Generative AI and Perceptual Harms: Who's Suspected of using LLMs?

KOWE KADOMA, Cornell University, USA
DANAË METAXA, University of Pennsylvania, USA
MOR NAAMAN, Cornell Tech, USA

Large language models (LLMs) are increasingly integrated into a variety of writing tasks. While these tools can help people by generating ideas or producing higher quality work, like many other AI tools they may risk causing a variety of harms, disproportionately burdening historically marginalized groups. In this work, we introduce and evaluate *perceptual harm*, a term for the harm caused to users when others perceive or suspect them of using AI. We examined perceptual harms in three online experiments, each of which entailed human participants evaluating the profiles for fictional freelance writers. We asked participants whether they suspected the freelancers of using AI, the quality of their writing, and whether they should be hired. We found some support for perceptual harms against for certain demographic groups, but that perceptions of AI use negatively impacted writing evaluations and hiring outcomes across the board.

CCS Concepts: • **General and reference** → *Empirical studies*; • **Human-centered computing**;

1 INTRODUCTION

The public release of generative artificial intelligence (AI) technologies like ChatGPT, Claude, Stable Diffusion, and Midjourney has lead to widespread adoption. The general public is using generative AI for a variety of tasks from inspiring ideas [28, 85, 95] and revising text [16] to creating images [20, 96] and popular music [13, 17, 24]. 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 worker's productivity and reduce the skill gap between between employees [35, 74]. In education, generative AI can support engaging, personalized learning [3, 11].

Despite the 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 [83, 91]. Many of the identified harms stem from biases that arise during the model development, when decisions about training data have downstream impacts on the models outputs [47, 68]. Furthermore, the majority of these harms directly impact the people interfacing with AI systems. For example, users may encounter stereotypical text in language models [1] or stereotypical images in text-to-image models [27, 29, 87] when prompting the model 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 type of potential implicit harm: perceptual harm. 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 models outputs, they can be categorized as a type of societal harm [83] which reflects the adverse macro-level effects of algorithmic systems that exacerbate inequality.

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 enforced and that people cannot reliably identify AI-generated content [19, 46]. Previous research in AI-Mediated Communication [33] has demonstrated, in multiple settings, that the suspicion that a communication partner used AI to create content results in reduced evaluations of that partner [45, 71]. Furthermore, when people

Authors' addresses: Kowe Kadoma, kk696@cornell.edu, Cornell University, Ithaca, NY, USA; Danaë Metaxa, metaxa@seas.upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Mor Naaman, mor.naaman@cornell.edu, Cornell Tech, New York, New York, USA.

use AI, they are seen as less intentional [38], and their work is seen as less enjoyable and having lower quality [23]. We categorize these results from prior work as examples of perceptual harms, but these papers did not reflect on any differential treatment. In this context, we offer a framework to understand the various aspects of perceptual harms and how to study the impacts they may have on different groups.

We measure perceptual harms across three axes: suspicion of AI use; evaluation of content; and outcomes. These axes were chosen based on literature highlighting people's inability to distinguish between AI-generated and human-generated content, which can influence how they assess quality [46, 53]. Additionally, we want to understand the potential real world impact of these perceptions on outcomes for the people being evaluated. We propose that perceptual harms may function independently across these three axes. For example, men might be more readily suspected of using AI than women, but women's writing could still be judged more negatively than men's when AI use is suspected. Investigating perceptual harms across these three dimensions, this paper 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 studies, each of which addresses all of the research questions, with each study focusing on a different identity category: Study 1 focuses on gender, Study 2 on race, and Study 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 important context of study. Online freelance marketplaces are popular platforms where customers can hire workers for a variety of tasks [34]. Online marketplaces have a significant economic impact by providing additional income for many workers, with some relying on them as their primary source of income [34]. While LLMs could help freelancers complete tasks more efficiently (and potentially increase earnings), they have also increased competition within the marketplace as traditional customers are opting to use the LLM as opposed to human services [41, 62]. In an already competitive marketplace, where prior work has shown women and Black workers face discrimination in online platforms [34], perceptual harms could further disenfranchise these groups.

This paper addresses the research questions above using a carefully designed online experiment 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, so participants saw the (fictional) profiles from all demographic groups (e.g., women and men in Study 1) in a randomized order. After reading the writing content on each freelancer's profile, 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 Study 1) and non-dominant groups (foreign nationals in Study 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 of how to consider and evaluate perceptual harms in Generative AI and LLM applications having outlined a new way to think about the adverse impacts of these technologies.

2 RELATED WORK

We situate this study in two main areas of related literature: biases in sociotechnical systems broadly, and research specifically about AI-mediated communication. Having described this prior work, we also explain our hypothesized findings.

2.1 Harms in Sociotechnical Systems

Sociotechnical systems—systems that are a combination of technical and social components—have been extensively studied for biases, especially with the integration of machine learning and AI [82, 83]. In the context of computing, many potential harms caused by these systems are seen as byproducts of the models' construction, and in particular the training data used [47, 68, 83, 91]. For example, language models can produce stereotypical text or demeaning language which is often directed at historically marginalized groups like women [54, 66] and racial and ethnic minorities [1, 72]. Similarly in image generation, text-to-image models can produce stereotypical images or fail to recognize particular identities [27, 29, 87].

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 [83]. 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 [69, 92]. Additionally, users of marginalized backgrounds may also not feel included by the tool, which can undermine their sense of agency [50]. Prior work has even found that visual cues in UIs can signal belongingness (or lack thereof) to users of different gender groups [70].

In this paper, we propose a new genre of potential harm, one that is situated neither in these technologies themselves, nor 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).

2.2 AI-Mediated Communication

The broader relationship between AI technologies and people—including those using it to generate content, and those consuming that content—has received some attention in fields outside computing, like Communication. A growing body of work names this entire sphere of interaction *AI-mediated communication* (AI-MC) [33].

Some AI-MC work has specifically focused on AI tools and trust (not to be confused with computing research on trustworthy AI, a term describing the development of transparent and data-privacy-respecting technologies [58, 59]). In AI-MC, researchers have empirically shown that the perception a self-presentation profiles was written AI can negatively impact the writer's trustworthiness [45]. Similarly, in the context of emailing, recipients' trust decreased when they were told AI was involved in the writing process [63].

Notably, this decrease in trust is true despite extensive literature showing that most people cannot reliably differentiate between human-written text and AI-generated text [14, 18, 43, 46].

However, when people perceive content to be AI-generated, it is seen as grammatically flawed with verbose language [25, 46].

Although prior work has not conceptualized these trends under a single umbrella, we propose that the differential impacts on those perceived as using AI, especially at the current moment when the use of AI in settings like the workplace is frowned upon, can be described as a new kind of harm: perceptual harm. Moreover, expanding on prior work that has focused on individual-level impacts, we see the potential for stereotypes about different social groups to result in harms against entire identity groups and not just individuals.

