press release), ensuring content that is faithful to the original while still mimicking AI-like characteristics. That kind of process is also similar to many human-AI co-writing situations where people write the majority of the text themselves and selectively incorporate AI suggestions [13].

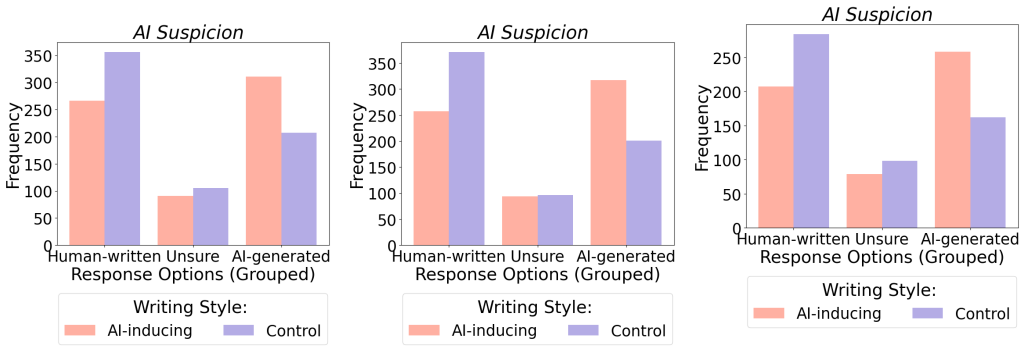
Specifically, to arouse AI suspicion, we modified one sentence in the beginning, middle, and end of half of the press releases to include verbose language or vivid descriptions. We also used a thesaurus to replace common words with less common synonyms. For example, in a product launch press release for a mobile game, we modified the sentence "Meet friendly villagers along the way and help them rebuild their homes and workshops" to "Encounter affable villagers and assist them in reconstructing their homes and workshops." We call these profiles *AI-inducing* profiles and evaluate below whether they had a different effect than the control set of profiles (controlling for any changes in quality evaluation that result from these edits). Table 3.3 presents an example of AI-inducing sentences. We validated the effect of writing style post hoc, as shown in Figure 3.3. In all three studies, the AI-inducing writing style was more likely to be seen as AI-generated compared to the control. A linear mixed model to predict AI suspicion with writing style as a fixed effect and participant as a random effect confirmed the visual finding. The result is statistically significant across all three studies ($p < 0.001$).

3.3.3 Measures

We designed the experiment to examine the perceptual harms caused by generative AI. Our direct measures included AI suspicion, variables related to the content and its quality, and a measure of potential job (hiring) outcome.

Table 3.3: **AI-inducing Sentences**. We manipulated half of the press releases to be AI-inducing by modifying three sentences in the sample. Noticeable differences in the sentences are highlighted in blue.

Original Sentence	**Modified Sentence**
Klarna, a leading global retail bank, payments, and shopping service is excited to announce its new collaboration with OpenAI, which will level up the shopping experience.	ReBank, a preeminent entity in the global retail banking, payments, and shopping services sector, is excited to announce its novel integration of generative AI, which will enhance the shopping paradigm.
Trade the bustling city for the peaceful countryside and a world of mystery with the release of InnoGames' new exploration and farming simulation game *Sunrise Village*.	Inugamis is elated to unveil *Sunset Village*, an enchanting mobile game that invites players to exchange the frenetic city life for the serene countryside and a realm brimming with mystery.
RainFocus is the next-generation event marketing platform built to capture and analyze unprecedented amounts of first-party data for exceptional events and optimized engagement throughout the customer journey.	Focus is the vanguard of next-generation event marketing platforms, meticulously engineered to capture and analyze unparalleled amounts of first-party data, thereby facilitating exceptional events and optimized engagement throughout the customer journey.



(a) Experiment 1–Gender (b) Experiment 2–Race (c) Experiment 3–Nationality

Figure 3.3: **Participants Assessment of Writing Styles.** The AI-inducing writing style was more suspected of being AI-generated compared to the control.

AI Suspicion

Prior work on trustworthiness and AI-mediated communication uses the term 'AI Score' to describe whether an evaluator thought the content had been AI-

generated [92]. In our work, we asked participants to review the profiles of online freelancers and state whether 'freelancers may have used generative AI'. By adding the statement about the freelancer's potential AI use, we primed participants to be suspicious of the extent to which the freelancer created the content. We therefore refer to 'AI Score' as 'AI suspicion' as we believe this term captures our intended measure. The response options for this measure range from *definitely human-written* (1) to *definitely AI-generated* (5) on a 5-point Likert. We also captured participants' explanations for their AI suspicion scores through a free response.

Quality Evaluations

We used three different measures for quality evaluations. First, participants were asked to provide a rating for the *overall quality* of the written content. Response options ranged from 1, indicating a poor piece of writing, to 5, indicating an excellent piece of writing. We did not provide examples for what constitutes each score, allowing participants to interpret these ratings based on their own judgment.

We draw from education literature, specifically writing evaluations for English as a Second Language (ESL) learners, to develop additional and more concrete writing evaluation measures. These content-based measures capture the strength of the ideas or information conveyed in the message [167, 169]. To develop our content measures, we omitted items from Rothschild [167] that focused on details in the writing [169, 167]. Instead, we focus on first impressions and idea clarity with the following statements: *the ideas and details expressed in the press release create an impression on the reader* and *all ideas are clear and fully developed*. The response options ranged from *strongly disagree* (1) to *strongly agree* (5). We simplify our analysis by performing a row-wise average across the responses to the two statements to create the *content index* measure.

While the content index captures *what* ideas are conveyed in the writing, the

structure measures capture *how* the idea is expressed [167]. We modified the statements from Rothschild [167] to omit unnecessary details about sentence mechanics (e.g., explaining a run-on sentence or a fragment). We measure structure with the following statements: *each sentence is complete*, *sentences are joined in the most effective/meaningful way*, and *there are almost no grammatical errors*. The response options were the same as those for the content questions. Similarly, we create a *structure index* by performing a row-wise average across the three statements to simplify our analysis.

Hiring Likelihood

We wanted to understand the potential outcomes due to perceptual harms, namely, if the perceived use of generative AI would impact hiring decisions. We asked participants how likely they would hire the freelancer to complete a similar task on a scale from 0 (very unlikely) to 100 (very likely).

3.3.4 Experiment Procedure

In each of our experiments, participants saw a demographically balanced set of four (2 demographic groups x 2 writing styles) (fictional) freelance profiles in a randomized order. After reading the content on each profile, participants were asked to evaluate if the writing sample was generated by AI (RQ1—AI suspicion), as well as evaluate the writing's overall quality and its quality in terms of content and structure (RQ2—content evaluation). Participants were then asked about the likelihood they would hire the freelancer for a similar task (RQ3—outcomes). After the experiment, we asked participants about their demographic background and familiarity & attitudes toward artificial intelligence.

3.3.5 Participant Recruitment

We recruited our participants using a US-representative sample of English-speaking adults on the crowdsourcing service Prolific. To determine the number of participants in each experiment, we calculated the sample size for a medium effect size using a linear mixed model with 80% power based on a pilot experiment. Each experiment had its own unique set of participants. We recruited 350 participants for Experiment 1 and Experiment 2 and 300 participants for Experiment 3. In all of our experiments, to ensure high-quality responses for our analysis, we excluded participants who failed attention checks or correctly guessed the purpose of the experiment. We had usable data from 334 participants in Experiment 1, the same number in Experiment 2, and 272 participants in Experiment 3. Across all three experiments, approximately 25% of participants were between 55-64 years old, 64% identified as White, 37% received a bachelor's degree, and 40% reported being somewhat familiar with coding. More than 90% of participants said they were familiar with generative AI technologies like ChatGPT or Gemini, and roughly 80% of participants reported using generative AI tools. Table 3.4 shows the full details on participant's demographics.

3.3.6 Deviation from Preregistration

For full transparency, we expand here where we deviated from our preregistered submissions in terms of hypotheses. First, we only preregistered Experiments 1 and 3. Second, H2 was different in the Experiment 3 preregistration, where we hypothesized that controlling for suspicion, people presented as foreign nationals will receive *higher* quality evaluations. In hindsight, we are not sure this direction of the hypothesis was well-justified. For simplicity, we aligned the hypothesis with the other H2 hypotheses above (H2 was not confirmed either way). Finally,

Table 3.4: **Participant Demographics**. An overview of the participants' demographic backgrounds in all three experiments.

	Experiment 1	Experiment 2	Experiment 3
Age			
18-24 years old	10.18%	12.28%	12.87%
25-34 years old	19.76%	17.96%	18.38%
35-44 years old	18.56%	16.77%	15.81%
45-54 years old	15.57%	16.47%	17.28%
55-64 years old	25.75%	24.25%	24.26%
65+ years old	10.18%	12.28%	11.40%
Sex			
Female	49.40%	49.70%	48.90%
Male	48.50%	47.60%	48.53%
Non-binary / third gender	1.80%	2.10%	2.21%
Prefer to self-describe	0.30%	0.60%	0.368%
Race			
Asian	6.29%	6.59%	6.62%
Black/African American	13.17%	16.17%	12.5%
Hispanic	9.28%	10.48%	8.09%
Native American	3.29%	1.50%	0.74%
Pacific Islander	0%	0.30%	0%
Prefer to self-describe	4.49%	3.29%	5.15%
White/Caucasian	63.47%	61.68%	66.91%
Sexuality			
Heterosexual or straight	80.54%	83.23%	79.04%
Bisexual	12.87%	9.28%	13.98%
Gay or Lesbian	4.50%	4.50%	5.15%
Prefer to self-describe	2.10%	3.00%	1.84%
Education			
Less than high school	0%	0.30%	0%
Some college but no degree	22.46%	20.36%	18.01%
Associate degree in college (2-year)	14.07%	11.38%	9.19%
Bachelor's degree in college (4-year)	35.03%	35.33%	42.65%
Master's degree	17.07%	18.86%	15.44%
Professional degree (JD, MD)	2.40%	1.80%	1.100%
Doctoral degree	0.60%	2.40%	1.10%

for Experiment 1, we only preregistered Hypotheses 1 and 2 and did not include Hypothesis 3 regarding outcomes. We developed the outcomes after the preregistration but before the Experiment 1 launch.

3.4 Results

Our results provide some evidence of perceptual harms. Our three experiments confirmed that people of different groups are suspected of AI use at varying rates (H1), but we did not find conclusive evidence that perceptual harms impacted the quality evaluations of the writing content (H2) or resulted in different hiring outcomes for people of different groups (H3). The experiments do, however, provide strong evidence that people associate content suspected as AI with lower quality regardless of the author's gender, race, or nationality, which in turn negatively impacts job opportunities.

3.4.1 Experiment 1: Gender

Experiment 1 examined how perceptual harms may affect different gender groups. We found that men were suspected of using AI more than women, and this difference between genders was exacerbated when the writing style was AI-inducing. However, we did not observe gender to have a significant impact on participants' evaluations of writing quality or their hiring recommendations.

AI Suspicion

Figure 3.4a shows the frequency (y-axis) of different survey responses to the AI Suspicion question (x-axis) for each gender. Although participants' responses were recorded on a 5-point scale, we grouped the "definitely" and "probably" responses on each side (Section 4.3.3) to create a 3-point scale reflecting the participants' leaning in their rating: human-written, unsure, or AI-generated. For example, when the evaluated freelancer was presented as a man, 42.2% of participants thought the writing was AI-generated (as indicated by the rightmost purple column in the figure). In contrast, when the freelancer was presented as a woman (in the rightmost

(a) AI Suspicion: overall frequency of responses by gender

(b) AI Suspicion: interaction between writing style and gender

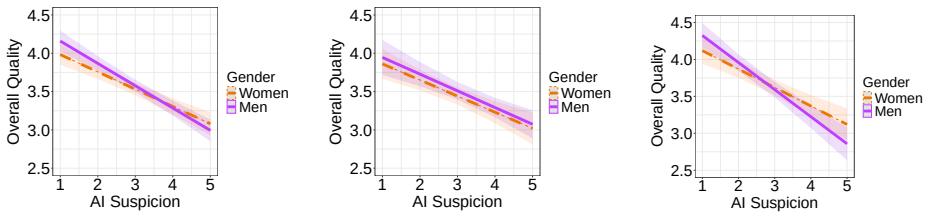
Figure 3.4: Participants' AI Suspicion evaluation by the presented gender of the evaluated freelancer.

orange column), just 35.3% of participants thought the writing was AI-generated. To statistically evaluate this difference, we used a linear mixed model with gender as a fixed effect; we included participants as a random effect to account for the repeated measures design and individual variability. The model confirmed that men freelancers were statistically significantly more likely to be suspected of using AI compared to women ($\beta = 0.105$, SE $= 0.050$, $z = -2.085$, $p = 0.037$, 95% CI [0.006, 0.203]).

We next examined how inducing AI suspicion via the language used in the writing samples affected AI suspicion. Figure 3.4b illustrates the interaction of gender (men and women) and writing style (AI-inducing or control). On the left of the figure, when the writing sample was written in a normal style (control), women (dashed, orange) and men (solid, purple) freelancers were suspected of using AI at about the same rate. However, when we manipulated the writing style to sound AI-stylized (AI-inducing), men were more suspected of using AI than women. This effect was confirmed by a linear mixed model analysis with AI suspicion as the dependent variable and the fixed effects of gender, writing style (AI-inducing or control), and their interaction with participants as a random

effect. Our results show that gender had a significant main effect on AI suspicion ($\beta = 0.392$, SE $= 0.094$, $z = 4.182$, $p < 0.001$, 95% CI [0.209, 0.576]), and there was also a significant interaction effect between gender and writing style ($\beta = -0.515$, SE $= 0.133$, $z = -3.872$, $p < 0.001$, 95% CI [-0.775, -0.254]), as reflected in the figure. The effect of writing style alone on AI suspicion also trended towards significant ($\beta = -0.178$, SE $= 0.094$, $z = -1.893$, $p = 0.058, n.s.$, 95% CI [-0.362, 0.006]).

Quality Evaluations



(a) All Writing Styles (b) AI-inducing (c) Control

Figure 3.5: The negative effect of AI Suspicion on Quality Evaluations, by gender.

Figure 3.5 shows the relationship between AI suspicion (x-axis) and overall quality ratings (y-axis) when participants evaluated freelancers presented as men (solid, purple) or women (dashed, orange). To examine the effect of the writing style on quality, we present three versions of the data: the full data from both conditions (Fig 3.5a, on the left), as well as the results from the two conditions separately, the AI-inducing style treatment (Fig 3.5b, in the middle) and the control (Fig 3.5c, on the right). Across all the conditions, the figure shows a strong inverse relationship: higher AI suspicion is associated with lower quality ratings. The figure also shows that there is no clear difference between quality evaluations for men versus women when controlling for suspicion.

We used a linear mixed model to statistically analyze the overall quality data of

all writing styles (represented in Fig 3.5a). In the model, we include AI suspicion, gender, and writing style as fixed effects. We also include the interaction effects of AI suspicion and gender, writing style and gender, writing style and AI suspicion, and the three-way interaction between gender, AI suspicion, and writing style as fixed effects. Lastly, we add participants as a random effect. The model confirms the visual trends in the figure. AI suspicion had a significant main effect on quality ($\beta = -0.193$, SE $= 0.041$, $z = -4.733$, $p < 0.001$, 95% CI $[-0.273, -0.113]$); however, gender did not have a significant main effect ($\beta = 0.125$, SE $= 0.195$, $z = 0.640$, $p = 0.522, n.s.$, 95% CI $[-0.257, 0.507]$). While writing style was not significant ($\beta = 0.345$, SE $= 0.183$, $z = 1.882$, $p = 0.60, n.s.$, 95% CI $[-0.014, 0.704]$), the trend indicates that a larger sample could expose quality differences where the AI-inducing writing style is seen as lower in quality (controlling for suspicion). The two-way interactions between gender and writing style ($\beta = 0.159$, SE $= 0.267$, $z = 0.595$, $p = 0.552, n.s.$, 95% CI $[-0.365, 0.683]$), gender and AI suspicion ($\beta = -0.020$, SE $= 0.058$, $z = -0.349$, $p = 0.727, n.s.$, 95% CI $[-0.134, 0.094]$) were not significant. Lastly, there was not a significant three-way interaction between gender, AI suspicion, and writing style ($\beta = -0.082$, SE $= 0.085$, $z = -0.959$, $p = 0.337, n.s.$, 95% CI $[-0.249, 0.085]$). Similar to the analysis for overall quality, we also consider the effect of AI suspicion on the dependent variables of content index and structure index (Section 4.3.3). We use a linear mixed model for each measure, using the same independent variables as the overall quality model reported above.