2.3 Hypotheses

While most harms impact historically marginalized groups, in the context of AI suspicion, we hypothesize the dominant group will be suspected. While counterintuitive, there is evidence to suggest that white men will be suspected of using AI since they are overrepresented in the technology sector [15, 39, 56]. However, in Study 3 (nationality), we hypothesize the non-dominant group will be suspected. While nationality is a political construct, it is deeply intertwined with culture and language [94]. In the United States, although there is no official language, English remains the dominant medium of communication [31]. Prior work has shown that non-native English speakers are less likely to produce English without mistakes and greatly benefit from AI assistance [8, 30, 42]. In this case, then, we believe the assumptions about nationality and language will overpower the effect we expect in Study 1 and 2.

For the content evaluation and outcomes measurements, we hypothesize the non-dominant group will receive lower evaluations and will be less likely to be recommended for hiring, controlling for whether they were suspected of using AI. This hypothesis aligns with prior work that shows discrimination in writing evaluation [49] and hiring [5].

Our hypotheses across all studies are thus:

- H1: In race and gender contexts, the dominant group will be more suspected of using AI; however, in the context of nationality, the non-dominant group will be more suspected of using AI
- H2: When both groups are suspected of using AI, the non-dominant group will receive lower evaluation scores
- H3: When both groups are suspected of using AI, the non-dominant group is less likely to be hired compared to the dominant group

3 METHODS

We addressed our research questions using a realistic within-subjects online experiment where we asked crowdworkers 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 3.1) and their content (Section 3.2), before describing our measurements (Section 3.3) and experimental procedure (Section 3.4) in more detail. We then describe participant recruitment in Section 3.5. The research design was approved by [University Redacted] IRB, and pre-registered on OSF[1].

3.1 Profile Creation

We manipulated the demographic information of the "freelancers" in our profile presentations to participants by using specific names and photos. Our presentation parallels the design of popular freelancing sites like Fiverr or TaskRabbit, which do not explicitly state workers' gender, race, or

[1]https://osf.io/69dn8/?view_only=561dcd19f6a94d22984fa52be4eafcf6

nationality but feature names and photos on each profile. The selected names in Study 1 (gender) and Study 2 (race) came from a list compiled by Gaddis [26]. The list provides a set of first names from the New York State birth record data categorized by gender, race, and socioeconomic status (SES). The surnames come from the US Census Bureau, which provides data on the racial composition of the last names. For Study 1, we randomly selected two distinctly masculine first names and distinctly feminine first names that are of the same race (White) and SES (high) from the initial list from Gaddis [26]. The first names were then paired with common surnames that have a high population-level occurrence for whites in the US Census. We repeated the same procedure, with masculine names and a high SES, to select the first and last names for the Black profiles in Study 2.

In Study 3 (nationality) we used a list of names compiled by Hogan [36]. 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 study, we randomly chose two distinctly East and Southeast Asian male names from this list. The final set of names used in each study is shown in Table 1.

Table 1. **Profile Names**. Each column contains the profile names used in each study.

Study 1	Study 2	Study 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

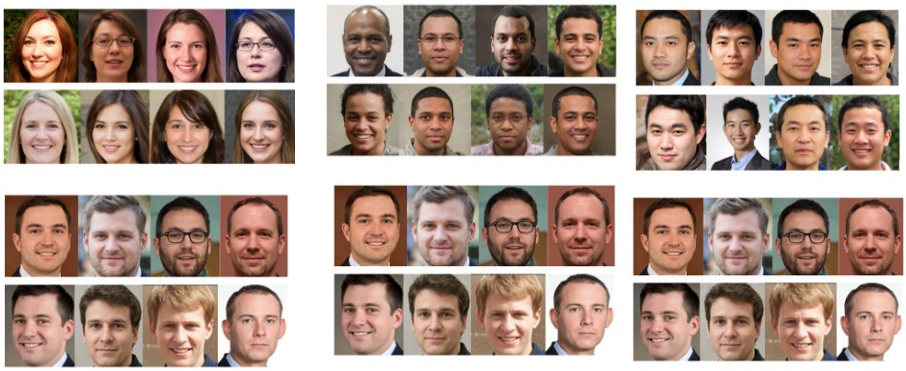
To create the visual representations of our freelancers, we used the AI-based image generator ThisPersonDoesNotExist [80]. We created a distinct set of eight images in a professional headshot style for each demographic group in our studies. The photos featured (fake) individuals in their mid- to late-30s, which is the age of most workers on freelance platforms [81]. 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.

The interface screen shown to participants was modeled after the profile pages of workers on Guru, a popular freelance site. Figure 2 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.

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 [81, 88]. Within marketing, we chose press releases for our evaluation content, rather than other writing tasks, because they are indented to be read by a general audience. Additionally, press releases have elements of creativity in self-promotion while also being informative [10] 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 various, real-life press releases. We then simplified the structure of the existing templates by removing datelines, subject lines, logos, and company contact information. The



(a) Study 1 Profile Photos (b) Study 2 Profile Photos (c) Study 3 Profile Photos

Fig. 1. **Profile Photos**. The two image sets, one per demographic group, used in each study.

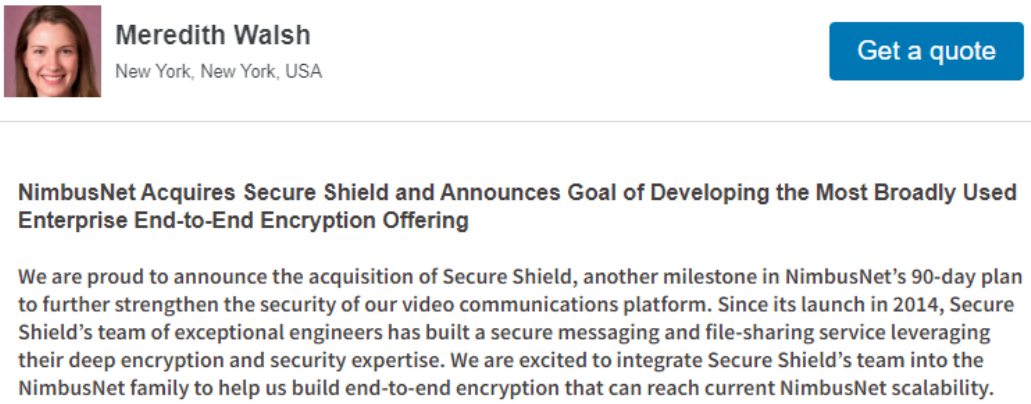


Fig. 2. **Screenshot of the Profile Interface.** The top panel contains the profile photo, name, and location of the (fake) freelancing marketing professional. The sample press release, supposedly written by this freelancer, is below.

resulting set of templates for each press release type can be found in Table 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.

The content in all of the press releases was human-written; however, in order to arouse participants' suspicion of AI use, we manually modified half of the press releases across all press release types to sound more AI-like. We do this modification manually for several reasons. First, LLMs are known to "hallucinate", producing plausible but inaccurate or unfaithful information [40, 83, 91]. Using fully AI-generated press releases could have introduced confounding variables, such as factual inaccuracies or unintended deviations from the structure of existing press release content, complicating our experimental analysis. Second, 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 [46]. We used these heuristics, namely, verbose language and rare or long words [25, 46], to modify half of the press releases to sound more AI-like. Given the fact that in most human-AI co-writing situations, people typically write the majority of the text themselves and selectively incorporate AI suggestions [6], we opted to partially modify the human-written press releases (as opposed to modifying the whole press release), ensuring content that this faithful to the original while still mimicking AI-like characteristics.

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 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 become "Encounter affable villagers and assist them in reconstructing their homes and workshops." We call these profiles *AI-inducing* profiles and evaluate whether they had a different effect than the control set of profiles (controlling for any changes in quality evaluation that results from these edits).

Table 2. **Press Release Templates**. The structure of each press release per column.

Product Launch	Event	Acquisition or Partnership	New Hire
- Title - Product overview - Uniqueness of the product	- Title - Event challenges - Future projection	- Title - Company overview - Message from CEO	- Title - Overview of previous leadership - Announcement of new leadership

3.3 Measures

We designed the experiment to examine the perceptual harms caused by generative AI. To do that, our direct measures included AI suspicion, variables related to content evaluation, and a measure of potential job (hiring) outcome.

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 [45]. In our work, we asked participants to review the profiles of online freelancers and stated the '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 motivation for their AI suspicion scores through a free response.

Overall Quality. Participants were asked to provide a rating for the overall quality of the evaluation 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.

Content. We draw from education literature, specifically writing evaluations for English as a Second Language (ESL) learners, to develop our more concrete writing evaluation measures. Content captures ideas or information conveyed in the message [78, 79]. To develop our content measures, we omitted statements that focused on details in the writing [78, 79]. 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.

Structure. While content captures *what* ideas are conveyed in the writing, structure captures *how* the idea is expressed [78]. We modified the statements from Rothschild [78] 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. 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).

3.4 Experiment Procedure

In each experiment, participants saw a demographically balanced set of four (2 demographic group x 2 writing style) (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 the likelihood they would hire the freelancer for a similar task (RQ3—outcomes). At the conclusion of the study, we asked participants about their demographic background and familiarity and attitudes toward artificial intelligence.

3.5 Participant Recruitment

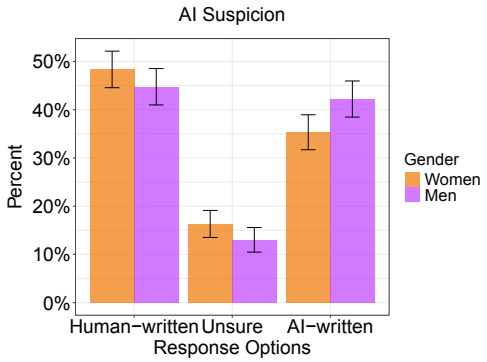
We recruited our participants from a US-representative sample of English-speaking adults on the crowdsourcing service Prolific. The sample size for each study was calculated based on linear mixed model (LMM) with 80% power. Each study had its own unique set of participants. We recruited 350 participants for Study 1 and Study 2 and 300 participants in Study 3. In all studies, we excluded participants who failed attention checks or correctly guessed the purpose of the study to ensure high quality responses in our analysis. We had usable data from 334 participants in both Study 1 and Study 2 and 272 participants in Study 3. Across all three studies, 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 details on participant's demographics can be found in Section A.2 of the Appendix.

4 RESULTS

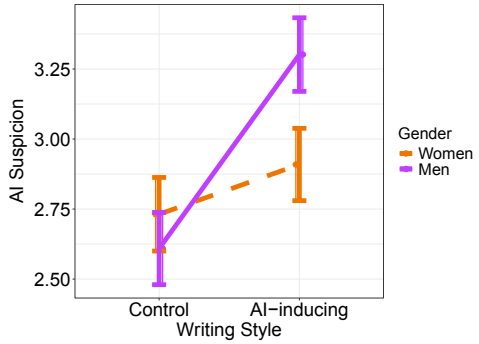
Our results provide some evidence of perceptual harms. Our three studies confirmed that different groups are suspected of AI use at varying rates (H1), but we did not find conclusive evidence that perceptual harms impacted evaluations of writing content (H2), or resulted in different hiring outcomes for different groups (H3). The experiments do, however, provide strong evidence that people associate AI-stylized writing with lower quality regardless of the author's race or nationality, which in turn negatively impacts job opportunities. The one notable exception to this pattern was when AI-stylized writing was attributed to women, in which case AI suspicion did not increase much; participants seemed willing to believe a woman had produced the poorly-received writing herself.

4.1 Study 1: Gender

Study 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.



(a) Participants are more likely to suspect AI use when evaluating men



(b) AI-stylized writing led participants to suspect the freelancer had used AI when the freelancer was a man, but not when the freelancer was a woman.

Fig. 3. **Participant's AI Suspicion**. Figure 3a depicts the distribution of the responses. Figure 3b depicts the interaction between writing style and gender.

AI Suspicion. Figure 3a 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 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 right most, purple column in the figure). In contrast, when the freelancer was a woman (in the right most, orange column), just 35.3% of participants thought the writing was AI generated. Evaluating this difference statistically, we used a linear mixed model with gender as fixed effect and included participant as a random effect to account for the repeated measures design and individual variability. This model confirmed that men freelancers were statistically significantly more likely to be suspected of using AI compared to women ($p = 0.037$).

We next examined how inducing AI suspicion via the language used in the writing samples affected AI suspicion. Figure 3b illustrates the interaction of gender (men and women) and writing style (AI-inducing or control) affected AI suspicion. 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 fixed effects of gender, writing style (AI-inducing or control), and their interaction, and with participant as a random effect. Our results show that gender had a significant main effect ($p < 0.001$) on AI suspicion, and there was also a significant interaction effect between gender and

writing style ($p < 0.001$), as reflected in the figure. The effect of writing style alone on AI suspicion also trended towards significant (p=0.058).

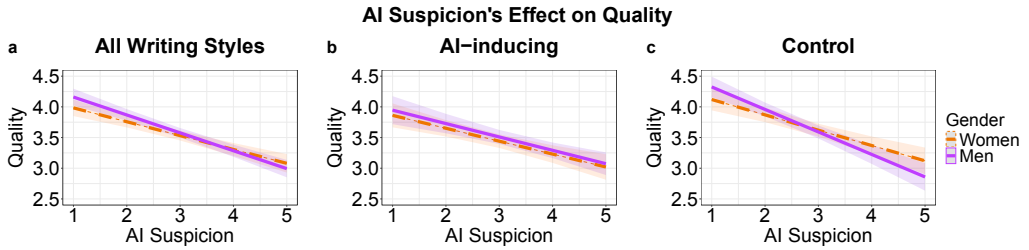


Fig. 4. **The Effect of AI Suspicion on Quality Evaluations.** AI Suspicion negatively impacts evaluations of writing quality.

Content Evaluations and Hiring.