The results are largely the same for the *content* dependent variable. The AI suspicion negatively impacted content evaluation ($\beta = -0.179$, SE $= 0.036$, $z = -4.939$, $p < 0.001$, 95% CI $[-0.250, -0.108]$), and both genders were evaluated similarly ($\beta = 0.082$, SE $= 0.173$, $z = 0.475$, $p = 0.635, n.s.$, 95% CI $[-0.258,$

0.422]). However, there was a significant interaction between AI suspicion and writing style ($\beta = -0.105$, SE = 0.053, $z = -1.973$, $p = 0.048$, 95% CI [−0.210, −0.001]). Lastly, the AI-inducing writing style was seen as lower quality content compared to the control ($\beta = 0.457$, SE = 0.163, $z = 2.806$, $p = 0.005$, 95% CI [0.138, 0.776]) For the *structure* variable, the writing style had a significant main effect ($\beta = 0.485$, SE = 0.162, $z = 2.989$, $p = 0.003$, 95% CI [0.167, 0.804]), with the control writing style seen as having a better writing structure than the AI-inducing writing style. AI suspicion negatively impacted the structure evaluations ($\beta = -0.081$, SE = 0.036, $z = -2.227$, $p = 0.026$, 95% CI [−0.152, −0.010]), and both genders were viewed similarly ($\beta = 0.260$, SE = 0.173, $z = 1.504$, $p = 0.132, n.s.$, 95% CI [−0.079, 0.599]). There were no interaction effects for the structure variable.

Hiring Likelihood

Finally, we show a strong inverse relationship between AI suspicion and participants' hiring recommendations in Figure 3.6. The figure shows that when AI use was not suspected, participants rated their likelihood to hire around 70-80 on a 100-point scale. However, when AI use was heavily suspected, hiring likelihood dropped by half, to around 40/100, with men and women similarly impacted.

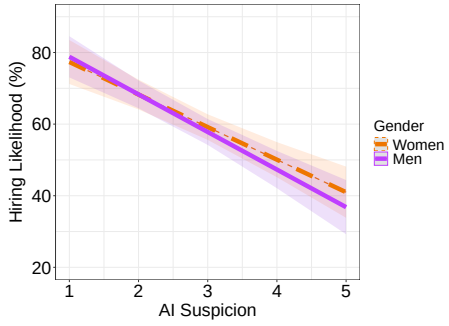


Figure 3.6: The negative effect of AI Suspicion on Hiring Likelihood, by gender.

While the hiring likelihood trend looks somewhat different between the genders, a linear mixed model analysis did not show significant differences. We used a linear mixed model with AI suspicion, gender, writing style, all two-way interactions, and the three-way interaction as fixed effects to predict the hiring likelihood score[1]. The model confirmed the relationship, with AI suspicion having a significant effect on hiring ($\beta = -7.854$, SE = 1.173, $z = -6.697$, $p < 0.001$, 95% CI $[-10.152, -5.555]$). The writing style effect again was trending towards significance, suggesting that there may be differences between the AI-inducing writing style and the control ($\beta = 9.496$, SE = 5.264, $z = 1.804$, $p = 0.071, n.s.$, 95% CI $[-0.820, 19.812]$). Gender ($\beta = 4.600$, SE = 5.600, $z = 0.821$, $p = 0.411, n.s.$, 95% CI $[-6.376, 15.575]$) did not have a significant main effect, and there were no significant interaction effects.

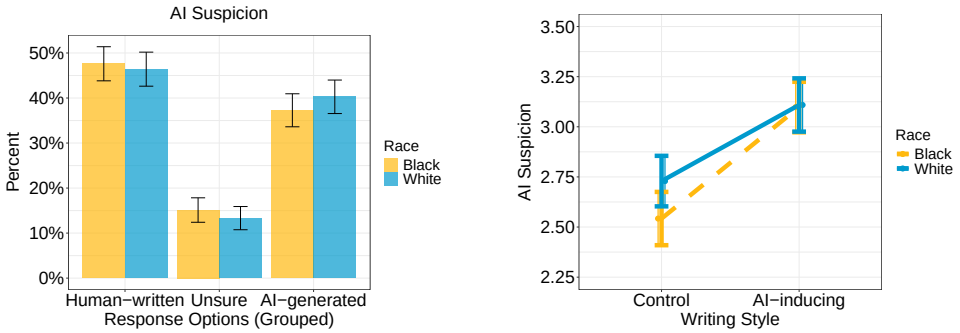
3.4.2 Experiment 2: Race

Experiment 2 examines the impact of perceptual harms on freelancers from two different racial groups, Black and White. We did not find racial differences in participants' suspicions of AI use, quality evaluation, or their likelihood to hire people from these different groups.

AI Suspicion

We did not find racial differences in AI suspicion. Figure 3.7a shows that freelancers presented as White were only suspected of using AI 40.2% of the time (rightmost, blue column), and when the freelancers were presented as Black, they were suspected 37.2% of the time (rightmost, gold). We analyzed the data in a similar manner to Experiment 1, with race as a fixed effect and participant as a

[1]Quality was highly correlated with the hiring measure and was not included in the model; we were interested to see if the variables that predict these two measures are different.

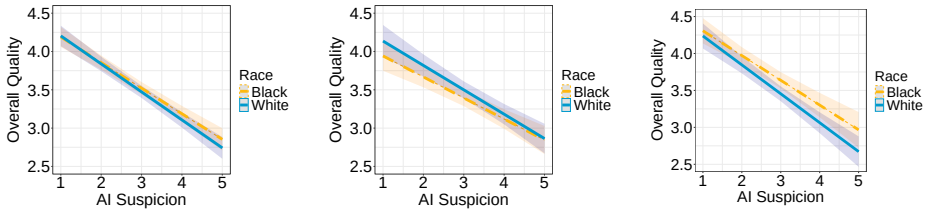


(a) AI Suspicion: overall frequency of responses by race

(b) AI Suspicion: interaction between writing style and race

Figure 3.7: **Participant's AI Suspicion**. Participants' AI Suspicion evaluation by the presented race of the evaluated freelancer.

random effect, which confirmed race did not have a significant effect on AI suspicion ($\beta = 0.042$, SE $= 0.050$, $z = 0.833$, $p = 0.405, n.s.$, 95% CI $[-0.057, 0.141]$). We also examined the relationship between writing style and AI suspicion between the racial groups, as seen in Figure 3.7b. The figure reflects that there may be slightly less suspicion towards Black freelancers (dashed, gold) in the control condition. Regardless, AI suspicion increases substantially for both groups in the AI-inducing writing style condition. A linear mixed model predicting AI suspicion from race, writing style, and their interaction (fixed effects) as well as participants (random effect) confirmed that only writing style had a significant effect on AI suspicion ($\beta = -0.556$, SE $= 0.093$, $z = -5.974$, $p < 0.001$, 95% CI $[-0.739, -0.374]$). Race did not have a significant main effect ($\beta = 0.010$, SE $= 0.093$, $z = 0.111$, $p = 0.912, n.s.$, 95% CI $[-0.172, 0.193]$). There was also no significant interaction between race and writing style ($\beta = 0.177$, SE $= 0.132$, $z = 1.335$, $p = 0.182, n.s.$, 95% CI $[-0.083, 0.436]$).



(a) All Writing Styles (b) AI-inducing (c) Control

Figure 3.8: The negative effect of AI Suspicion on Quality Evaluations, by race.

Quality Evaluations

Next, we move on to the quality evaluations and hiring judgments, as we did in Experiment 1. Figure 3.8 illustrates a strong inverse relationship between AI suspicion and overall quality (quality evaluations drop when AI suspicion grows) without notable differences by race. We confirm this result with a linear mixed model that predicts overall quality for all writing styles (Fig 3.8a) from AI suspicion, race, writing style, and all interaction effects with participants as a random effect. We observe that AI suspicion had a significant main effect ($\beta = -0.285$, SE $= 0.039$, $z = -7.335$, $p < 0.001$, 95% CI $[-0.362, -0.209]$), and there were no significant racial differences ($\beta = 0.234$, SE $= 0.189$, $z = 1.240$, $p = 0.215, n.s.$, 95% CI $[-0.136, 0.603]$). Writing style had a significant effect ($\beta = 0.416$, SE $= 0.177$, $z = 2.357$, $p = 0.018$, 95% CI $[0.070, 0.763]$), with the control receiving higher overall quality evaluations compared to the AI-inducing style when controlling for AI suspicion. There were no significant interactions in the model.

We did not see racial differences in overall quality evaluations, nor did we see evidence of racial differences in the other quality measures, namely content and structure. Similar to Experiment 1, we analyze our measures with a linear mixed model that includes a three-way interaction between race, writing style, and AI suspicion. For the content measure, we see that AI suspicion had a significant main effect ($\beta = -0.269$, SE $= 0.035$, $z = -7.649$, $p < 0.001$, 95% CI $[-0.338, -0.200]$).

Surprisingly, writing style did not have a main effect ($\beta = 0.100$, SE $= 0.161$, $z = 0.620$, $p = 0.535$, 95% CI $[-0.215, 0.415]$). There were no significant interactions between our independent measures on the content variable. In the structure index, AI suspicion ($\beta = -0.129$, SE $= 0.034$, $z = -3.773$, $p < 0.001$, 95% CI $[-0.196, -0.062]$) and writing style ($\beta = 0.377$, SE $= 0.155$, $z = 2.425$, $p = 0.015$, 95% CI $[0.072, 0.681]$) had a significant main effect, with the control receiving higher structure scores than the AI-inducing style (controlling for suspicion). Similar to the quality and content variables, there were no significant interactions between our independent measures.

Hiring Likelihood

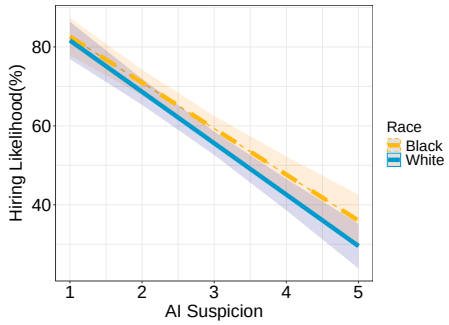


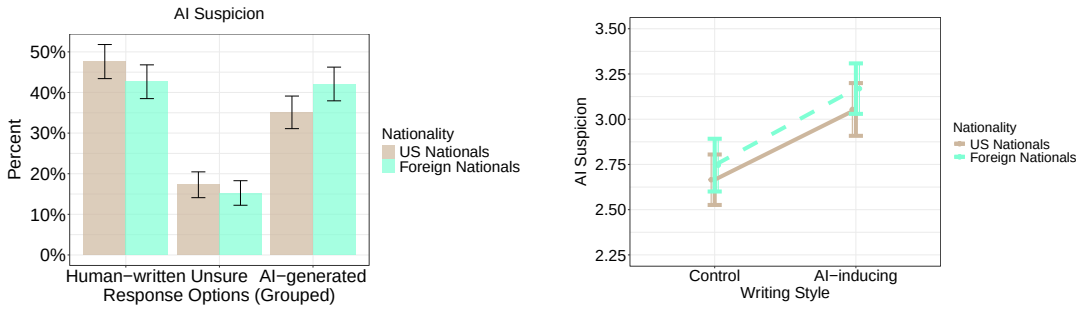
Figure 3.9: The negative effect of AI Suspicion on Hiring Likelihood, by race.

We look at the potential impact on participants' willingness to hire these freelancers. Figure 3.9 shows that participants are less likely to want to hire a freelancer they suspect of using AI. This trend is consistent across both racial categories. We confirmed the trend with a linear mixed model like we did in Experiment 1. We found that AI suspicion ($\beta = -9.998$, SE $= 1.062$, $z = -9.412$, $p < 0.001$, 95% CI $[-12.081, -7.916]$) and writing style ($\beta = 10.833$, SE $= 4.811$, $z = 2.251$, $p = 0.024$, 95% CI $[1.402, 20.263]$) had a significant main effect on hiring. There were no significant racial differences ($\beta = 5.008$, SE $= 5.145$, $z = 0.974$,

$p = 0.330, n.s.$, 95% CI [−5.075, 15.092]), nor were there significant interaction effects.

3.4.3 Experiment 3: Nationality

Our third and final experiment examines the potential for perceptual harms against individuals of foreign nationalities in the U.S. In particular, we compare freelance profiles suggestive of East Asian identity with freelancers presented as White Americans (or of similar origin). We find that freelancers presented as East Asian are suspected of using AI more than those presented as White Americans. However, we did not observe differences in quality evaluations or job outcomes.



(a) AI Suspicion: overall frequency of responses by nationality

(b) AI Suspicion: interaction between writing style and nationality

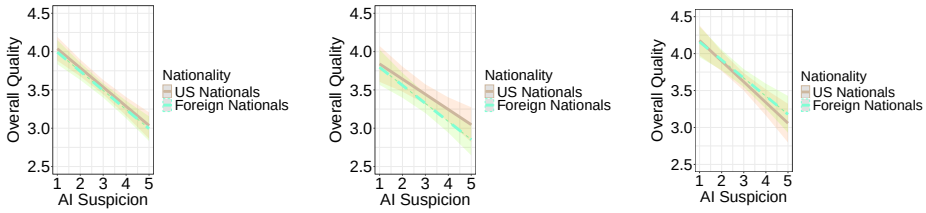
Figure 3.10: Participants' AI Suspicion evaluation by the presented nationality of the evaluated freelancer.

AI Suspicion

East Asian freelancer profiles were somewhat more likely to be suspected of using AI, as shown in Figure 3.10a. The figure shows that 42.0% of foreign nationals (rightmost, turquoise column) were suspected of using AI compared to 35.1% for U.S. nationals (rightmost, tan column). We confirmed this difference as significant ($\beta = 0.119$, SE $= 0.055$, $z = 2.181$, $p = 0.029$, 95% CI [0.012, 0.227]) using a

linear mixed model that included nationality as a fixed effect and participants as a random effect. However, we note that the differences in AI suspicion due to nationality are no longer significant when accounting for the writing style treatment. Figure 3.10b shows that writing in an AI-inducing writing style increases AI suspicion, with both U.S. and foreign nationals' profiles being affected similarly. While the figure still shows that foreign nationals (dashed, turquoise line) tend to be more readily suspected, a fuller linear mixed model did not confirm this effect as significant. Specifically, the linear mixed model (fixed effects: nationality, writing style, and their interaction; random effect: participants) confirmed that writing style has a significant main effect ($\beta = -0.388$, SE $= 0.103$, $z = -3.772$, $p < 0.001$, 95% CI $[-0.589, -0.186]$) on AI suspicion. Neither nationality ($\beta = 0.116$, SE $= 0.103$, $z = 1.125$, $p = 0.261, n.s.$, 95% CI $[-0.086, 0.317]$) nor the interaction between nationality and writing style ($\beta = -0.036$, SE $= 0.146$, $z = -0.244$, $p = 0.807, n.s.$, 95% CI $[-0.321, 0.250]$) had an effect.

Quality Evaluations



(a) All Writing Styles (b) AI-inducing (c) Control

Figure 3.11: The negative effect of AI Suspicion on Quality Evaluations, by nationality.

Like in Experiments 1 and 2, we see an inverse relationship between overall quality and AI suspicion. Figure 3.11 shows quality evaluations as a function of AI suspicion. Both sets of freelancer profiles perform similarly and show the

same steep drop in quality as AI suspicion increases. A linear mixed model predicting overall quality for all writing styles (Figure 3.11a) that included all main effects, two-way interactions, and a three-way interaction between nationality, AI suspicion, and writing style with participants as a random effect, confirmed the statistically significant impact of AI suspicion ($\beta = -0.191$, SE $= 0.046$, $z = -4.111$, $p < 0.001$, 95% CI $[-0.281, -0.100]$) and writing style ($\beta = 0.462$, SE $= 0.214$, $z = 2.162$, $p = 0.031$, 95% CI $[0.043, 0.880]$) on quality. There were no significant differences between nationalities ($\beta = -0.015$, SE $= 0.213$, $z = -0.072$, $p = 0.943, n.s.$, 95% CI $[-0.433, 0.402]$), and there were no significant interaction effects.

Considering the effect of AI suspicion on our other quality evaluation measures, we find similar results to our previous studies: AI suspicion negatively impacts the content and structure evaluations. This difference is statistically significant for content ($\beta = -0.160$, SE $= 0.043$, $z = -3.739$, $p < 0.001$, 95% CI $[-0.244, -0.076]$) and structure ($\beta = -0.118$, SE $= 0.041$, $z = -2.879$, $p = 0.004$, 95% CI $[-0.198, -0.038]$) according to linear mixed models that take into account the three-way interaction between nationality, AI suspicion, and writing style. For both content and structure evaluations, nationality and writing style do not have a main effect, and there were no significant interaction effects.