Figure 4 shows the relationship between AI suspicion (x-axis) and overall quality ratings (y-axis) when participants evaluated men (solid, purple) and 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 4a, on the left), as well as the results from the two conditions separately, the AI-inducing style treatment (Fig 4b, in the middle) and the control (Fig 4c, on the right). The figure shows a strong inverse relationship: higher suspicion is associated with lower quality ratings, though there is no clear difference between quality evaluations for men versus women. We used a linear mixed model to statistically analyze the overall quality data of all writing styles (represented in Fig 4a). 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 also as fixed effects. Lastly, we add participant as a random effect. The model confirms the visual trends in the figure. AI suspicion had a significant main effect on quality ($p < 0.001$); however, gender did not have a significant main effect ($p = 0.522$, *n.s.*). While writing style was not significant ($p = 0.060$, *n.s.*), the trend indicates that a larger sample could expose quality differences where the AI-inducing writing style is lower in quality, controlling for suspicion. There were no significant interactions in the model.

Similar to the analysis for overall quality, we also consider the effect of AI suspicion on the content and structure dependent variables (Section 3.3), and the results are largely the same. We use a linear mixed model for each measure, using the same independent variables as the overall quality model reported above. For the *content* dependent variable, the AI suspicion negatively impacted content evaluation ($p < 0.001$), and both genders were evaluated similarly ($p = 0.264$, *n.s.*). Lastly, the AI-inducing writing style was seen as lower quality content compared to the control ($p = 0.005$). The results for the *structure* variable are similar.

Lastly, we show a strong inverse relationship between AI suspicion and participants' hiring recommendations in Figure 5. 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. While the hiring likelihood trend looks somewhat different between the genders, our model did not show significant differences. We used a linear mixed model with AI suspicion, gender, writing style, and their three-way interaction as a fixed effects, predicting the hiring likelihood

score[2]. The model confirmed the relationship, with AI suspicion having a significant effect on hiring ($p < 0.001$). The writing style effect again was trending towards significance, suggesting that there may be differences between the AI-inducing writing style and the control ($p = 0.071$, *n.s.*). Gender ($p = 0.411$, *n.s.*) did not have a significant main effect, and there were no significant interaction effects.

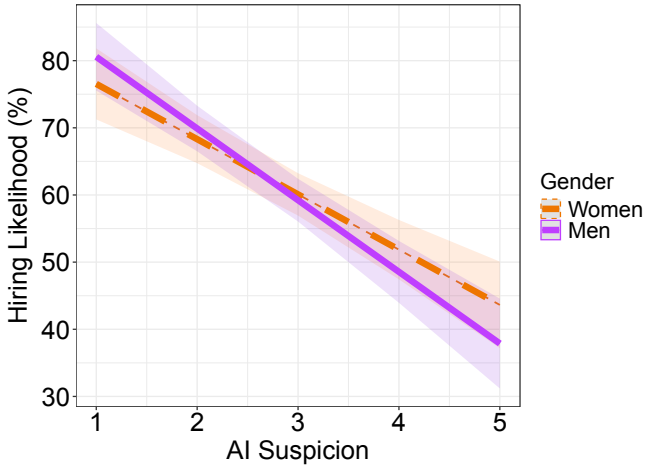


Fig. 5. **The Effect of AI Suspicion on Hiring.** Participants are less likely to hire a freelancer whom they suspect of using AI.

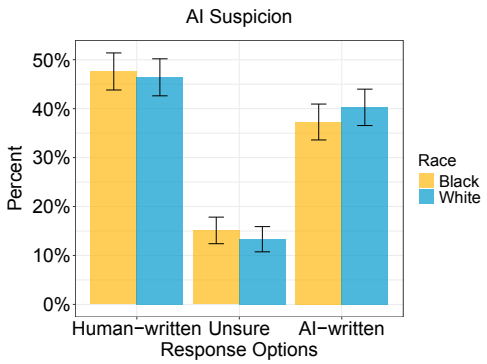
4.2 Study 2: Race

Study 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 in their likelihood to hire.

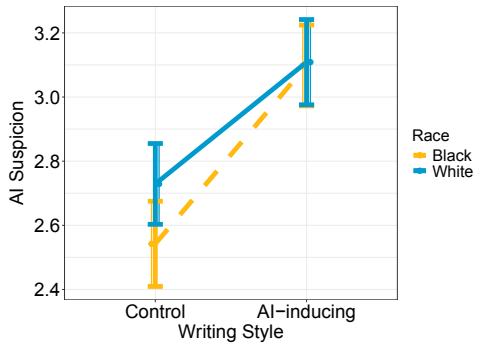
AI Suspicion. We did not find racial differences in AI suspicion. Figure 6a shows that White freelancers were only suspected of using AI 40.2% of the time (rightmost, blue column) Black freelancers were suspected 37.2% of the time (rightmost, gold). Analyzing this data statistically as we did in Study 1, with race as a fixed effect and participant as a random effect, confirmed race did not have a significant effect on AI suspicion ($p = 0.405$, *n.s.*). We also examined the relationship between writing style and AI suspicion between racial groups, as seen in Figure 6b. The figure reflects that there may be slightly less suspicion towards Black freelancers (dashed, gold) in the control condition, but the most dramatic trend is that 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 participant (random effect) confirmed that only writing style had a significant main effect on AI suspicion ($p < 0.001$).

Content Evaluations and Hiring. Next we move on to evaluations of content and hiring judgments, as we did in Study 1. Figure 7 illustrates a strong inverse relationship between AI suspicion and overall quality (quality evaluations drop sharply at higher levels of AI suspicion), without

[2]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) Though White freelancers appear to be more suspected of using AI, the difference is not significant



(b) White freelancers and Black freelancers receive the same amount of suspicion when the writing is AI-inducing.

Fig. 6. **Participant's AI Suspicion**. Figure 6a depicts the distribution of the responses. Figure 6b depicts the interaction between writing style and race.

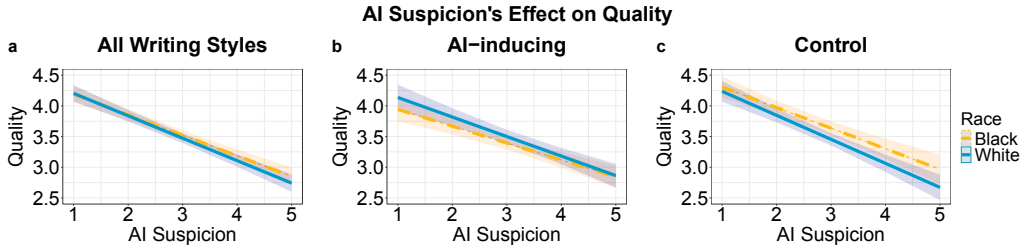


Fig. 7. **The Effect of AI Suspicion on Quality Evaluations.** AI Suspicion negatively impacts writing quality.

notable differences by race. We confirm this result with a linear mixed model that predicts overall quality for all writing styles (Fig 7a) from AI suspicion, race, writing style, and all interaction effects with participant as a random effect. We observe that AI suspicion had a significant main effect ($p < 0.001$), and there were no significant racial differences ($p = 0.215$, *n.s.*). Writing style had a significant effect ($p = 0.018$), with the control receiving higher overall quality evaluations compared to the AI-inducing style when controlling for AI suspicion.

We did not see racial differences in overall quality evaluations, nor did we see evidence of racial differences in the other quality measures, content and structure. Similar to Study 1, we analyze our measures with a linear mixed model that includes a three-way interaction between race, writing style, and AI suspicion. As expected, in both content and structure evaluations, AI suspicion had a significant main effect ($p < 0.001$). Writing style did not have a significant main effect on the content measure ($p = 0.535$, *n.s.*). However, writing style had a significant main effect on structure evaluations ($p = 0.015$) with the control evaluated as better in terms of structure compared to the AI-inducing style (controlling for suspicion).

Next, looking at potential impacts on participants' willingness to hire these freelancers, Figure 8 shows that participants are less likely to want to hire a freelancer they suspect of using AI. This trend consistent across both racial categories. We confirmed the trend with a linear mixed model

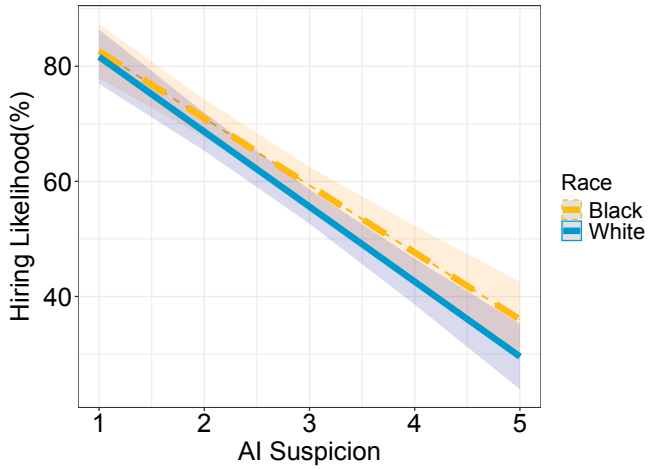
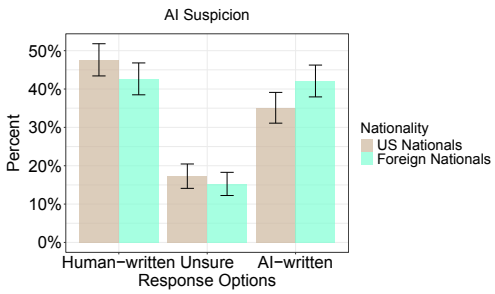


Fig. 8. **The Effect of AI Suspicion on Hiring.** Participants are less likely to hire a freelancer they suspect of using AI.

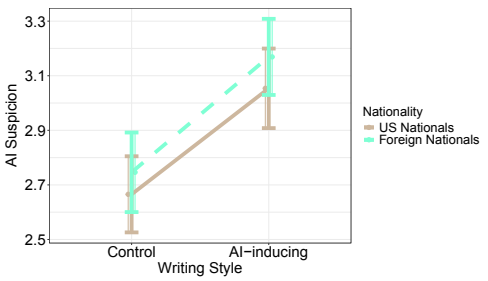
like in Study 1. We found that AI suspicion ($p < 0.001$) and writing style ($p = 0.024$) had a significant main effect on hiring. There were no significant racial differences ($p = 0.330$, *n.s.*), nor were there significant interaction effects.

4.3 Study 3: Nationality

Our third and final study 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 White Americans. We find that East Asian freelancers are suspected of using AI more than White Americans. However, we did not observe differences in quality evaluations or job outcomes.



(a) Participants are more likely to suspect foreign nationals of using AI



(b) Participants are more likely to suspect the AI-inducing writing style

Fig. 9. **Participant's AI Suspicion**. Figure 9a depicts the distribution of the responses. Figure 9b depicts the interaction between writing style and nationality.

AI Suspicion. East Asian freelancer profiles were somewhat more likely to be suspected of using AI, as shown in Figure 9a. The figure shows, for example, that 42.0% (the rightmost turquoise were suspected of using AI, compared to 35.1% for U.S. nationals (rightmost tan column). We confirmed this difference as significant ($p < 0.001$) with a linear mixed model (nationality as a fixed effect and participant as a random effect). However, we do not see differences in AI suspicion due nationality when accounting for the writing style treatment. Figure 9b shows that writing in an AI-inducing writing style increases AI suspicion, with both U.S. and foreign nationals' profiles being affected similarly. A linear mixed model (fixed effects: nationality, writing style, and their interaction; random effect: participant) confirmed that writing style has a significant main effect ($p < 0.001$) on AI suspicion. Neither nationality ($p = 0.216$, *n.s.*) nor the interaction between AI suspicion and writing style ($p = 0.807$, *n.s.*) had an effect.

Content Evaluations and Hiring. Like in Study 1 and Study 2, we see an inverse relationship

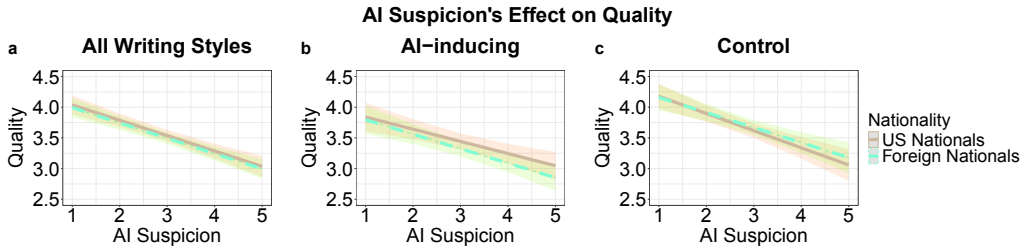


Fig. 10. **The Effect of AI Suspicion on Quality Evaluations.** AI suspicion negatively impacts quality.

between quality and AI suspicion; Figure 10 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 10a) that included a three-way interaction between nationality, AI suspicion, and writing style with participant as a random effect confirmed the statistically significant impact of AI suspicion ($p < 0.001$) and writing style ($p = 0.031$) on quality. There were no significant differences between nationalities ($p = 0.943$, *n.s.*), and there were no significant interaction effects.

Considering the effect of AI suspicion on our other quality evaluation measures, we find similar results to studies 1 and 2: AI-suspicion negatively impacts the content and structure evaluations. This difference is statistically significant for content ($p < 0.001$) and structure ($p = 0.004$) 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.

Turning a final time to participants' reported likelihood to hire, results are again similar to the two previous studies. As AI suspicion increases, hiring likelihood decreases, as seen in Figure 11. 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 and their three-way interaction as fixed effects; participant as random effect). The model shows that AI suspicion had a statistically significant effect on hiring ratings ($p < 0.001$). The effect of writing style was close to significant ($p = 0.054$, *n.s.*). There were no effects of nationality ($p = 0.982$, *n.s.*), and there were no significant interaction effects.