Hiring Likelihood

Lastly, we look at participants' reported likelihood to hire. The results are again similar to the two previous studies. As AI suspicion increases, hiring likelihood decreases, as seen in Figure 3.12. The figure shows that between the lowest and highest levels of AI suspicion, hiring ratings drop from around 75 out of 100 to about 40 for both East Asian and White American freelance profiles. We confirm these results with a final linear mixed model (AI suspicion, nationality, writing style, all

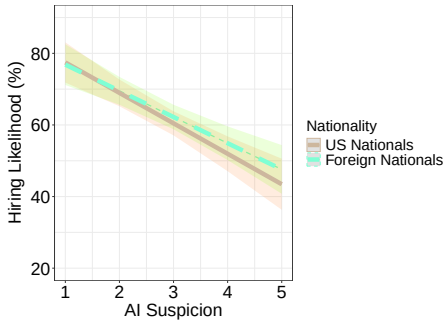


Figure 3.12: The negative effect of AI Suspicion on Hiring Likelihood, by nationality.

two-way interactions, and the three-way interaction as fixed effects; participants as a random effect). The model shows that AI suspicion had a statistically significant effect on hiring ratings ($\beta = -7.145$, SE $= 1.264$, $z = -5.651$, $p < 0.001$, 95% CI $[-9.623, -4.667]$). The effect of writing style was close to significant ($\beta = 11.193$, SE $= 5.821$, $z = 1.923$, $p = 0.054, n.s.$, 95% CI $[-0.216, 22.601]$). There were no effects of nationality ($\beta = -0.130$, SE $= 5.812$, $z = -0.022$, $p = 0.982, n.s.$, 95% CI $[-11.521, 11.262]$), and there were no significant interaction effects.

3.5 Discussion

Our study is the first to characterize an emergent potential harm in generative AI systems—perceptual harms. By introducing perceptual harms, our work extends the previous research in computing and AI harms, which has focused on harms as a result of AI's development and deployment [174, 202]. Perceptual harms, by our definition, can negatively impact individuals and groups independently of whether they are actually using AI. Rather, it is the *appearance or perception* of AI use that results in a negative judgment against the perceived AI user. In this study, we examine perceptual harms with textual content; however, perceptual harms could also occur in other mediums such as images, videos, and audio. Similar to writ-

ing tasks, creators may use text-to-image or text-to-audio [69] models to enhance their creative productivity or produce novel content [211]. For example, although research has demonstrated that knowing whether a piece of art was created by a human or AI-generated influences how people perceive the same piece of art [80], it remains unclear how merely suspecting that AI was involved impacts people's judgments about both the artwork and the creator, and how this evaluation may be different for people of different social groups.

Although most work on harms indicates burdens against historically marginalized groups, our work demonstrates that, in this case, perceptual harms can also affect socially dominant groups. The findings across our three experiments are somewhat mixed, but taken together, they provide some context about the conditions under which perceptual harms may occur. In this discussion, we try to address the question of why there were differential impacts in AI suspicion in our experiment and propose potential solutions.

One possible explanation for the disparate impact of perceptual harms we observed may stem from real or stereotypical expected differences in terms of technology use. In Experiment 1, where participants evaluated profiles of different genders, men were suspected of using AI more than women. This result is consistent with gender attitudes toward technology. Previous studies have shown that men have a positive view toward computer-related technologies whereas women generally have a more critical view [105, 91]; it is possible (and should be confirmed in future work) that men were more suspected of using AI due to real gendered differences in technology uptake. The effect may also be driven by stereotypes, with participants viewing male freelancers as more technologically savvy and, therefore, more likely to be using this relatively new technology. The possible impact of technology-related stereotypes may also be consistent with the results of Experiment 3. Ex-

periment 3 showed that freelancers with East Asian names were suspected of using AI more than White Americans. Asians and Asian Americans are often seen as so-called "model minorities" and are overrepresented in STEM [134, 30, 196]. In fact, early work has shown that Asians adopted generative AI more than White folks [22]. However, in addition to technology-related stereotypes, it is possible that the East Asian names were suspected of using AI due to language assumptions. Participants may have assumed the East Asian profiles were non-native English speakers who had AI assistance due to limited language proficiency. Again, this is a question that could be addressed in future follow-up work.

While AI suspicion was different between most of the groups in our experiments (supporting Hypothesis 1), the hypotheses about the impact of the group on quality evaluations (Hypothesis 2) and outcomes (Hypothesis 3) were not confirmed. Taken together, the results of the three experiments show that the perceived use of AI is the main factor impacting the writing evaluation and hiring outcomes. Regardless of the group, when people thought the freelancer had AI assistance, the quality, content, and structure evaluations were negatively impacted. This effect could be due to the negative associations of AI-writing [93, 108]. For example, it is possible that people who use AI may be seen as having undesirable traits in an employee (e.g., "lazy" or "incapable").

While the current results suggest that stereotypes may cause perceptual harms, the evolving societal norms surrounding AI and the context of its use will likely influence perceptual harms in the future. Since the inception of artificial intelligence, there have been many ideas of what AI is and what it can or cannot do [50]. In the 1990s, many people associated AI with chess supercomputers and IBM's Deep Blue, and by the 2010s, people thought of AI as autonomous vehicles [50]. While public sentiment toward AI was and still is largely positive [50],

there are growing concerns about issues such as loss of control, ethical challenges, and the impact of AI on employment [50, 99, 150]. The release of ChatGPT in 2022 marked a technological turning point, making generative AI mainstream as many other companies developed similar models. Shortly after ChatGPT's release, many Americans believed that AI would reduce job opportunities, with just 10% of the US adults believing AI does more good than harm [162]. While fewer Americans now perceive AI as harmful compared to a year ago, many remain cautious about its applications, especially in workplace settings [162]. Notably, people who are more knowledgeable about AI are less likely to express concerns about its effects [162]. As AI becomes increasingly normalized and people are more informed of its capabilities, the negative consequences of perceptual harms, in terms of content evaluation and loss of opportunities, may diminish or shift.

The impact of perceptual harms will depend not only on the normalization of AI but also on the specific contexts in which generative AI is suspected. For example, in artist communities, AI suspicion could lead to worse content evaluation and loss of income and opportunities [96]. Whereas, in interpersonal work communication tasks—like confirming times with high-profile professionals or expressing sympathy in a distressing scenario—AI suspicion could lead to the loss of trust between the sender and the recipient [92, 127, 79]. The context in which generative AI is suspected can also affect the magnitude of potential outcomes. For example, in mass communication like journalism, AI suspicion could further erode the public's trust of not only individual journalists but journalistic institutions as a whole [129, 52]. In contrast, in interpersonal work communication, AI suspicion might result in a more localized loss of trust. On the other hand, if the use of AI becomes more normalized in the future, the outcomes in these three scenarios may shift. In art and journalism, for example, generative AI tools *could* come to be seen as helpful

aids that enhance artists and journalists rather than undermine their credibility. In workplace settings, using AI assistance may come to be expected—like the use of spellchecking today—and avoiding its use may reflect badly on the contributor.

Beyond suspicion, we may see shifts in other social dynamics between individuals as people start being aware of the potential suspicion. Increasingly, people creating content may react in various ways, preemptively trying to avoid being perceived as using AI. Similarly, content creators may seek new ways to signal ownership and authenticity. These kinds of responses will likely be also subject to direct or indirect biases. On the other side, content evaluators may turn to a new set of tools to validate the legitimacy of work. For instance, in education, where concerns about generative AI and academic integrity are prevalent [184, 157, 208], teachers may adopt varying approaches to addressing AI-related suspicion [121], or ask students to submit a history tracking all work on their assignments. In turn, to signal effort and authenticity, students may emphasize the time they spent completing an assignment or deliberately alter their writing style to differ from AI-generated text.

Based on our findings, we encourage HCI researchers and designers to explore potential strategies to help mitigate perceptual harms. We note that perceptual harms are fundamentally a social problem caused by stereotypes and societal views of AI; therefore, previous approaches (e.g., collecting more training data from various dialects or diverse speakers to mitigate representational harms) to addressing AI's sociotechnical harms [202, 201, 175] may not fully address the problem. Perceptual harms are deeply rooted in social, cultural, and systemic dynamics that often transcend the scope of technological solutions. We recognize that while design recommendations can often be helpful, they risk oversimplifying the nuanced and multifaceted nature of social harms. To address this challenge, we need to

critically examine the trade-offs involved in potential mitigation strategies, aiming to balance practicality with a deeper understanding of these complexities. One potential solution to mitigate AI suspicion could be to introduce AI disclosure labels to AI-generated content. A disclosure statement could eliminate differences in AI suspicion; however, there is evidence to suggest disclosure would not reduce the negative perceptions around AI use [10, 122, 80, 92] and may not reduce the negative effects people could receive for using AI. A potential design intervention could go beyond an AI-disclosure label (i.e., stating that AI was involved) to signal the *extent* to which AI was involved in the creative process. For example, there could be a future system similar to InkSync [114] to demonstrate AI's involvement by providing a log of AI suggestions and tracked AI-generated content in the document. However, such detailed reports may still result in perceptual harms, and may also be rejected by users whose agency may be challenged and may feel monitored [118].

Increasing tech awareness and literacy of the capabilities of AI tools could reduce perceptual harms in terms of content evaluation and outcomes as people have a better understanding of what AI is, how it works, and what it can do [162, 14, 128]. Many people are still unsure of what AI can or cannot do [99, 50, 128]. Additional data we collected from our experiments was consistent with previous work showing people's inconsistent heuristics for detecting AI text [93]. When reviewing the justification for the AI suspicion responses in our studies, we found that one person's rationale for why the content was AI-generated was often another person's rationale for why the content was human-written. For example, like previous work [93], some of our participants associated punctuation errors or grammar mistakes with AI, while others associated these with human writing. If people had a more accurate understanding of AI's abilities, it might shift how they perceive

others' use of AI tools [162]. Rather than falling into extremes of either algorithmic aversion—dismissing AI's potential benefits—or algorithmic over-appreciation, a nuanced perspective of AI's abilities would allow people to more accurately assess human capabilities without unfairly attributing success or failure solely to AI [161, 128].

Finally, we note that this work has some limitations that should be considered. First, while we recruited a US-representative sample for our experiment, it still relied on crowd workers who are Western, Industrialized, Educated, Rich, and Democratic (WEIRD) and whose views may differ from the US general public. This US-centric study's results may not directly extend to other cultural contexts; however, we have no reason to believe that *some* of our findings would not extend more broadly as research from the Global South demonstrates perceptions of generative AI that influence how it is used and could shape how individuals view others' use of the technology [2]. Second, since our experiment highlighted AI use, it may have primed the participants to think about *and* consider it negatively. We attempted to phrase the question as neutrally as possible when describing the task to participants. In addition, our work's generalizability is limited by the fact it is based on online experiments performed in one context (freelance marketplaces). The results may not generalize to other contexts, mediums (e.g., AI-generated images or audio), or demographic groups. Finally, while we attempted to assess potential outcomes (e.g., hiring or not), the hiring measure remained hypothetical and not a behavioral measure. Of course, there are several factors not examined in this study that influence hiring decisions. A future audit study (e.g., [12]) can help provide more robust data on the impact of perceptual harms of AI.

3.6 Conclusion

This study extends the responsible computing literature by proposing the concept of perceptual harms: when the appearance (or perception) of AI use, regardless of whether it was used, results in differential treatment between people of various social groups. We propose that perceptual harms occur along three axes: suspected AI use, which can lead to differences in quality evaluations and outcomes. Through a series of online experiments, we show that dominant social groups are sometimes more likely to be suspected of using AI. Consistent with prior work, we see that perceptions of AI use negatively impacted quality evaluations and hiring outcomes, but these metrics were not different between groups after controlling for AI suspicion.

While this chapter focuses on the audience's evaluation (and what it means for the user), it raises a complementary question: if users may be evaluated based on suspected AI use regardless of actual behavior, how can AI systems be designed to better support users when they *do* engage with these tools? The next chapter addresses this question by shifting the perspective from the audience to the user's experience when collaborating with an AI-assistant. It examines how the style of an AI assistant can impact the user's sense of control and ownership during the co-writing process. In doing so, the chapter advances a framework for inclusive AI-assisted writing that mitigates quality-of-service harms.

CHAPTER 4

SUPPORTING INCLUSION AND AGENCY IN AI-ASSISTED WORK

Given large language models' (LLMs) increasing integration into workplace software, it is important to examine how these tools can best support workers. While there is extensive work documenting how AI writing assistants can improve workers' efficiency on the whole, there is also research demonstrating how these tools may not work as well for certain individuals. For example, stylistic biases in the language suggested by LLMs may cause feelings of alienation and result in increased labor for individuals or groups whose style does not match as they must put in for effort to correct the models suggestions. This chapter examines how such writer-style bias impacts feelings of inclusion, control, and ownership over the work when co-writing with LLMs. Through an online experiment with over 700 participants, we found that the style of the AI model minimally impacts perceived inclusion. However, our exploratory analysis shows that individuals with higher perceived inclusion did perceive greater agency and ownership. We found this effect more strongly impacted participants of minoritized genders, and feelings of inclusion mitigated a loss of control and agency when accepting more AI suggestions.

4.1 Introduction

Large language models (LLMs) have the potential to transform workplace communication but can also exacerbate existing societal biases. LLMs are already seeing widespread use in assisting in everyday workplace tasks like writing emails and authoring documents. Recent scholarship has shown that workers can use LLMs to increase their productivity and produce higher-quality work [154]. De-

spite these potential benefits, LLMs also have drawbacks, including the potential to reproduce social biases [202, 11]. For instance, when LLMs generate text about particular identity groups, they may display social biases in the content and style of the resulting text [199], causing *representational harm* through stereotyping and demeaning language [202, 174]. Beyond representational harms, in this work we focus on another category of harm: *quality-of-service harm* [174, 15], in which systems unequally serve different groups, which may "result in feelings of alienation, increased labor, and service or benefit loss" for people of different identities [174]. For example, if an LLM produces text that is stylistically better aligned with one gender group over another, the result is a model that disparately benefits certain users over others.

Such stylistic biases, whether real or perceived by users, can impact how individuals and groups of end-users feel about and interact with AI tools. When an LLM's text style better suits one group over another, such mismatch may impact users' feelings of inclusion when using that system. Inclusion could affect other psychological factors like users' sense of the human experiences of control and ownership [168] during the co-writing process. For example, lower perceptions of inclusion may lead users to accept fewer AI suggestions; prior work has shown the relationship between accepting AI suggestions and feelings of agency and control [141]. Given the potential importance of feelings of inclusion on people's use of AI tools, we set out to examine the relationship between writer-style bias and perceptions of inclusion, control, and ownership while co-writing with AI.

In particular, we focus here on the potential impact of AI writing assistants that provide auto-complete suggestions as the user is typing text, for example, while writing an email. In this context, we ask the following questions:

- RQ1: Can the style of AI auto-complete suggestions impact people's sense

of inclusion, control, and ownership over the text they write?

- RQ2: How do these perceptions differ by gender?

More specifically, in this work, we investigate the impact of a specific communication style—assertiveness—on writers and its impact on minoritized gender groups. We build on a rich history examining gender disparities, particularly in the workplace where assertive language can be very consequential [3, 137] and associated with better work outcomes [133, 3]. While there are many cultural differences that guide expectations of gender presentation in the workplace, scholars in gender-based communication describe male talk in Western workplaces as being assertive, powerful, and authoritative [146, 137, 192]. If writing assistants making suggestions in a more assertive style are not as well-aligned with the communication style of women and members of other minoritized gender groups, the discrepancy in style alignment could cause an additional burden for women and gender minorities by increasing writing effort and task completion time.

This study addresses the research questions above using an online experiment where participants completed a hypothetical but common and consequential workplace writing task: writing to a manager to ask for a promotion. Participants received auto-complete suggestions from one of two LLM-powered AI writing assistants, one assistant offering suggestions in an assertive style and the other in a hesitant style. We then asked participants about their feelings of inclusion, control, and ownership in the writing experience.

We found that the intervention—getting suggestions in an assertive or hesitant style from an AI assistant—impacted participants' perceptions of inclusion, control, and ownership. We did not find the assistant style treatments directly resulted in gender differences in our main measures. We do find evidence that gender plays some role: minoritized genders' perceptions of control and ownership are more

greatly impacted by their (likely pre-existing) feelings of inclusion than men's. In light of these findings, we offer a new conceptual model for understanding the impact of task- and user-style alignment on people's feelings of overall agency in AI-mediated environments.