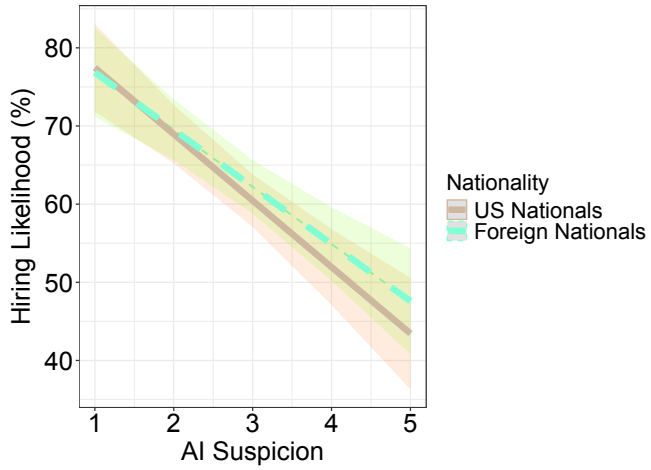


Fig. 11. **The Effect of AI Suspicion on Hiring.** Participants are less likely to hire a freelancer they suspect of using AI.

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 harms, which has focused on harms as a result of AI's development and deployment [83, 91]. Perceptual harm, 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 negative judgment against the perceived AI user. In this study, we examine perceptual harms in the context of textual content; however, perceptual harms could also occur in different mediums such as images, videos, and audio. Similar to writing tasks, creators may use text-to-image or text-to-audio [32] models to enhance their creative productivity or produce novel content [96]. 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 [38], 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 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 studies 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 Study 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 [44, 52]; it is possible (and should be confirmed in future work) that men were more suspected of AI use due to real gendered differences in technology uptake.

This 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. This possible impact of technology-related stereotypes may also be consistent with the results of Study 3. This last study 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 [12, 67, 89]. In fact, early work has shown that Asians adopted generative AI more than White folks [9]. However, in addition to technology-related stereotypes, it is possible 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.

While AI suspicion was different between most of the groups in our studies (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 studies 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 could be due to the negative associations with AI-writing [46, 53]. Furthermore, it is possible that people who use AI may be seen as having undesirable traits in an employee (e.g., "lazy" or "incapable"). These results suggest that stereotypes about technological savviness may cause perceptual harms, in the form of AI suspicion, against groups seen as likely to uptake these new technologies, including groups like White men who usually stand to benefit from social biases.

While the current results suggest that stereotypes may cause perceptual harms, the evolving societal norms surrounding AI and the 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 [21]. 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 [21]. While public sentiment toward AI was and still is largely positive [21], there are growing concerns about issues such as loss of control, ethical challenges, and the impact of AI on employment [21, 51, 73]. 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 [77]. While fewer Americans now perceive AI as harmful compared to a year ago [77], many remain cautious about its applications, especially in workplace settings [77]. Notably, people who are more knowledgeable about AI are less likely to express concerns about its effects [77]. 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. The outcomes of perceptual harms can differ based on the context. In artist communities, AI suspicion could lead to worse content evaluation and loss of income and opportunities [48]. 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 [37, 45, 63]. The magnitude of the outcomes of perceptual harms also varies by context. 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 [22, 65]. In contrast, in interpersonal work communication, AI suspicion might result in a more localized loss of trust. In the future, as AI use becomes normalized, 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.

As AI norms evolve and the outcomes of perceptual harms become more prominent, we are already seeing different forms of actions taken that may further shift social dynamics between individuals. Individuals respond to others' perceived use of AI-generated content in various ways. Increasingly, people will also react, likely preemptively, to being perceived themselves as using AI. With the prevalence of AI-generated content, content evaluators may turn to a new set of tools to validate the legitimacy of work; whereas content creators may find other ways to signal ownership and authenticity. These kinds of responses will likely be also subject to direct or indirect biases. For instance, in education, where concerns about generative AI and academic integrity are prevalent [75, 86, 93], teachers may adopt varying approaches to addressing AI-related suspicion. Some teachers may be critical of all writing assignments and apply AI detection software to every assignment, while others may take a more targeted approach, using the software only for assignments they suspect are AI-generated [60]. 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 mitigate perceptual harms. We note that perceptual harms are fundamentally a social problem caused by stereotypes and societal views of AI; therefore, previous solutions to AI's sociotechnical harms [84, 90, 91] 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, we 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 [4, 38, 45, 61] 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 [55] 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 [57].

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 [7, 64, 77]. Many people are still unsure of what AI can or cannot do [21, 51, 64]. Our data, while intended to create ambiguity, was consistent with previous work showing people's inconsistent heuristics for detecting AI text [46]. 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 another person's rationale for why the content was human-written. For example, like previous work [46], 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 [77]. 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 [64, 76].

Our work has some limitations that should be considered. First, while we recruited a US-representative sample for our experiment, it still relied on crowdworkers 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]. Since our experiment highlighted AI use, it may have primed the participants to think about *and* to consider it as negative. We attempted to phrase the question as neutrally as possible when describing the task to participants as a "profile evaluation task" and stating "some of the profiles may or may not be AI generated" to avoid priming participants. In addition, our work's generalizeability 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 a 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., [5]) can help provide more robust data on the impact of perceptual harms of AI, but creating the experimental manipulation would be difficult for such an experiment.

6 CONCLUSION

This paper 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 social groups. We propose that perceptual harms occur along three axes: suspected AI use, which can lead differences in content evaluation, and outcomes. Through a series of online experiments, we show, somewhat surprisingly, that dominant social groups are often more likely to be suspected of using AI. Consistent with prior work, we see that perceptions of AI use negatively impacted content evaluations and hiring outcomes, but these metrics were not different between groups after controlling for AI suspicion. We encourage the research community to further explore perceptual harms and how they may change technology's perception changes. As AI technologies become mainstream and all types of people are seen as equally likely to use it, who might perceptual harms impact? Conversely, if there is broad public uptake and AI technologies are seen as virtuous, will there be harms, or will the norms of its use and evaluation change? These are important questions for future work.

7 ACKNOWLEDGMENTS

This research was supported by a gift to the LinkedIn-Cornell Bowers CIS Strategic Partnership. The material is also based upon work supported by the National Science Foundation under Grant No. CHS 1901151/1901329.

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. https://doi.org/10.1145/3461702.3462624

[2] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2024. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. arXiv:2409.11360 [cs.HC] https://arxiv.org/abs/2409.11360

[3] David Baidoo-Anu and Leticia Owusu Ansah. 2023. 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.

[4] Lucas Bellaiche, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli. 2023. Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications* 8, 1 (2023), 42.

[5] Marianne Bertrand and Sendhil Mullainathan. 2004. 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.

[6] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. 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* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 436–452. https://doi.org/10.1145/3581641.3584060

[7] Maalvika Bhat and Duri Long. 2024. Designing Interactive Explainable AI Tools for Algorithmic Literacy and Transparency. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 939–957. https://doi.org/10.1145/3643834.3660722

[8] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[9] Jenna Butler, Sonia Jaffe, Nancy Baym, Mary Czerwinski, Shamsi Iqbal, Kate Nowak, Sean Rintel, Abigail Sellen, Mihaela Vorvoreanu, Najeeb G. Abdulhamid, and et al. 2024. Microsoft New Future of Work Report 2023. https://www.microsoft.com/en-us/research/publication/microsoft-new-future-of-work-report-2023/

[10] Paola Catenaccio. 2008. 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.

[11] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 43.

[12] Grace A Chen and Jason Y Buell. 2018. Of models and myths: Asian (Americans) in STEM and the neoliberal racial project. *Race Ethnicity and Education* 21, 5 (2018), 607–625.