4.2 Background

As LLMs become more robust with the ability to inspire ideas, revise text, and generate short stories, the writing assistants powered by large language models are increasingly seen as co-writers. There is a growing body of work studying the interaction between users and these co-writing systems [116, 143]. Recent scholarship in this area has focused on understanding how writers evaluate and integrate the suggestions provided by LLMs into their cognitive writing processes [13], and how writers proactively engage with system-generated content [178].

Along with their potential benefits, however, these models carry potential risks—risks that may not be equally distributed across different user groups. Prior work has identified several categories of harms including social systems harms, allocative harms, representational harms, quality of service harms, and interpersonal harms [174, 202]. Prior work has pointed to the existence of such harms in the context of auto-complete suggestions. For example, societal biases in the suggestions can cause information harms (a subset of social systems harms) by influencing what content is written [160, 90], and even shifting writers' attitudes, potentially without their awareness [90, 102]. While we do not know of work showing representational harms in auto-complete environments, given the potential for content influence [90, 102] and the known representational biases that exist in LLM text generation [174, 202], the issue is likely to persist in the auto-complete context as well.

Quality-of-service harms are another concern identified by researchers. These harms reflect developer's choices and result in unintended performance disparities based on identity [174, 15]. For example, prior work has shown that automatic speech recognition (ASR) systems have significantly higher error rates for Black speakers than White speakers [109]. These error rates can have psychological and behavioral impacts on Black users such as feelings of frustration and disappointment; they can cause users to have to modify their linguistic patterns [136]. Such quality-of-service harms can lead to alienation and result in greater labor burdens and loss of benefits for marginalized communities [174]. In this work, we extend research on this topic to consider writer-style biases in auto-complete suggestions and their potential to cause differential impacts on inclusion, control, and ownership for different social groups, with a specific focus on gender.

4.3 Methods

We addressed our research questions using an online experiment where participants completed a common workplace writing task. We asked participants to write a hypothetical email to their manager asking for a promotion. We chose the promotion task since it is a situation that requires high assertiveness in context in which minoritized genders face several challenges [3, 4, 133]. The experiment had three conditions. Participants in two of the treatments received auto-complete suggestions from a language model. In one of these treatments, the model made *assertive* auto-complete suggestions, and in the other, it made *hesitant* auto-complete suggestions. A control group of participants wrote without the assistance of any model.

After the writing task, participants answered a demographic questionnaire. Participants in the AI conditions answered an additional post-task questionnaire

regarding their perceptions of inclusion, ownership, and agency while writing with the AI assistant. The questions and 5-point Likert scale answers were adapted from previous work [141, 90]. The research design was approved by Cornell's IRB. The design was preregistered and is available on OSF[1].

4.3.1 Writing App

We used a custom experimental platform combining a rich-text editor and a writing assistant, as seen in Figure 4.1. The platform code was adapted from prior experimental work on auto-complete systems [90]. The suggestions work such that when the participant pauses typing, the system displays suggestions for completion that are 20 words or less. The user can then choose to accept the suggestion by pressing *tab* or *right arrow key,* or they can ignore the suggestion by continuing to type. Pressing *tab* accepts the subsequent word in the suggestion, and a user can press multiple times to accept the full suggestion.

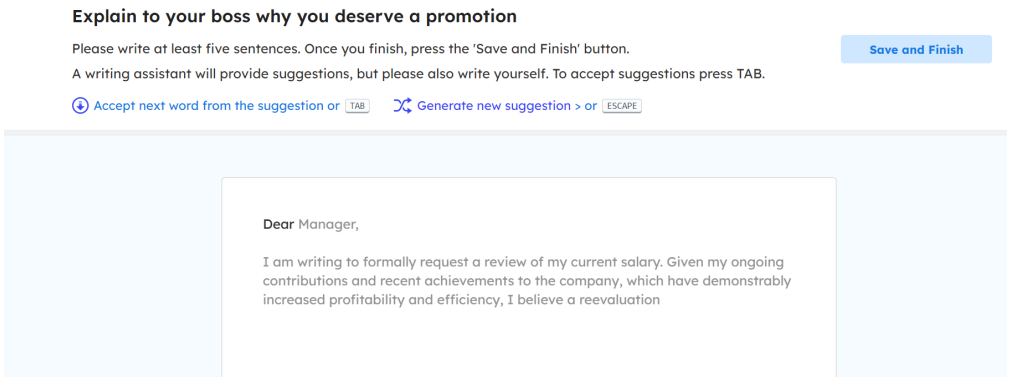


Figure 4.1: **Screenshot of the writing task.** Instructions are given in the panel at the top. Participants can hit 'tab' to accept the suggestion or 'esc' to generate a new suggestion. Written text is in black and suggestions appear in gray text.

To generate the interactive text suggestions for the experiment, we used GPT-4 via an API call with prompts carefully designed to display assertive or hesitant text.

[1]https://osf.io/kwbev?view$_o nly = c5639dcc6292482b8bd79f2086b38d24$

To achieve that goal, we inserted the prompt "In a [self-assured/hesitant] manner ask for a raise" before the participants' written text. After the entered text, we added the affirming instruction "add one sentence continuing the email in a [self-assured/hesitant] manner asking for a raise." This way, the suggestions maintained the desired style regardless of the text already entered by the participant. An example of the model's suggestions can be seen in below in Table 4.1.

Table 4.1: **Sampled Suggestions.** We prompted GPT-4 to generate suggestions in self-assured style and hesitant style. We sampled a few suggestions participants saw while they were writing. Noticeable differences in the suggestions are highlighted.

Prompt	**Sampled Suggestions**
Come up with a continuation to the input text in a **self-assured** manner asking for a raise	"Be justifiable to review my current compensation package. I believe an increase commensurate with my increased..."
	"Over the past year, I have consistently met and exceeded my performance objectives, taking on additional responsibilities without..."
	"My compensation should reflect this. I would appreciate the opportunity to discuss this with you."
	"It is fair to align my compensation with these mounting responsibilities. I have done extensive research on industry."
Come up with a continuation to the input text in a **hesitant** manner asking for a raise	"I certainly don't want to come across as ungrateful or discontented."
	"...my performance and contributions, if needed. I am not asking for an immediate response, but rather a thoughtful..."
	"In comparison to the industry standard, my current compensation seems a bit... well, it's not entirely..."
	"I thought it might be time to discuss the possibility of... well, perhaps a slight adjustment in my..."

4.3.2 Participant Recruitment

We recruited 750 participants from a gender-balanced sample of English-speaking US-based adults on the crowdsourcing service Prolific. The sample size was calculated using a two-way ANOVA with a power of 80% based on a pre-test survey. To ensure response quality, we required participants to have approval ratings greater than 95% on Prolific. We excluded participants who did not state their gender and manually verified that each writing sample was faithful to the writing task. There were five writing samples that were removed as they did not perform the requested task. In the end, we had usable data from 738 participants.

Our experiment considers two distinct gender categories: men, as category that is usually associated with greater societal influence and dominance, and minoritized genders who typically encounter societal disparities. While describing our results, we will refer to **W**omen, **N**on-**B**inary or third gender, and "preferred to describe" (also known as **S**elf-**D**escribed) as **WNBSD** participants. Of the 738 participants whose data is used, 48.9% identified as men and 51.1% as WNBSD: 47.7% identified as women, 3% identified as non-binary or third gender, and 0.4% preferred to describe their gender.

The majority of the participants were between 25-34 years old (56%), identified as white (61.9%), received a bachelor's degree (38.2%), reported working a full-time job (59.4%), and reported working entirely in person (40.6%).

4.3.3 Measures

We used the following measures to examine how participants interacted with the AI assistant and their perceptions of the writing task with the AI assistant.

AI reliance. Prior work on how users interact with AI writing assistants uses the measure of mutuality to describe the level of *interaction* the writer has with

the AI assistant [141, 116]. However, in this work, we are interested in the final text output, not the event-based interactions. We introduce a new measure, *AI reliance*, as we believe this measure captures the construct more intuitively. AI reliance is the fraction of AI-written characters in the entire text. The measure uses a score from 0 to 1, where 1 means the assistant wrote the entire text (reflecting complete reliance on the assistant) and 0 means the human wrote the entire text (no reliance).

Inclusion. We draw on the concept of self-extension, which explains how objects become integral to one's self-concept [9], to develop the notion of inclusion in AI-mediated communication. We focus on identity self-extension where the technology is a reflection of the user [166, 141]. This literature suggests that if a user feels the technology reflects themselves, they will feel included. We therefore measure inclusion with the following statements: *the writing assistant was made for people like me*, and *the writing assistant sounded like I would write myself.* Response options ranged from *strongly disagree* (1) to *strongly agree* (5).

Control. Prior work has shown that feeling in control is not limited to one action or outcome. Feeling in control can encompass engaging in an action, control over the outcome, or both [178, 141]. For example, some users may feel a lack of control since they cannot choose which suggestions they are seeing (control during the writing process). However, the fact that they can select which suggestions are incorporated into the message can impact their sense of control over the process *and* the final version [13]. We therefore measure control by asking participants *how much control did you feel over the process of writing the message* and *how much control did you feel over the final version of the message* [141]. Response options ranged from *not at all* (1) to *a great deal* (5).

Ownership. One aspect of ownership involves "mineness"—the idea that a

specific target belongs to an individual. The target can include physical objects but can also encompass intangible objects such as words and thoughts. In our experiment, we consider two targets—the message itself and the message style. Therefore, we also investigate ownership over the writing content and style. Our questions included *to what extent do you feel like the message you wrote is yours* and *thinking back on the message writing activity, how much did the message sound like you* [141]. The response options were the same as those for the control questions.

4.4 Results

Our results expose the impact of writing style on perceptions of inclusion, control, ownership, and agency and hint at potential differences between genders in that respect. Our primary, preregistered analysis finds some significant effects of AI writing style on perceptions of inclusion, control, and ownership. A follow-up, exploratory analysis suggests that perceptions of inclusion mediate the loss of agency as AI reliance increases. Additionally, the analysis suggests that heightened perceptions of inclusion lead to increased agency, particularly among minoritized genders.

4.4.1 The Overall Writing Experience

We first provide an in-depth analysis of the AI suggestions before providing an overview of their effects on the writing task. For our analysis, we randomly sampled 100 participants from each AI treatment condition (self-assured or hesitant). We selected suggestions that appeared at the beginning, middle, and end of the writing task for a total of five suggestions per participant. We used LIWC-22 on the auto-

complete suggestions to analyze the following features: positive tone, negative tone, pro-social behavior, politeness, conflict, clout, authenticity, and social behavior as a whole. Clout refers to the social status or confidence people display in writing. Authenticity refers to how much a person self-regulates, with low authenticity indicating the person is socially cautious and high authenticity indicating a person is spontaneous. Based on prior literature [146, 137, 192], self-assured language and hesitant language should differ along these metrics.

With the exception of the positive tone metric, t-tests revealed significant differences between the means of the self-assured model suggestions and hesitant model suggestions across all of the metrics (negative tone: $t(102.26) = -6.73$, $p < 0.001$; pro-social behavior: $t(192.84) = 3.29$, $p = 0.0012$; politeness: $t(154.30) = 3.77$, $p = 0.0024$, conflict: $t(99) = -2.27$, $p = 0.025$; clout: $t(194.28) = 3.15$, $p = 0.002$; authenticity: $t(181.7) = -4.00$, $p < 0.001$; social behavior: $t(194.74) = 3.24$, $p = 0.001$). Positive tone was comparable between the two styles ($M_{\text{self-assured}} = 5.32$, $SD_{\text{self-assured}} = 2.33$; $M_{\text{hesitant}} = 4.97$, $SD_{\text{hesitant}} = 1.90$). In contrast, the hesitant suggestions were substantially more negative (M = 0.59, SD = 0.86) than the self-assured suggestions (M = 0.01, SD = 0.11). The self-assured suggestions also used more prosocial language (M = 2.15, SD = 1.64) than the hesitant suggestions (M = 1.44, SD = 1.39), and they were markedly more polite (M = 0.89, SD = 1.25) compared to hesitant suggestions (M = 0.35, SD = 0.69). The hesitant suggestions showed slightly more conflict-related language (M = 0.06, SD = 0.27), whereas the self-assured suggestions contained virtually none (M = 0.00, SD = 0.00). The self-assured suggestions also exhibited higher clout scores (M = 24.69, SD = 21.53) than the hesitant suggestions (M = 15.70, SD = 18.73), while the hesitant suggestions were rated as more authentic (M = 88.36, SD = 13.41) than the self-assured suggestions (M = 79.29, SD = 18.26). Finally, the self-assured

suggestions used more social language (M = 13.57, SD = 3.54) than the hesitant suggestions (M = 12.04, SD = 3.11).

We then examined whether the suggestions impacted the tone of the message overall using three human annotators. The annotators evaluated all of the final messages on a scale from 1-10, where 1 indicated no assertiveness and 10 indicated high assertiveness. We then performed a row-wise average to determine the tone of the message overall. A one-way ANOVA showed that the AI treatment significantly influenced how assertive participants perceived the messages to be ($F(2, 732) = 120.62, p < 0.001$). Tukey HSD post hoc tests indicated that participants rated the control messages (M = 6.17, SD = 0.88) as significantly more assertive than those written in the hesitant AI style (M = 5.24, SD = 1.14). The self-assured AI style (M = 6.52, SD = 0.76) also differed significantly from the control condition, and it was rated as significantly more assertive than the hesitant style.

Lastly, we examined the effects of the suggestions on the writing experience, namely the length of text written by participants, the writing time, and participants' degree of AI reliance. The AI treatment had a significant effect on the text length ($F(2, 732) = 67.86$, $p < 0.001$). Post hoc comparisons using the Tukey HSD test indicated that the difference between the control (M = 98.53, SD = 35.64) and the hesitant AI style (M = 151.80, SD = 73.60) was significant. The difference between the control and self-assured AI style (M = 168.38, SD = 89.15) was significant, as well as the difference between the hesitant AI style and the self-assured AI style. Furthermore, the AI treatment had a significant effect on the time spent on the task ($F(2, 732) = 6.35$, $p = 0.002$). Using a Tukey HSD test, we observe that participants who wrote with the self-assured AI suggestions (M = 284.75, SD = 209.90) spent significantly less time than participants in the control (M = 363.65, SD = 292.67). There were no differences between the control and the hesitant AI

style or the hesitant AI style and the self-assured AI style. We also observe significant impacts of AI writing style on AI reliance: participants who wrote with the self-assured writing style model relied on the model more (M=0.55, SD=0.34) than participants who wrote with the hesitant writing style model (M=0.45, SD=0.35). These differences were also significant ($t(486) = 3.16$, $p = 0.002$).

4.4.2 Inclusion

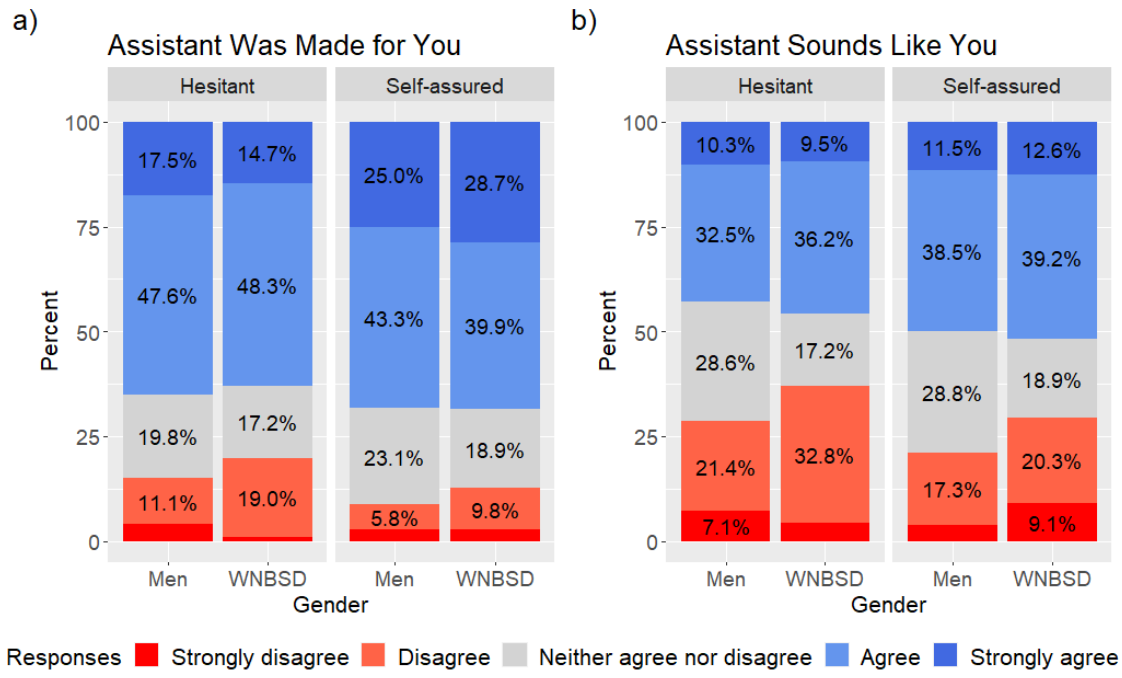


Figure 4.2: **Participants' assessment of inclusion.** Participants are more likely to say the assistant was made for them than sounds like them. Responses less than 5% are not labeled on the figure.

We examine the impact of AI writing style on our key measures, beginning with perceptions of inclusion. Figure 4.2 shows the frequency of different survey responses for each gender condition (different columns) for the inclusion statements: *the writing assistant was made for people like me* ("Assistant Was Made for You" on the left), and *the writing assistant sounded like I would write myself*

("Assistant Sounds Like You" on the right). In the condition with the hesitant AI writing style (left side of Figure 4.2a), 65.1% of men (first column) and 63% of WNBSD participants (second column) either agreed or strongly agreed (blue and dark blue bars in the column) that the writing assistant was made for people like them. As the figure also shows, in the self-assured AI writing style group (right side of Figure 4.2a), 68.3% of the men and 68.6% of WNBSD participants either agreed or strongly agreed with this statement. In general, participants more readily agreed with the "made for you" statement (Figure 4.2a) compared to the "sounds like you" statement (Figure 4.2b). The AI treatment group had a significant effect ($F(1, 485) = 5.605$, $p = 0.018$). Post hoc comparisons using the Tukey HSD test indicated that participants who wrote with the self-assured AI writing style (M = 0.817, SD = 1.01) felt like the assistant was made for them more participants who wrote with the hesitant AI writing style (M = 0.603, SD = 1.00). Gender did not have a significant effect ($F(1, 485) = 0.125$, $p = 0.723, n.s.$) nor was there an interaction effect between the AI treatment group and gender ($F(1, 485) = 0.131$, $p = 0.717, n.s.$).

In Figure 4.2b, of the participants who received the hesitant treatment (left), 42.8% of the men and 45.7% of the WNBSD participants either agreed or strongly agreed that the assistant sounded like how they would write themselves. In the condition with the self-assured AI writing style, on the right side of Figure 4.2b, 50% of the men and 51.8% of the WNBSD participants either agreed or strongly agreed. There were no significant differences for gender ($F(1, 485) = 0.498$, $p = 0.480, n.s.$), AI treatment group ($F(1, 485) = 2.307$, $p = 0.130, n.s.$), or the interaction between gender and the AI treatment groups ($F(1, 485) = 0.120$, $p = 0.730, n.s.$) for this question.

4.4.3 Control

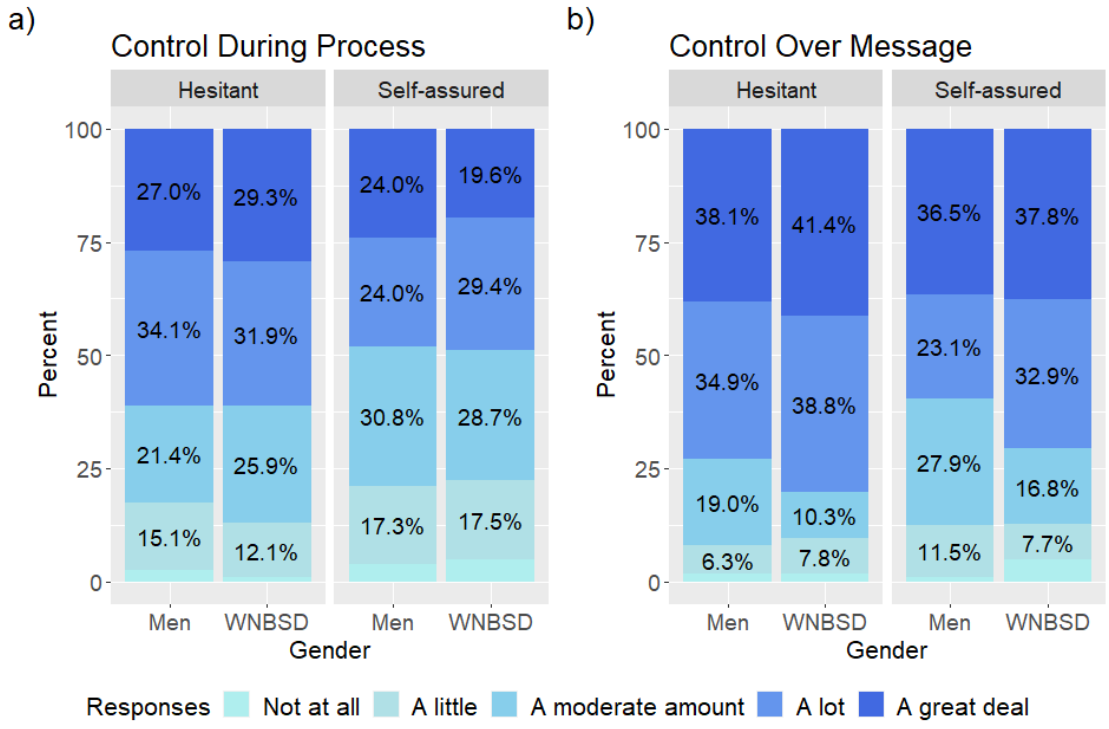


Figure 4.3: **Participants assessment of control.** Participants assisted by the hesitant writing style model are more likely to feel greater control over the final version of the message and the writing process than participants assisted by the self-assured writing style model. Responses less than 5% are not labeled on the figure.

Figure 4.3 shows the frequency of different survey responses for each gender (different columns) for the following questions: *how much control did you feel over the process of writing the message* ("Control During Process" on the left), and *how much control did you feel over the final version of the message* ("Control Over Message" on the right). In the condition with the hesitant AI writing style (left side of Figure 4.3a), 61.1% of the men (first column) and 61.2% of WNBSD participants (second column) reported feeling a lot or great deal of control (the darker blue and darkest blue bars in the columns) during the writing process. In the condition with the self-assured AI writing style (right side of Figure 4.3a),

48% of men and 49% of the WNBSD participants expressed a lot or a great deal of control during the writing process. A two-way ANOVA revealed that the difference between the AI treatment groups was significant ($F(1, 485) = 8.17$, $p = 0.004$), meaning that participants in the hesitant condition reported more control over the process. There were no significant differences for gender ($F(1, 485) = 0.017$, $p = 0.90, n.s.$) or the interaction between gender and the AI treatment groups ($F(1, 485) = 0.51$, $p = 0.48, n.s.$) for this question.

Looking at Figure 4.3b and control over the final version, in the condition with the hesitant AI writing style (left), 73% of men and 80.2% of WNBSD participants reported a lot or a great deal of control over the final version of the message. The right side of Figure 4.3b shows 59.6% of men and 70.7% of WNBSD participants who wrote with a self-assured writing style model reported a lot or a great deal of control over the final version of the message. For this question as well, a two-way ANOVA revealed that the difference between the AI treatment groups was significant ($F(1, 485) = 4.00$, $p = 0.046$), with participants in the hesitant AI writing style condition reporting greater control over the final version of the message. At the same time, there were no significant differences in relation to gender ($F(1, 485) = 0.78$, $p = 0.38, n.s.$) or interactions between gender and the AI treatment group ($F(1, 485) = 0.0007$, $p = 0.97, n.s.$) for this question. In both treatment conditions, participants generally reported higher levels of control over the final version of the message (Figure 4.3b) than the during the writing process (Figure 4.3a).

4.4.4 Ownership

Figure 4.4 illustrates the frequency of different survey responses for each gender condition (columns) for the following questions: *to what extent do you feel like the*

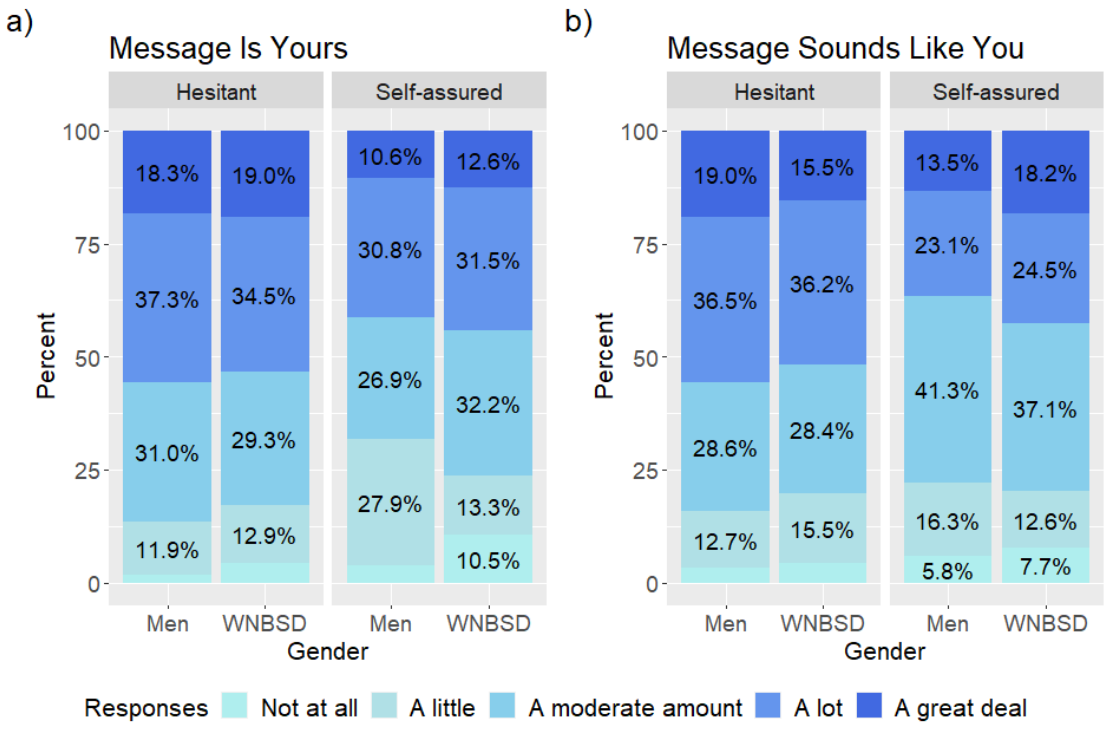


Figure 4.4: **Participants assessment of ownership.** Participants assisted by a hesitant model are more likely to say that they wrote the message and the message sounds like them. Responses less than 5% are not labeled on the figure.

message you wrote is yours ("Message Is Yours"), and *thinking back on the message writing activity, how much did the message sound like you* ("Message Sounds Like You"). In the hesitant AI writing style condition (left side of Figure 4.4a), 55.6% of the men and 53.5% of WNBSD participants expressed a lot or a great deal of ownership over the message (the dark blue and darkest blue bars in the column). In the self-assured AI writing style condition (right side of Figure 4.4a), participants experienced a higher level of ownership, with 41.4% of men and 44.1% of the WNBSD participants expressing a similar degree of ownership over the message. A two-way ANOVA revealed that this difference between the AI treatment groups was significant ($F(1, 485) = 12.89$, $p < 0.001$). There was no significant difference for gender ($F(1, 485) = 0.009$, $p = 0.92, n.s.$) or the interaction between gender and the AI treatment groups ($F(1, 485) = 0.50$, $p = 0.47, n.s.$) for this question.

Reflecting on the style of the message, Figure 4.4b shows a similar trend: participants in the hesitant AI writing style condition (left) were more likely to say the final message reflected their style. Once again, a two-way ANOVA revealed that the difference between the AI treatment groups was significant ($F(1, 485) = 4.63$, $p = 0.032$). There was no significant difference for gender ($F(1, 485) = 0.008$, $p = 0.93, n.s.$) or the interaction between gender and the AI treatment groups and gender ($F(1, 485) = 1.39$, $p = 0.24, n.s.$) for this question.

Overall, our results show AI writing style impacts control and ownership. Participants who wrote with the hesitant model expressed greater control over the writing process and the final message and felt the message sounded more like them than participants who wrote with the self-assured model. We saw a similar result for ownership, with participants in the hesitant treatment group expressing greater ownership over the message and stating the message sounds more like them than participants in the self-assured treatment group. However, our analysis did not reveal an impact of the AI writing style on perceived inclusion. Additionally, we did not see any effects of gender or interaction between gender and treatment group across inclusion, control, and ownership questions. We further explore these factors in the analysis below.

4.4.5 Exploratory Analysis: Inclusion and Agency

To better understand the relationship between the different measures and the participants' interaction with the AI model, we perform an exploratory analysis that brings together multiple factors using a linear regression model. To perform this analysis, we combined various measures into simplified high-level constructs. First, we use a compound measure of agency as our dependent variable. The concept of agency has been defined in several different ways but broadly captures the idea

of "the capacity to alter a situation" [149, 24]. The variables that reflect that in our study were control and ownership. While distinct, the literature shows these concepts are closely related: whether a person has control greatly influences their perceived ownership [141, 18]. We confirmed the relationships between our post-task measures with a correlation matrix, as seen in Table 4.2. The table shows the two ownership questions (*message sounds like me*, *message is mine*), two control questions (*control during process*, *control over message*), and two inclusion questions (*assistant made for me*, *assistant sounds like me*) are each highly correlated, as indicated by the bold text. We also note that the control and ownership questions were indeed highly correlated. Thus, we created a score for agency by performing a row-wise average across the control and ownership questions.

Table 4.2: **Post-task question Correlation Table.** The highest correlations are between *message is mine* and *message sounds like me* (the two ownership questions), *control during process* and *control over message* (the two control questions), and *assistant made for me* and *assistant sounds like me (the two inclusion questions).*

	message is mine	control over message	control during process	message sounds like me	assistant made for me	(assistant sounds like me)	ai_reliance
message is mine	1.00						
control over message	0.53	1.00					
control during process	0.45	**0.65**	1.00				
message sounds like me	**0.67**	0.51	0.45	1.00			
assistant made for me	0.23	0.21	0.21	0.37	1.00		
assistant sounds like me	0.13	0.17	0.22	0.21	**0.56**	1.00	
ai_reliance	-0.37	-0.27	-0.22	-0.28	0.13	0.16	1.00

The independent variables in the model consisted of a simplified inclusion variable, the AI reliance score, the AI writing style, and demographic variables. To simplify the model and analysis, we combined the two highly correlated inclusion measures into one variable by performing a row-wise average across the two inclusion questions. The model also included the measure of AI reliance as defined above, namely how much the participants accepted the AI suggestions in the writing process. Finally, we include in our model the AI writing style of the model that participants wrote with (the AI treatment), as well as more detailed demographic

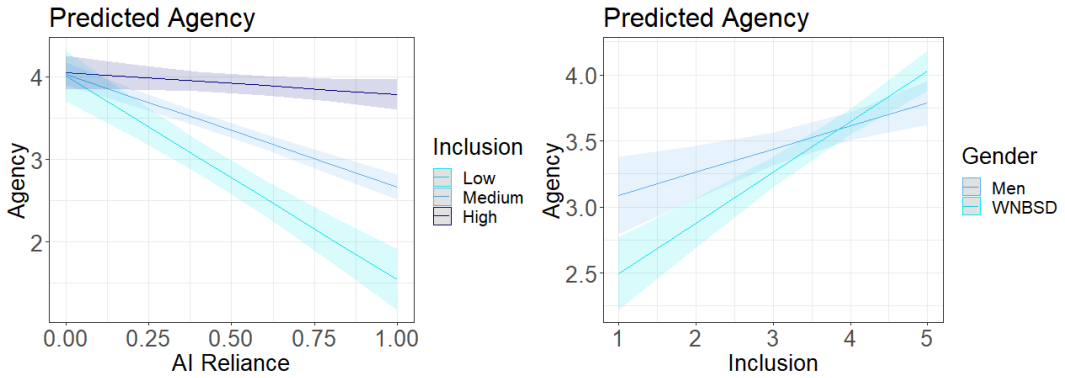
variables: age, gender, race, and income. The race and gender features were operationalized as binary variables where 1 reflected the privileged group and 0 reflected the minoritized group (e.g., for gender, 1=men, 0=WNBSD). We also mapped the AI writing style as a binary variable (0=hesitant model, 1=self-assured model). We operationalized the age and income demographic features as ordinal variables. Table 4.3 shows the correlation analysis for the predictors in our model.

Table 4.3: **Predictor Correlation Table.** Correlation table for the predictors in the OLS linear regression

	Age	Race	Income	Gender	AI Reliance	Inclusion	AI Style
Age	1.00						
Race	0.16	1.00					
Income	0.22	0.06	1.00				
Gender	-0.02	-0.05	0.08	1.00			
AI Reliance	0.04	-0.01	0.09	0.07	1.00		
Inclusion	-0.08	-0.12	-0.08	0.02	0.17	1.00	
AI Style	-0.00	0.03	0.04	-0.10	0.14	0.10	1.00

Table 4.4 shows the parameter estimates of a linear regression model predicting perceived agency. We tested several versions of the model that accounted for additional interactions (like AI writing style and AI reliance), and we present the model with the best fit. The model shows a main effect of AI writing style (Table 4.4, line 7) on agency. The negative value of the parameter signifies that writing with a self-assured model contributes to a loss of agency. We also see main effects of gender (line 4), showing men contributed to higher levels of agency. Perceptions of inclusion (line 6) also positively impacted agency, and AI reliance (line 5) had a negative impact. These effects were qualified by interaction effects between AI reliance and perceived inclusion (line 9) and gender and perceived inclusion (line 10), which we explore next.

We explore the interaction effects to better understand their impact on perceived agency. Figure 4.5a depicts the interaction between AI reliance and perceived inclusion. AI reliance is the x-axis, and perceived agency is the y-axis. The



(a) Participants who feel higher level of in-clusion by the writing assistant maintainof inclusion agency even when AI reliance increases

(b) Gender minorities have a stronger sense

Figure 4.5: **Interaction Effects in the Linear Regression**. Figure 4.5a depicts the interaction between Inclusion and AI Reliance. Figure 4.5b depicts the interaction between Gender and Inclusion.

graph compares the groups that showed high (dark blue), medium (blue), and low (light blue) levels of inclusion. Participants with a high level of inclusion, as indicated by the dark blue line in the figure, maintain their sense of agency even with increasing reliance on the writing assistant. Participants with a low (and to a lesser degree, medium) sense of inclusion lose their sense of agency rapidly when they rely more on AI.

Figure 4.5b explores how the interaction between inclusion (x-axis) and gender impacts agency. The graph compares men (blue line) and minoritized genders (WNBSD, light blue line). While inclusion and agency have a direct relationship for all gender identities, the slope for WNBSD participants is greater than the slope for men, indicating that perceived inclusion is more strongly tied to perceived agency for minoritized genders.

4.5 Discussion

Our findings enhance the current models of "writing with AI" by introducing new factors and furthering our understanding of their relationships. The findings also hint at the role that gender may play in such a context, though this line of investigation requires additional substantiation. Overall, the findings offer implications for research and design of co-writing interactions with AI.

By introducing the construct of *inclusion*, our work extends the previous research on AI-MC, which has examined perceptions of agency while interacting with an AI writing assistant [141, 140]. We measure the concept of inclusion after the co-writing process by asking participants if they could envision themselves in the model and whether the model's writing style matched their own. We show that this concept contributes to the perception of agency, a combined metric reflecting ownership and control over the writing process and the final product.

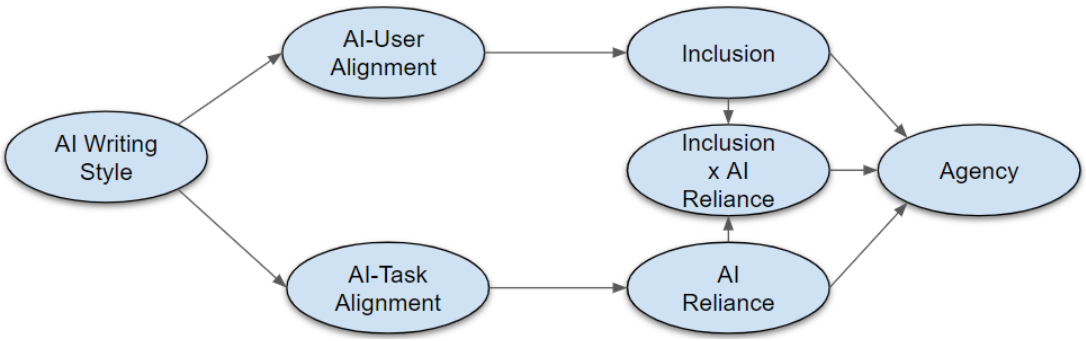


Figure 4.6: **Conceptual Model of Agency:** A proposed model of the constructs and measured variables that contribute to feelings of agency.

Informed by these findings, Figure 4.6 presents a new conceptual model of the factors contributing to agency in AI co-writing environments. This model is tentative: not all of these concepts were measured, let alone tested, in our work here. The model suggests that there are two paths for feeling agency in the co-writing process and its outcome. On the top, the AI writing style affects AI-user

alignment—which we define as the similarity between the AI writing style and the user's own preferences that may be derived from their identity. High AI-user alignment leads to increased perceived inclusion, positively contributing to perceived agency. In our research, we did not measure AI-user alignment. Instead, we built on the literature suggesting that minoritized genders are less likely to be aligned with assertive AI suggestions [137, 76]—which was not the case in this study. At the bottom of the model, the AI writing style affects AI-task alignment, which we define as the perceived usefulness of the AI assistant for the task (e.g., writing to ask for a promotion with an assertive AI). If there is high AI-task alignment, the user will rely on the model to complete the task. However, such reliance, as we have shown, can *negatively* contribute to the user's perceived agency. The conceptual model includes an interaction between AI reliance and inclusion, as suggested by our exploratory results.

Our model raises several open questions that have implications for designers. The most salient question is the potential competition between AI-User alignment, achieved through personalized models, and AI-Task alignment, achieved through generic models. Or how can developers strike a balance between models that are inclusive of various users but are also effective in completing the task? We note that even if the models align with both the user and task, there is remaining conflict: personalized models will ultimately contribute to agency by enabling users to express ideas in ways that reflect their communicative style, but generic models may encourage greater reliance on system outputs which will *undermine* agency. Since many generic models are already optimized for a task, here we focus on the benefits and risks in operationalizing personalized models.

By learning users' preferences and communication styles, personalized models can improve perceived usefulness at the individual level. Tailored suggestions may

reduce friction, increase efficiency, and help users articulate ideas in ways that feel authentic to their voice [106]. Beyond individual gains, personalization also carries societal benefits. Systems that adapt to diverse linguistic styles may better support communities historically marginalized by standardized systems [106].

However, these benefits are accompanied by significant risks. First, personalized models can introduce essentialism and profiling by inferring user characteristics and inadvertently reinforcing stereotypes [106]. As a result, users may also experience personalized double binds in which they must choose between personalization that risks reinforcing stereotypes or generic responses that erase their identity [200]. Prior work shows that personalized binds are unevenly distributed, with women reporting greater concern than men both about receiving overly generic responses and about encountering stereotypical outputs when interacting with chatbots [200]. Second, personalization introduces privacy risks and governance challenges. Creating personalized models requires continuous data collection and integration across multiple infrastructural layers, including retained chat histories, cross-session memory mechanisms, organization-level user profiling, and search-based personalization derived from browsing or query histories [106, 202, 200]. Taken together, personalized models offer both promising opportunities and significant risks. Our exploratory framework highlights constructs that were not validated in the main study. With this in mind, we turn to our primary results to examine whether AI–user alignment through personalization translated into differences in perceived inclusion and agency across gender groups.

The idea of inclusion via AI-User alignment captures our hypothesis that gender identity will play a role in perceived inclusion and agency. We hypothesized that minoritized genders would feel more included in the hesitant AI writing style condition, at least compared to men. However, our main findings did not show

such gender effects. Although existing literature highlights language use differs by gender in negotiation contexts, it often relies on experimental methods (such as laboratory studies) that encompass factors beyond the scope of our experiment like job status, vocal pitch, or physical presence [3, 133]. Many of the laboratory studies show it is the small details, like the tone of voice or the wording, which makes gender differences salient [192]. The absence of physical cues, as shown in computer-mediated negotiations, reduces gender-based communication bias [183].

While gender differences did not account for our expected main effect, our exploratory analysis also suggests there are still, potentially, *some* differences between genders. A noteworthy finding in this analysis was the significant interaction between gender and inclusion in relation to perceived agency. The interaction suggests that for minoritized genders, perceptions of inclusion have a greater effect on the perception of agency compared to men. If they hold, these findings lend support to our initial hypothesis: that people of different genders may react differently to writing assistants of different styles. This finding suggests that it may be useful to devise, develop, and test language models that also optimize on generating perceptions of inclusion by people of different demographic backgrounds.

Our work explored the impact of AI auto-complete suggestions on feelings of inclusion, control, and ownership over the written text and considered the implications for users of various backgrounds. The work has a number of limitations. First, in asking people to write as if asking for a promotion, our study focused on an important and realistic workplace task. However, other workplace writing tasks should also be studied to generalize the results. Furthermore, we did not measure individual differences in assertiveness, which could have provided additional insight on the effect of the treatment. Pre-task measurement of such attitudes, though, could have introduced bias to the experimental task. We focus on English-speaking

US residents. Additional work is needed to understand how users from various linguistic backgrounds and cultures interact with such AI agents. Despite these limitations, this study showed some of the first evidence that minoritized genders' perceptions of control and ownership could vary in some ways from men's as a response to writer-style biases. We offer a new conceptual model for understanding the impact of these factors on people's feelings of overall agency in AI-mediated environments.

Table 4.4: **OLS Linear Regression Analysis.** Predicting perceived agency based on demographic features, AI writing style, AI reliance, and perceived inclusion. The constant corresponds to the baseline perceived agency.

	Dependent variable:
	Agency
Age	0.024
	(0.020)
Race	0.013
	(0.070)
Income	0.023
	(0.034)
Gender	0.571*
	(0.261)
'AI Reliance'	−3.451***
	(0.367)
Inclusion	0.136*
	(0.066)
'AI Writing Style'	−0.244**
	(0.090)
Gender:'AI Reliance'	0.251
	(0.191)
'AI Reliance':Inclusion	0.666***
	(0.098)
Gender:Inclusion	−0.204**
	(0.071)
Gender:'AI Writing Style'	0.050
	(0.133)
Constant	3.661***
	(0.248)
Observations	489
R^2	0.354
Adjusted R^2	0.339
Residual Std. Error	0.715 (df = 477)
F Statistic	23.747*** (df = 11; 477)
Note:	*p<0.05; **p<0.01; ***p<0.001

CHAPTER 5

CONCLUSIONS

In this concluding chapter, I discuss opportunities to build upon the studies presented in this dissertation. I then present two key contributions of my research: demonstrating the importance of audience's perceptions when technology is embedded in communicative settings and the importance of human-centered evaluations. Lastly, I show how the contributions of this research can guide the design and evaluation of future, more advanced AI technologies toward more inclusive and equitable outcomes.

While the studies presented in this dissertation were designed with careful attention to experimental rigor, there are several opportunities to improve upon this work. In Chapter 2, we stated that audiences' perception of subtitle errors could amplify speakers' feelings of frustration, thereby exacerbating quality-of-service harms. However, the study did not measure speakers' experiences. Future work could empirically assess the speakers' experience of live subtitling during a presentation and the audience's perception of the subtitles to validate the prior statement. Additionally, while the experiment captured audience members' negative perceptions of subtitle errors, it did not examine whether audiences attributed these errors to the speaker or to the underlying system. Exploring who is blamed when errors occur could provide insight into the dynamics between the speaker and the audience when there are subtitles. For example, if the audience blames the system for the errors, maybe the speaker's feelings of frustration are alleviated. However, if the audience primarily blames the speaker when there are errors, the speaker's negative feelings could be amplified. It is also possible that the audience's judgments may not significantly impact the speaker's perception of their presentation or themselves. These are all important considerations for future work.

More broadly, these limitations point to the constraints of controlled experimental methods in capturing how dynamics unfold in real-world settings.

Building on this limitation, the dissertation would benefit from extending its findings beyond controlled online experiments to more ecologically valid, real-world contexts. In Chapter 3, we conducted an online experiment as an initial characterization of perceptual harms. Since the completion of that study, increasing attention has been paid to the use of AI tools in professional workflows. As a result, future research could extend this work through real-world field experiments in office environments to document perceptual harms as they unfold in practice. Similarly, the online experiment described in Chapter 4 captures only a snapshot of the users' experiences when co-writing with an AI assistant. Future work could instead adopt a diary-study methodology in which participants engage with an AI writing assistant in their everyday activities and reflect on their experiences over time, including whether and how the tool affects their sense of inclusion. Taken together, these proposed extensions highlight the importance of moving beyond controlled online experiments toward more ecologically valid, longitudinal approaches for understanding the complex and evolving impacts of AI systems.

Crucially, moving into real-world contexts is not only a matter of ecological validity, but also of uncovering the trade-offs that emerge when AI tools are deployed in practice. Across the studies in this dissertation, AI tools were implicitly treated as beneficial along specific instrumental dimensions. However, these assumed gains were not directly measured alongside the social phenomena under investigation in the real world. Future studies could jointly evaluate measurable improvements in productivity or task quality and the interpersonal or organizational costs that may accompany them. Such work would help clarify whether efficiency gains persist in practice, and under what conditions they are offset by social frictions. As an initial

first step to disentangle these trade-offs in a real world setting, researchers could investigate AI workslop. One emerging concern in workplaces is the phenomenon often described as "workslop" or AI-generated content that appears polished but lacks substantive value. Receiving low-quality or confusing AI-generated work can impose downstream costs on collaborators, who must invest additional time to interpret, correct, or re-do the work [153]. Beyond productivity losses, exposure to workslop may also damage trust, prompting recipients to reassess their colleagues' competence, creativity, or reliability, and in some cases to avoid future collaboration altogether [153]. Despite the workslop phenomena, companies are still pushing for AI adoption claiming that it improves worker efficiency and can increase profitability [19]. These dynamics suggest that the future of AI in the workplace will not be defined solely by whether tools increase individual efficiency, but by how they reshape social judgments, coordination, and professional norms in light of company incentives.

In light of these dynamics, understanding AI's impact in the workplace requires approaches that foreground human perception, evaluation, and experience. This dissertation contributes to that understanding by demonstrating that audience perceptions—often mediated by technology—play a consequential role in users' experiences and outcomes. When technology becomes part of how work is produced and evaluated, it also becomes part of how people are judged. As a result, harms can emerge not solely from system error or bias in model outputs, but from the interpretations and social meanings attached to those outputs. As shown in Chapter 2, users are penalized for technological errors. Chapter 3 extends this concept by illustrating how individuals can be disadvantaged when the technology use is merely perceived.

A second contribution of this dissertation is to show that incorporating human-

centered measures, alongside technical accuracy, enables more holistic and grounded evaluations of AI systems. Oftentimes, technical evaluations overlook the nuance and sensitivity with which people interact with technology, leading to discrepancies in outcomes even when measured performance appears equal across groups. Chapter 2 showed there are differential outcomes between various users when accounting for the *perceived quality* even when the error rates are similar. Chapter 4 demonstrated that although users spent comparable amounts of time with the writing assistant, their sense of inclusion—how acknowledged and supported they felt by the model's suggestions—significantly influenced their perception of agency, with inclusion having a particularly strong effect for users of different genders. These results underscore that technical accuracy alone cannot guarantee equitable or empowering interactions, making human-centered measures essential for more robust evaluation.

These contributions are especially salient as AI systems continue to evolve. Since this work began, there have been several advancements in AI models. One of the most notable has been the rise of AI agents like OpenAI's Operator (or ChatGPT agent mode) and Amazon Rufus. Broadly speaking, agentic AI systems are those in which a model can accomplish a provided task with minimal instruction and human oversight, often making decisions "independently" [25]. With increasingly agentic AI systems, there has also been a shift from viewing AI as a tool to a helpful team member that brings new ideas, identifies risks and suggests next steps [126, 82]. As AI systems become more embedded in decision-making and task execution, the social and experiential dimensions highlighted in this dissertation become increasingly consequential for how such systems are evaluated and deployed.

In this context, the findings on audience perceptions suggest that users may

be judged not only by the quality of their work but also by how others interpret their relationship to AI agents. Some individuals may be penalized simply because colleagues suspect they relied on an agent, while others may face no such scrutiny. In the context of agentic AI, this means that evaluation frameworks must account for how users are perceived within their teams or organizations, and how these perceptions shape outcomes such as credibility, trust, and access to opportunities. Focusing solely on system performance would obscure these social forms of harm. Similarly, the results on human-centered metrics underscore the need to assess the experiential impact of AI agents on users. When an agent generates content or executes tasks on a user's behalf, the agent's language and behavior can meaningfully influence the user's sense of participation. If users do not feel acknowledged or included by the agent, they may experience diminished agency, which can affect engagement and confidence when interacting with an agent. Evaluations of agentic AI should therefore incorporate measures of users' subjective experiences—such as inclusion and sense of control—alongside task-level metrics.

BIBLIOGRAPHY

[1] Abid, A., Farooqi, M., and Zou, J. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2021), AIES '21, Association for Computing Machinery, p. 298–306.

[2] Agarwal, D., Naaman, M., and Vashistha, A. Ai suggestions homogenize writing toward western styles and diminish cultural nuances, 2024.

[3] Amanatullah, E., and Morris, M. Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others. *Journal of personality and social psychology 98* (2010), 256–67.

[4] Amanatullah, E. T., and Tinsley, C. H. Ask and ye shall receive? how gender and status moderate negotiation success. *Negotiation and Conflict Management Research 6* (2013), 253–272.

[5] Apone, T., Brooks, M., and O'Connell, T. Caption accuracy metrics project. caption viewer survey: Error ranking of real-time captions in live television news programs, 2010.

[6] Arroyo Chavez, M., Feanny, M., Seita, M., Thompson, B., Delk, K., Officer, S., Glasser, A., Kushalnagar, R., and Vogler, C. How users experience closed captions on live television: Quality metrics remain a challenge. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[7] Baidoo-Anu, D., and Ansah, L. O. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI 7*, 1 (2023), 52–62.

[8] Baumann, J., Castelnovo, A., Crupi, R., Inverardi, N., and Regoli, D. Bias on demand: A modelling framework that generates synthetic data with bias. Association for Computing Machinery.

[9] Belk, R. W. Extended Self in a Digital World. *Journal of Consumer Research 40* (2013), 477–500.

[10] Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., and Seli, P. Humans

versus ai: whether and why we prefer human-created compared to ai-created artwork. *Cognitive Research: Principles and Implications 8*, 1 (2023), 42.

[11] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery.

[12] Bertrand, M., and Mullainathan, S. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*, 4 (2004), 991–1013.

[13] Bhat, A., Agashe, S., Oberoi, P., Mohile, N., Jangir, R., and Joshi, A. Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2023), IUI '23, Association for Computing Machinery, p. 436–452.

[14] Bhat, M., and Long, D. Designing interactive explainable ai tools for algorithmic literacy and transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (New York, NY, USA, 2024), DIS '24, Association for Computing Machinery, p. 939–957.

[15] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in ai. Tech. Rep. MSR-TR-2020-32, Microsoft, May 2020.

[16] Blaschke, V., Winkler, M., Foerster, C., Wenger-Glemser, G., and Plank, B. A multi-dialectal dataset for german dialect asr and dialect-to-standard speech translation. In *Proceedings of Interspeech 2025* (Aug. 2025), ISCA, pp. 913–917.

[17] Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., and D'Mello, S. K. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (New York, NY, USA, 2021), ICMI '21, Association for Computing Machinery, p. 268–277.

[18] Braun, N., Debener, S., Spychala, N., Bongartz, E., Sörös, P., Müller, H. H. O., and Philipsen, A. The senses of agency and ownership: A review. *Frontiers in Psychology 9* (2018).

[19] Brynjolfsson, E., Rock, D., and Syverson, C. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics 13*, 1 (2021), 333–372.

[20] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (2018), PMLR, pp. 77–91.

[21] Buschek, D., Zürn, M., and Eiband, M. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021), CHI '21, Association for Computing Machinery.

[22] Butler, J., Jaffe, S., Baym, N., Czerwinski, M., Iqbal, S., Nowak, K., Rintel, S., Sellen, A., Vorvoreanu, M., Abdulhamid, N. G., and et al. Microsoft new future of work report 2023, Apr 2024.

[23] Catenaccio, P. Press releases as a hybrid genre: Addressing the informative/promotional conundrum. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 18*, 1 (2008), 9–31.

[24] Cesafsky, L., Stayton, E., and Cefkin, M. Calibrating agency: Human-autonomy teaming and the future of work amid highly automated systems. In *Ethnographic Praxis in Industry Conference Proceedings* (2019), Wiley Online Library.

[25] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2023), FAccT '23, Association for Computing Machinery, p. 651–666.

[26] Chan, C. K. Y., and Hu, W. Students' voices on generative ai: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education 20*, 1 (2023), 43.

[27] Chan, W. S., Kruger, J.-L., and Doherty, S. Comparing the impact of automatically generated and corrected subtitles on cognitive load

and learning in a first-and second-language educational context. *Linguistica Antverpiensia, New Series–Themes in Translation Studies 18* (2019).

[28] Chandras, J. Multilingualismin india. *Education about Asia 25*, 3 (2020).

[29] Chatterji, R. The state of enterprise ai, Dec 2025.

[30] Chen, G. A., and Buell, J. Y. Of models and myths: Asian (americans) in stem and the neoliberal racial project. *Race Ethnicity and Education 21*, 5 (2018), 607–625.

[31] Chu, H., Kim, J., Kim, S., Lim, H., Lee, H., Jin, S., Lee, J., Kim, T., and Ko, S. An empirical study on how people perceive ai-generated music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (2022), pp. 304–314.

[32] Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online, Aug. 2021), C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Association for Computational Linguistics, pp. 7282–7296.

[33] Clark, J. M., and Paivio, A. Dual coding theory and education. *Educational psychology review 3*, 3 (1991), 149–210.

[34] Correll, S., Benard, S., and Paik, I. Getting a job: Is there a motherhood penalty? *American Journal of Sociology 112*, 5 (2007), 1297–1338.

[35] Council, A. I., Jun 2022.

[36] Crawford, K. Artificial intelligence's white guy problem. *New York Times* (June 2016).

[37] Cui, W., Zhu, S., Zhang, M. R., Schwartz, H. A., Wobbrock, J. O., and Bi, X. Justcorrect: Intelligent post hoc text correction techniques on smartphones. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2020), UIST '20, Association for Computing Machinery, p. 487–499.

[38] Cullen, A., and Harte, N. Thin slicing to predict viewer impressions of ted talks. In *AVSP* (2017), pp. 58–63.

[39] Cunningham, J., Blodgett, S. L., Madaio, M., Daumé Iii, H., Harrington, C., and Wallach, H. Understanding the impacts of language technologies' performance disparities on African American language speakers. In *Findings of the Association for Computational Linguistics: ACL 2024* (Bangkok, Thailand, Aug. 2024), L.-W. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, pp. 12826–12833.

[40] Curtis, K., Jones, G. J., and Campbell, N. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (New York, NY, USA, 2015), ICMI '15, Association for Computing Machinery, p. 35–42.

[41] Dervakos, E., Filandrianos, G., and Stamou, G. Heuristics for evaluation of ai generated music. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 9164–9171.

[42] Desai, A., Alharbi, R., Hsueh, S., Ladner, R. E., and Mankoff, J. Toward language justice: Exploring multilingual captioning for accessibility. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025), CHI '25, Association for Computing Machinery.

[43] Deshpande, K. V., Pan, S., and Foulds, J. R. Mitigating demographic bias in ai-based resume filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA, 2020), UMAP '20 Adjunct, Association for Computing Machinery, p. 268–275.

[44] Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., and Choi, Y. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin, Ireland, May 2022), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, pp. 7250–7274.

[45] Dourish, P., and Bellotti, V. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work* (New York, NY, USA, 1992), CSCW '92, Association for Computing Machinery, p. 107–114.

[46] Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., and Welsch, R. The ai ghostwriter effect: When users do not perceive ownership of ai-generated text but self-declare as authors. *ACM Trans. Comput.-Hum. Interact. 31*, 2 (feb 2024).

[47] (EBG), E. B. G. Report on efforts to improve the quality of closed captioning, 2014.

[48] Epstein, Z., Hertzmann, A., of Human Creativity, I., Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., et al. Art and the science of generative ai. *Science 380*, 6650 (2023), 1110–1111.

[49] Ezema, K., Chandler, C., Southwell, R., Cholendiran, N., and D'Mello, S. "it feels like we're not meeting the criteria": Examining and mitigating the cascading effects of bias in automatic speech recognition in spoken language interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025), CHI '25, Association for Computing Machinery.

[50] Fast, E., and Horvitz, E. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence* (2017), vol. 31.

[51] Feng, D., Yun, B., and Yi Wang, A. From junior to senior: Allocating agency and navigating professional growth in agentic ai-mediated software engineering. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2026), CHI '26, Association for Computing Machinery.

[52] Fink, K. The biggest challenge facing journalism: A lack of trust. *Journalism 20*, 1 (2019), 40–43.

[53] Foundation, N. S., May 2024.

[54] Frid, E., Gomes, C., and Jin, Z. Music creation by example. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–13.

[55] Friedman, B., and Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst. 14*, 3 (July 1996), 330–347.

[56] Fu, Y., Foell, S., Xu, X., and Hiniker, A. From text to self: Users' perception of aimc tools on interpersonal communication and self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[57] Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., and Giles, H. A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European journal of social psychology 42*, 1 (2012), 120–133.

[58] Gaddis, S. M. How black are lakisha and jamal? racial perceptions from names used in correspondence audit studies. *Sociological Science 4* (2017), 469–489.

[59] Gautam, S., Venkit, P. N., and Ghosh, S. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776* (2024).

[60] Gernsbacher, M. Video captions benefit everyone. policy insights from the behavioral and brain sciences, 2 (1), 195-202, 2015.

[61] Gero, K. I., Liu, V., and Chilton, L. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (New York, NY, USA, 2022), DIS '22, Association for Computing Machinery, p. 1002–1019.

[62] Ghosh, S., and Caliskan, A. 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. *arXiv preprint arXiv:2310.19981* (2023).

[63] Giering, O., and Kirchner, S. Artificial intelligence and autonomy at work: empirical insights from germany. *Journal for Labour Market Research 59*, 1 (2025), 20.

[64] Giglio, A. D., and Costa, M. U. P. d. The use of artificial intelligence to improve the scientific writing of non-native english speakers. *Revista da Associação Médica Brasileira 69*, 9 (2023), e20230560.

[65] Gluszek, A., and Dovidio, J. F. The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and social psychology review 14*, 2 (2010), 214–237.

[66] Gothi, R., and Rao, P. Improving automatic speech recognition withnbsp;dialect-specific language models. In *Speech and Computer: 25th International Conference, SPECOM 2023, Dharwad, India, November 29 – December 2, 2023, Proceedings, Part I* (Berlin, Heidelberg, 2023), Springer-Verlag, p. 57–67.

[67] Gotoman, J., Luna, H., Sangria, J., Santiago Jr, C., and Barbuco, D. Accuracy and reliability of ai-generated text detection tools: a literature review. *American Journal of IR 4*, 0 (2025), 1–9.

[68] Government, U. S.

[69] Gozalo-Brizuela, R., and Garrido-Merchan, E. C. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655* (2023).

[70] Hancock, J. T., Naaman, M., and Levy, K. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication 25*, 1 (2020), 89–100.

[71] Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (New York, NY, USA, 2017), CSCW '17, Association for Computing Machinery, p. 1914–1933.

[72] Harrington, C. N., Garg, R., Woodward, A., and Williams, D. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery.

[73] Harris, C., Mgbahurike, C., Kumar, N., and Yang, D. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (Miami, Florida, USA, Nov. 2024), Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Association for Computational Linguistics, pp. 15166–15184.

[74] Haslberger, M., Gingrich, J., and Bhatia, J. No great equalizer: experimental evidence on ai in the uk labor market. *Available at SSRN* (2023).

[75] Heilman, M. E., and Caleo, S. Gender discrimination in the workplace. *The Oxford handbook of workplace discrimination* (2018), 73–88.

[76] Herring, S., and Martinson, A. Assessing gender authenticity in computer-mediated language useevidence from an identity game. *Journal of Language and Social Psychology - J LANG SOC PSYCHOL 23* (2004), 424–446.

[77] Hickman, L., Langer, M., Saef, R. M., and Tay, L. Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology* (2024).

[78] Hogan, B., and Berry, B. Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & community 10*, 4 (2011), 351–372.

[79] Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., and Jung, M. F. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports 13* (2023), 5487.

[80] Hong, J.-W. Bias in perception of art produced by artificial intelligence. In *Human-Computer Interaction. Interaction in Context: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II* (Berlin, Heidelberg, 2018), Springer-Verlag, p. 290–303.

[81] House, W. Preparing for the future of artificial intelligence. executive office of the president national science and technology council. committee on technology, 2016.

[82] Huang, T., Chen, S., Chen, M., May, J., Yang, L., Wan, M., and Zhou, P. Teaching language models to gather information proactively. *Findings of the Association for Computational Linguistics: EMNLP 2025* (2025), 15588–15599.

[83] Hui, X., Reshef, O., and Zhou, L. The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Available at SSRN 4527336* (2023).

[84] Hutiri, W. T., and Ding, A. Y. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022), FAccT '22, Association for Computing Machinery, p. 230–247.

[85] Hwang, G., Lee, J., Oh, C. Y., and Lee, J. It sounds like a woman: Exploring gender stereotypes in south korean voice assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI EA '19, Association for Computing Machinery, p. 1–6.

[86] Hwang, S. I., Lim, J. S., Lee, R. W., Matsui, Y., Iguchi, T., Hiraki, T., and Ahn, H. Is chatgpt a "fire of prometheus" for non-native english-speaking researchers in academic writing? *Korean Journal of Radiology 24*, 10 (2023), 952.

[87] Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650* (2019).

[88] Jacobs, A. Z., and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 375–385.

[89] Jaffe, S., Shah, N. P., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M., and Teevan, J. Generative ai in real-world workplaces. Tech. Rep. MSR-TR-2024-29, Microsoft, July 2024.

[90] Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023), CHI '23, Association for Computing Machinery.

[91] Jakesch, M., Buçinca, Z., Amershi, S., and Olteanu, A. How different groups prioritize ethical values for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), pp. 310–323.

[92] Jakesch, M., French, M., Ma, X., Hancock, J. T., and Naaman, M. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[93] Jakesch, M., Hancock, J. T., and Naaman, M. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences 120*, 11 (2023), e2208839120.

[94] Jang, Y., Sakashita, M., Niinuma, K., and Gupta, A. Evolving enactions of expertise: Software engineers' evaluation and demonstration of coding expertise with ai coding assistants. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2026), CHI '26, Association for Computing Machinery.

[95] Jiang, H., and Nachum, O. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics* (2020), PMLR, pp. 702–712.

[96] Jiang, H. H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., and Gebru, T. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2023), AIES '23, Association for Computing Machinery, p. 363–374.

[97] Kadoma, K., Aubin Le Quere, M., Fu, X. J., Munsch, C., Metaxa, D., and Naaman, M. The role of inclusion, control, and ownership in workplace ai-mediated communication. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[98] Kafle, S., and Huenerfauth, M. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing, 2017.

[99] Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., Newman, D. T., and Woodruff, A. Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), pp. 627–637.

[100] Kelly, S., Kaye, S.-A., and Oviedo-Trespalacios, O. What factors contribute to the acceptance of artificial intelligence? a systematic review. *Telematics and Informatics 77* (2023), 101925.

[101] Kemp, A. Frequent use of ai in the workplace continued to rise in q4, Jan 2026.

[102] Kidd, C., and Birhane, A. How ai can distort human beliefs. *Science 380* (2023), 1222–1223.

[103] Kim, S., Le, D., Zheng, W., Singh, T., Arora, A., Zhai, X., Fuegen, C., Kalinli, O., and Seltzer, M. Evaluating User Perception of Speech

Recognition System Quality with Semantic Distance Metric. In *Interspeech 2022* (2022), pp. 3978–3982.

[104] Kim, S., Park, Y. S., Ahn, D., Kwak, J. M., and Kim, J. Is the same performance really the same?: Understanding how listeners perceive asr results differently according to the speaker's accent. *Proceedings of the ACM on Human-Computer Interaction 8*, CSCW1 (2024), 1–22.

[105] Kim, T., Molina, M. D., Rheu, M., Zhan, E. S., and Peng, W. One ai does not fit all: A cluster analysis of the laypeople's perception of ai roles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–20.

[106] Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence 6*, 4 (2024), 383–392.

[107] Kobiella, C., Flores López, Y. S., Waltenberger, F., Draxler, F., and Schmidt, A. "if the machine is as good as me, then what use am i?" – how the use of chatgpt changes young professionals' perception of productivity and accomplishment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[108] Köbis, N., and Mossink, L. D. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in human behavior 114* (2021), 106553.

[109] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences 117* (2020), 7684–7689.

[110] Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference* (New York, NY, USA, 2023), CI '23, Association for Computing Machinery, p. 12–24.

[111] Kruger, J.-L., Hefer, E., and Matthew, G. Measuring the impact of subtitles on cognitive load: eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (New York, NY, USA, 2013), ETSA '13, Association for Computing Machinery, p. 62–66.

[112] Kuhn, K., Kersken, V., Reuter, B., Egger, N., and Zimmermann, G. Measuring the accuracy of automatic speech recognition solutions. *ACM Trans. Access. Comput. 16*, 4 (Jan. 2024).

[113] Kushalnagar, R. S., Behm, G. W., Kelstone, A. W., and Ali, S. Tracked speech-to-text display: Enhancing accessibility and readability of real-time speech-to-text. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (New York, NY, USA, 2015), ASSETS '15, Association for Computing Machinery, p. 223–230.

[114] Laban, P., Vig, J., Hearst, M., Xiong, C., and Wu, C.-S. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2024), UIST '24, Association for Computing Machinery.

[115] Landivar, L. C. Disparities in stem employment by sex, race, and hispanic origin. *Education Review 29*, 6 (2013), 911–922.

[116] Lee, M., Liang, P., and Yang, Q. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery.

[117] Lev-Ari, S., and Keysar, B. Why don't we believe non-native speakers? the influence of accent on credibility. *Journal of experimental social psychology 46*, 6 (2010), 1093–1096.

[118] Levy, K. Data driven: truckers, technology, and the new workplace surveillance.

[119] Li, L., Wang, G., and Freeman, G. "it's about research. it's not about language": Understanding and designing for mitigating non-native english-speaking presenters' challenges in live qa sessions at academic conferences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025), CHI '25, Association for Computing Machinery.

[120] Li, T., and Chmiel, A. Automatic subtitles increase accuracy and decrease cognitive load in simultaneous interpreting. *Interpreting 26*, 2 (2024), 253–281.

[121] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. Gpt detectors are biased against non-native english writers. *Patterns 4*, 7 (2023).

[122] Lim, S., and Schmälzle, R. The effect of source disclosure on evaluation of ai-generated messages. *Computers in Human Behavior: Artificial Humans 2*, 1 (2024), 100058.

[123] Lima, L., Furtado, V., Furtado, E., and Almeida, V. Empirical analysis of bias in voice-based personal assistants. In *Companion Proceedings of The 2019 World Wide Web Conference* (New York, NY, USA, 2019), WWW '19, Association for Computing Machinery, p. 533–538.

[124] Liu, J., Xu, X., Li, Y., and Tan, Y. " generate" the future of work through ai: Empirical evidence from online labor markets. *arXiv preprint arXiv:2308.05201* (2023).

[125] Liu, T., and Aryadoust, V. Orchestrating teacher, peer, and self-feedback to enhance learners' cognitive, behavioral, and emotional engagement and public speaking competence. *Behavioral Sciences 14*, 8 (2024).

[126] Liu, X. B., Fang, S., Shi, W., Wu, C.-S., Igarashi, T., and Chen, X. A. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2025), CHI '25, Association for Computing Machinery.

[127] Liu, Y., Mittal, A., Yang, D., and Bruckman, A. Will ai console me when i lose my pet? understanding perceptions of ai-mediated email writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery.

[128] Long, D., and Magerko, B. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020), CHI '20, Association for Computing Machinery, p. 1–16.

[129] Longoni, C., Fradkin, A., Cian, L., and Pennycook, G. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022), FAccT '22, Association for Computing Machinery, p. 97–106.

[130] Lorenzoni, A., Faccio, R., and Navarrete, E. Does foreign-accented speech affect credibility? evidence from the illusory-truth paradigm. *Journal of Cognition 7*, 1 (2024), 26.

[131] Lu, X., Li, S., and Fujimoto, M. Automatic speech recognition. In *Speech-to-speech translation*. Springer, 2019, pp. 21–38.

[132] Lucy, L., and Bamman, D. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding* (Virtual, June 2021), Association for Computational Linguistics, pp. 48–55.

[133] Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., and Hertel, G. A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological bulletin 141* (2015), 85–104.

[134] McGee, E. O., Thakore, B. K., and LaBlance, S. S. The burden of being "model": Racialized experiences of asian stem college students. *Journal of Diversity in Higher Education 10*, 3 (2017), 253.

[135] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR) 54*, 6 (2021), 1–35.

[136] Mengesha, Z., Heldreth, C., Lahav, M., Sublewski, J., and Tuennerman, E. "i don't think these devices are very culturally sensitive." - the impact of errors on african americans in automated speech recognition. *Frontiers in Artificial Intelligence 26* (2021).

[137] Merchant, K. *How Men And Women Differ: Gender Differences in Communication Styles, Influence Tactics, and Leadership Styles*. Claremont McKenna College, 2012.

[138] Metaxa-Kakavouli, D., Wang, K., Landay, J. A., and Hancock, J. Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proceedings of the 2018 CHI Conference on human factors in computing systems* (2018), pp. 1–6.

[139] Michel, S., Kaur, S., Gillespie, S. E., Gleason, J., Wilson, C., and Ghosh, A. "it's not a representation of me": Examining accent bias and digital exclusion in synthetic ai voice services. In *Proceedings of the*

2025 ACM Conference on Fairness, Accountability, and Transparency (New York, NY, USA, 2025), FAccT '25, Association for Computing Machinery, p. 228–245.

[140] Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., and Hohenstein, J. Ai-mediated communication: Language use and interpersonal effects in a referential communication task. *Proc. ACM Hum.-Comput. Interact. 5*, CSCW1 (apr 2021).

[141] Mieczkowski, H. N. *AI-Mediated Communication: Examining Agency, Ownership, Expertise, and Roles of AI Systems*. PhD thesis, 2022.

[142] Mileva, M., and Lavan, N. Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General 152*, 6 (2023), 1539.

[143] Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. CHI '23, Association for Computing Machinery.

[144] Mishra, T., Ljolje, A., and Gilbert, M. Predicting human perceived accuracy of asr systems. In *INTERSPEECH* (2011), pp. 1945–1948.

[145] Morgan, H. K., Purkiss, J. A., Porter, A. C., Lypson, M. L., Santen, S. A., Christner, J. G., Grum, C. M., and Hammoud, M. M. Student evaluation of faculty physicians: gender differences in teaching evaluations. *Journal of Women's Health 25*, 5 (2016), 453–456.

[146] Mullany, L. *Gendered discourse in the professional workplace*. Springer, 2007.

[147] Murphy, N. A., and Hall, J. A. Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement. *Frontiers in psychology 12* (2021), 667326.

[148] Nadeem, M., Bethke, A., and Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Online, Aug. 2021), Association for Computational Linguistics, pp. 5356–5371.

[149] Neff, G., and Nagy, P. Agency in the digital age: Using symbiotic agency to explain human-technology interaction. In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, Z. Papacharissi, Ed. Routledge, 2018, p. 11.

[150] Neri, H., and Cozman, F. The role of experts in the public perception of risk of artificial intelligence. *AI & society 35*, 3 (2020), 663–673.

[151] Newman, A., Chrispal, S., Dunwoodie, K., and Macaulay, L. Hidden bias, overt impact: A systematic review of the empirical literature on racial microaggressions at work. *Journal of Business Ethics 200*, 4 (2025), 753–772.

[152] Ngueajio, M. K., and Washington, G. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *HCI International 2022 – Late Breaking Papers: Interacting with EXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings* (Berlin, Heidelberg, 2022), Springer-Verlag, p. 421–440.

[153] Niederhoffer, K., Rosen, G., Lee, A., Liebscher, A., Rapuano, K., and Hancock, J. T. subscribe sign in latest magazine topics podcasts store reading lists data visuals case selections hbr executive generative ai ai-generated "workslop" is destroying productivity, Sep 2025.

[154] Noy, S., and Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381* (2023), 187–192.

[155] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*, 6464 (2019), 447–453.

[156] Pedersen, J. The far model: assessing quality in interlingual subtitling. *Jostrans: The Journal of Specialised Translation*, 28 (2017), 210–229.

[157] Perkins, M. Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond. *Journal of University Teaching and Learning Practice 20*, 2 (2023).

[158] Pipek, V., and Wulf, V. Infrastructuring: Toward an integrated perspective on the design and use of information technology. *J. Assoc. Inf. Syst. 10* (2009), 1.

[159] Plexico, L. W., Hamilton, M.-B., Hawkins, H., and Erath, S. The influence of workplace discrimination and vigilance on job satisfaction with people who stutter. *Journal of Fluency Disorders 62* (2019), 105725.

[160] Poddar, R., Sinha, R., Naaman, M., and Jakesch, M. Ai writing assistants influence topic choice in self-presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023), CHI EA '23, Association for Computing Machinery.

[161] Rae, I. The effects of perceived ai use on content perceptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[162] Ray, J. Americans express real concerns about artificial intelligence, Aug 2024.

[163] Romero-Fresco, P. Accessing communication: The quality of live subtitles in the uk. *Language Communication 49* (2016), 56–69.

[164] Romero-Fresco, P., and Fresno, N. The accuracy of automatic and human live captions in english. *Linguistica Antverpiensia, New Series–Themes in Translation Studies 22* (2023).

[165] Romero-Fresco, P., and Van Gauwbergen, Y. Fit for what purpose? ner certification of automatic captions in english and spanish. *Applied Sciences 15*, 3 (2025), 1387.

[166] Ross, M. Q., and Bayer, J. B. Explicating self-phones: Dimensions and correlates of smartphone self-extension. *Mobile Media & Communication 9* (2021), 488–512.

[167] Rothschild, D., and Klingenberg, F. Self and peer evaluation of writing in the interactive esl classroom: An exploratory study. *TESL Canada Journal* (1990), 52–65.

[168] Ryan, R. M., and Deci, E. L. *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford publications, 2017.

[169] Sager, C. Sager writing scale.

[170] Sashaborm. Thispersondoesnotexist - random ai generated photos of fake persons, Apr 2021.

[171] Sears, B., Castleberry, N., Lin, A., and Mallory, C. Lgbtq people's experiences of workplace discrimination and harassment.

[172] Segal, E. New report provides reality check about freelancers in the workforce, Dec 2023.

[173] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2019), FAT* '19, Association for Computing Machinery, p. 59–68.

[174] Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2023), AIES '23, Association for Computing Machinery, p. 723–741.

[175] Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.-H., Qiwei, L., Rakshit, S., Si, C., Xie, Y., Bigham, J. P., Bentley, F., Chai, J., Lipton, Z., Mei, Q., Mihalcea, R., Terry, M., Yang, D., Morris, M. R., Resnick, P., and Jurgens, D. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024.

[176] Shimogori, N., Ikeda, T., and Tsuboi, S. Automatically generated captions: will they help non-native speakers communicate in english? In *Proceedings of the 3rd International Conference on Intercultural Collaboration* (New York, NY, USA, 2010), ICIC '10, Association for Computing Machinery, p. 79–86.

[177] Singh, A., and Joachims, T. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2018), KDD '18, Association for Computing Machinery, p. 2219–2228.

[178] Singh, N., Bernal, G., Savchenko, D., and Glassman, E. L. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Trans. Comput.-Hum. Interact. 30*, 5 (sep 2023).

[179] Slaughter, I., Greenberg, C., Schwartz, R., and Caliskan, A. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (Singapore, Dec. 2023), H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, pp. 8967–8989.

[180] Spence, J. L., Hornsey, M. J., Stephenson, E. M., and Imuta, K. Is your accent right for the job? a meta-analysis on accent bias in hiring decisions. *Personality and Social Psychology Bulletin 50*, 3 (2024), 371–386.

[181] Sridhar, K. K. Language in education: Minorities and multilingualism in india. *International Review of Education 42*, 4 (1996), 327–347.

[182] Stainback, K. Organizations, employment discrimination, and inequality. In *The Oxford Handbook of Workplace Discrimination*. Oxford University Press, 01 2018.

[183] Stuhlmacher, A. F., and Walters, A. E. Gender differences in negotiation outcome: A meta-analysis. *Personnel Psychology 52* (1999), 653–677.

[184] Sullivan, M., Kelly, A., and McLaughlan, P. Chatgpt in higher education: Considerations for academic integrity and student learning.

[185] Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., and Yang, S. Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication 29*, 1 (02 2024), zmad045.

[186] Suresh, H., and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (New York, NY, USA, 2021), EAAMO '21, Association for Computing Machinery.

[187] Synnaeve, G., Xu, Q., Kahn, J., Likhomanenko, T., Grave, E., Pratap, V., Sriram, A., Liptchinsky, V., and Collobert, R. End-to-end asr: from supervised to semi-supervised learning with modern architectures, 2020.

[188] Szarkowska, A., Ragni, V., Szkriba, S., Black, S., Kruger, J.-L., and Orrego-Carmona, D. 'that's not what they said!' the impact

of incongruities between the dialogue and intralingual subtitles on viewer experience. *Perspectives 0*, 0 (2024), 1–20.

[189] Tatman, R., and Kasten, C. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech* (2017).

[190] Taylor, G. Perceived processing strategies of students watching captioned video. *Foreign Language Annals 38*, 3 (2005), 422–427.

[191] Team, T. U., 2024.

[192] Thimm, C., Koch, S. C., and Schey, S. Communicating gendered professional identity: Competence, cooperation, and conflict in the workplace. *The handbook of language and gender* (2003), 528–549.

[193] Tursunbayeva, A., and Chalutz-Ben Gal, H. Adoption of artificial intelligence: A top framework-based checklist for digital leaders. *Business Horizons 67*, 4 (2024), 357–368.

[194] Union, I. T.

[195] Varma, R. India-born in the us science and engineering workforce. *American Behavioral Scientist 53*, 7 (2010), 1064–1078.

[196] Vue, Z., Vang, C., Vue, N., Kamalumpundi, V., Barongan, T., Shao, B., Huang, S., Vang, L., Vue, M., Vang, N., Shao, J., Coombes, C., Katti, P., Liu, K., Yoshimura, K., Biete, M., Dai, D.-F., Phillips, M. A., and Behringer, R. R. Asian americans in stem are not a monolith. *Cell 186*, 15 (2023), 3138–3142.

[197] Waes, L., Leijten, M., and Remael, A. Live subtitling with speech recognition. causes and consequences of text reduction. *Across languages and cultures 14*, 1 (2013), 15–46.

[198] Waldman, D. A., and Avolio, B. J. Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of applied psychology 76*, 6 (1991), 897.

[199] Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational*

Linguistics: EMNLP 2023 (Dec. 2023), Association for Computational Linguistics.

[200] Wang, A., Beeghly, E., Koyejo, S., and Ho, D. E. Personalization double binds: When user preferences meet group-based chatbot behaviors. *arXiv preprint* (2025).

[201] Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

[202] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2022), FAccT '22, Association for Computing Machinery, p. 214–229.

[203] Wenzel, K., Devireddy, N., Davison, C., and Kaufman, G. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023), CHI '23, Association for Computing Machinery.

[204] Wenzel, K., and Kaufman, G. Designing for harm reduction: Communication repair for multicultural users' voice interactions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[205] Wilken, P., Georgakopoulou, P., and Matusov, E. SubER - a metric for automatic evaluation of subtitle quality. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)* (Dublin, Ireland (in-person and online), May 2022), E. Salesky, M. Federico, and M. Costa-jussà, Eds., Association for Computational Linguistics, pp. 1–10.

[206] Wilson, K., and Caliskan, A. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2024), vol. 7, pp. 1578–1590.

[207] Wu, Y., and Kelly, R. M. Online dating meets artificial intelligence: How the perception of algorithmically generated profile text impacts attractiveness and trust. In *Proceedings of the 32nd Australian Conference on Human-Computer Interaction* (New York, NY, USA, 2021), OzCHI '20, Association for Computing Machinery, p. 444–453.

[208] Xie, Y., Wu, S., and Chakravarty, S. Ai meets ai: Artificial intelligence and academic integrity - a survey on mitigating ai-assisted cheating in computing education. In *Proceedings of the 24th Annual Conference on Information Technology Education* (New York, NY, USA, 2023), SIGITE '23, Association for Computing Machinery, p. 79–83.

[209] Xueliang, M., and Qingxia, D. Language and nationality. *Chinese Sociology & Anthropology 21*, 1 (1988), 81–104.

[210] Yuan, A., Coenen, A., Reif, E., and Ippolito, D. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2022), IUI '22, Association for Computing Machinery, p. 841–852.

[211] Zhou, E., and Lee, D. Generative artificial intelligence, human creativity, and art. *PNAS Nexus 3*, 3 (03 2024), pgae052.