[13] Hyeshin Chu, Joohee Kim, Seongouk Kim, Hongkyu Lim, Hyunwook Lee, Seungmin Jin, Jongeun Lee, Taehwan Kim, and Sungahn Ko. 2022. An empirical study on how people perceive AI-generated music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 304–314.

[14] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[15] Kate Crawford. 2016. Artificial Intelligence's White Guy Problem. *New York Times* (June 2016).

[16] Wenzhe Cui, Suwen Zhu, Mingrui Ray Zhang, H. Andrew Schwartz, Jacob O. Wobbrock, and Xiaojun Bi. 2020. JustCorrect: Intelligent Post Hoc Text Correction Techniques on Smartphones. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 487–499. https://doi.org/10.1145/3379337.3415857

[17] Edmund Dervakos, Giorgos Filandrianos, and Giorgos Stamou. 2021. Heuristics for evaluation of AI generated music. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 9164–9171.

[18] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. 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)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7250–7274. https://doi.org/10.18653/v1/2022.acl-long.501

[19] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. 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, Article 25 (feb 2024), 40 pages. https://doi.org/10.1145/3637875

[20] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[21] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

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

[23] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How do users evaluate the credibility of Web sites? a study with over 2,500 participants. In *Proceedings of the 2003 Conference on Designing for User Experiences* (San Francisco, California) *(DUX '03)*. Association for Computing Machinery, New

York, NY, USA, 1–15. https://doi.org/10.1145/997078.997097
[24] Emma Frid, Celso Gomes, and Zeyu Jin. 2020. Music creation by example. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[25] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 977, 17 pages. https://doi.org/10.1145/3613904.3641955
[26] S Michael Gaddis. 2017. How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science* 4 (2017), 469–489.
[27] Sanjana Gautam, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. From melting pots to misrepresentations: Exploring harms in generative ai. *arXiv preprint arXiv:2403.10776* (2024).
[28] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533
[29] Sourojit Ghosh and Aylin Caliskan. 2023. 'Person'== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *arXiv preprint arXiv:2310.19981* (2023).
[30] Auro Del Giglio and Mateus Uerlei Pereira da Costa. 2023. The use of artificial intelligence to improve the scientific writing of non-native english speakers. *Revista da Associação Médica Brasileira* 69, 9 (2023), e20230560.
[31] United States Government. [n. d.]. https://www.usa.gov/official-language-of-us
[32] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655* (2023).
[33] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.
[34] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1914–1933. https://doi.org/10.1145/2998181.2998327
[35] Matthias Haslberger, Jane Gingrich, and Jasmine Bhatia. 2023. No great equalizer: experimental evidence on AI in the UK labor market. *Available at SSRN* (2023).
[36] Bernie Hogan and Brent Berry. 2011. Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & community* 10, 4 (2011), 351–372.
[37] Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports* 13, 1 (2023), 5487.
[38] Joo-Wha Hong. 2018. 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* (Las Vegas, NV, USA). Springer-Verlag, Berlin, Heidelberg, 290–303. https://doi.org/10.1007/978-3-319-91244-8_24
[39] White House. 2016. Preparing for the future of artificial intelligence. Executive Office of the President National Science and Technology Council. Committee on Technology.
[40] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* (Nov. 2024). https://doi.org/10.1145/3703155 Just Accepted.
[41] Xiang Hui, Oren Reshef, and Luofeng Zhou. 2023. The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. *Available at SSRN 4527336* (2023).
[42] Sung Il Hwang, Joon Seo Lim, Ro Woon Lee, Yusuke Matsui, Toshihiro Iguchi, Takao Hiraki, and Hyungwoo Ahn. 2023. Is ChatGPT a "fire of prometheus" for non-native English-speaking researchers in academic writing? *Korean Journal of Radiology* 24, 10 (2023), 952.
[43] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650* (2019).
[44] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How different groups prioritize ethical values for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 310–323.
[45] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. 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*. 1–13.

[46] Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120.

[47] Heinrich Jiang and Ofir Nachum. 2020. Identifying and correcting label bias in machine learning. In *International conference on artificial intelligence and statistics*. PMLR, 702–712.

[48] Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 363–374. https://doi.org/10.1145/3600211.3604681

[49] David Johnson and Lewis VanBrackle. 2012. Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing* 17, 1 (2012), 35–54.

[50] Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danaë Metaxa, and Mor Naaman. 2024. The Role of Inclusion, Control, and Ownership in Workplace AI-Mediated Communication. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1016, 10 pages. https://doi.org/10.1145/3613904.3642650

[51] Patrick Gage Kelley, Yongwei Yang, Courtney Heldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T Newman, and Allison Woodruff. 2021. 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*. 627–637.

[52] Taenyun Kim, Maria D Molina, Minjin Rheu, Emily S Zhan, and Wei Peng. 2023. 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*. 1–20.

[53] Nils Köbis and Luca D Mossink. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in human behavior* 114 (2021), 106553.

[54] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. https://doi.org/10.1145/3582269.3615599

[55] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 23 pages. https://doi.org/10.1145/3654777.3676419

[56] Liana Christin Landivar. 2013. Disparities in STEM employment by sex, race, and Hispanic origin. *Education Review* 29, 6 (2013), 911–922.

[57] Karen Levy. 2022. Data driven: truckers, technology, and the new workplace surveillance. (2022).

[58] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[59] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (2022), 669–677.

[60] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. arXiv:2304.02819 [cs.CL] https://arxiv.org/abs/2304.02819

[61] Sue Lim and Ralf Schmälzle. 2024. The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100058. https://doi.org/10.1016/j.chbah.2024.100058

[62] Jin Liu, Xingchen Xu, Yongjun Li, and Yong Tan. 2023. " Generate" the Future of Work through AI: Empirical Evidence from Online Labor Markets. *arXiv preprint arXiv:2308.05201* (2023).

[63] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. 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 Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 474, 13 pages. https://doi.org/10.1145/3491102.3517731

[64] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376727

[65] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from Generative Artificial Intelligence Is Believed Less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/3531146.3533077

[66] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. https:

//doi.org/10.18653/v1/2021.nuse-1.5
[67] Ebony O McGee, Bhoomi K Thakore, and Sandra S LaBlance. 2017. The burden of being "model": Racialized experiences of Asian STEM college students. *Journal of Diversity in Higher Education* 10, 3 (2017), 253.
[68] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.
[69] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "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). https://www.frontiersin.org/articles/10.3389/frai.2021.725911/full?&utm_source=Email_to_authors_&utm_medium=Email&utm_content=T1_11.5e1_author&utm_campaign=Email_publication&field=&journalName=Frontiers_in_Artificial_Intelligence&id=725911
[70] Danaë Metaxa-Kakavouli, Kelly Wang, James A Landay, and Jeff Hancock. 2018. Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 1–6.
[71] Hannah Mieczkowski, Jeffrey T Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–14.
[72] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416
[73] Hugo Neri and Fabio Cozman. 2020. The role of experts in the public perception of risk of artificial intelligence. *AI & society* 35, 3 (2020), 663–673.
[74] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. https://doi.org/10.1126/science.adh2586 arXiv:https://www.science.org/doi/pdf/10.1126/science.adh2586
[75] Mike Perkins. 2023. 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).
[76] Irene Rae. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 978, 14 pages. https://doi.org/10.1145/3613904.3642076
[77] Julie Ray. 2024. Americans express real concerns about Artificial Intelligence. https://news.gallup.com/poll/648953/americans-express-real-concerns-artificial-intelligence.aspx
[78] Dennie Rothschild and Felicia Klingenberg. 1990. Self and peer evaluation of writing in the interactive ESL classroom: An exploratory study. *TESL Canada Journal* (1990), 52–65.
[79] Carol Sager. 1973. Sager Writing Scale. (1973).
[80] Sashaborm. 2021. Thispersondoesnotexist - random AI generated photos of fake persons. https://this-person-does-not-exist.com/en
[81] Edward Segal. 2023. New report provides reality check about freelancers in the workforce. https://www.forbes.com/sites/edwardsegal/2023/12/12/new-report-provides-reality-check-about-freelancers-in-the-workforce/
[82] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
[83] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. https://doi.org/10.1145/3600211.3604673
[84] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. arXiv:2406.09264 [cs.HC] https://arxiv.org/abs/2406.09264
[85] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 68 (sep 2023), 57 pages. https://doi.org/10.1145/3511599

[86] Miriam Sullivan, Andrew Kelly, and Paul McLaughlan. 2023. ChatGPT in higher education: Considerations for academic integrity and student learning. (2023).
[87] Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. 2024. Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication* 29, 1 (02 2024), zmad045. https://doi.org/10.1093/jcmc/zmad045 arXiv:https://academic.oup.com/jcmc/article-pdf/29/1/zmad045/56546560/zmad045.pdf
[88] The Upwork Team. 2024. https://www.upwork.com/resources/highest-paying-freelance-jobs
[89] Zer Vue, Chia Vang, Neng Vue, Vijayvardhan Kamalumpundi, Taylor Barongan, Bryanna Shao, Sunny Huang, Larry Vang, Mein Vue, Nancy Vang, Jianqiang Shao, CoohleenAnn Coombes, Prasanna Katti, Kaihua Liu, Kailee Yoshimura, Michelle Biete, Dao-Fu Dai, Mark A. Phillips, and Richard R. Behringer. 2023. Asian Americans in STEM are not a monolith. *Cell* 186, 15 (2023), 3138–3142. https://doi.org/10.1016/j.cell.2023.06.017
[90] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).
[91] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088
[92] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 109, 14 pages. https://doi.org/10.1145/3544548.3581357
[93] Ying Xie, Shaoen Wu, and Sumit Chakravarty. 2023. 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* (Marietta, GA, USA) *(SIGITE '23)*. Association for Computing Machinery, New York, NY, USA, 79–83. https://doi.org/10.1145/3585059.3611449
[94] Ma Xueliang and Dai Qingxia. 1988. Language and nationality. *Chinese Sociology & Anthropology* 21, 1 (1988), 81–104.
[95] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105
[96] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (03 2024), pgae052. https://doi.org/10.1093/pnasnexus/pgae052 arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/3/pgae052/57464715/pgae052.pdf

A APPENDIX

A.1 Treatment Validation

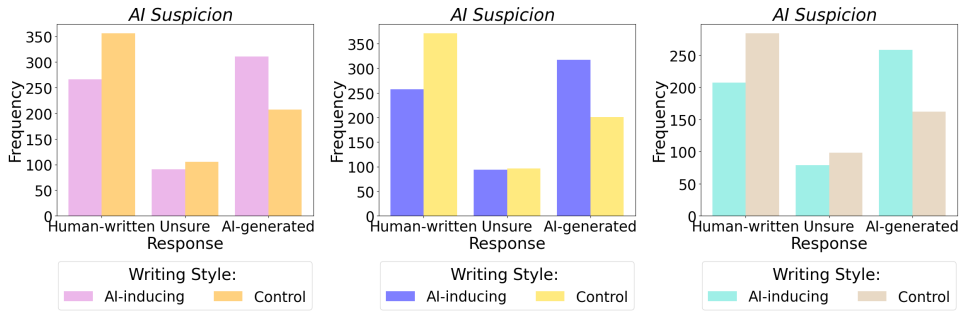
We manually manipulated half of the press releases in an AI-inducing style to raise participants' suspicion of AI use. Table 3 presents an example of AI-inducing sentences, which were created by replacing words with less common synonyms and employing verbose language. We validated the effect of writing style post hoc, as shown in Figure 12. 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 fixed effect and participant as a random effect confirmed the visual finding. The result is statistically significant across all three studies ($p < 0.001$).

A.2 Participant Demographics

We recruited a US representative sample from Prolific and present the demographic data of the participants in all three studies in Table 4. We also conduct an exploratory analysis to determine if age had a significant effect on AI suspicion and hiring. Since our largest age group was participants aged 55-64 years old, for each study in our experiment, we removed this group and re-ran our AI suspicion analysis and hiring analysis. The exploratory analysis shows the same trends for all three studies.

Table 3. **AI-inducing sentences**. We manipulated half of the press releases to be AI-inducing by modifying a three sentences in the sample. We provide a sample of AI-Inducing sentences below. Noticeable differences in the sentences are highlighted in red.

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) Study 1–Gender (b) Study 2–Race (c) Study 3–Nationality

Fig. 12. **Participants Assessment of Writing Styles.** The AI-inducing writing style was more suspected of being AI-generated compared to the control.

In Study 1, we conducted a linear mixed model analysis to predict AI suspicion based on gender and a separate model incorporating both gender and writing style as predictors. Our findings indicate a trend where men are more likely to be suspected of using generative AI ($p = 0.116$, *n.s.*); however, this trend is not statistically significant, likely due to reduced statistical power. Similarly, in the model including both gender and writing style, the same pattern emerges: men are more often suspected ($p < 0.001$), with suspicion being notably higher when the writing style is perceived as AI-generated ($p < 0.001$). Lastly, we conducted a linear mixed model to predict hiring likelihood based on gender, AI suspicion, and writing style. AI suspicion negatively impacted hiring likelihood ($p < 0.001$).

We performed the same analysis in Study 2 with race as our demographic variable. Like the findings in our main results section, while White freelancers were suspected of using AI more than Black freelancers, we do not see any significant racial differences ($p = 0.155$, *n.s.*). When we

include race and writing style in the AI suspicion model, we still do not see any significant racial differences ($p = 0.594$, *n.s.*). We do, however, see the effect of writing style ($p < 0.001$), where the AI-inducing style is much more likely to be suspected of being AI-generated. As expected, in our hiring prediction model, we see that AI suspicion negatively impacts hiring likelihood ($p < 0.001$).

In Study 3, we see that Foreign Nationals are more likely to be suspected of using generative AI ($p = 0.017$). When we include writing style in the AI suspicion model, we still observe the same trend ($p = 0.011$). Furthermore, we see that the AI-inducing style is more likely to be suspected of being AI-generated ($p = 0.008$). Similar to previous studies, we see that AI suspicion reduces hiring likelihood ($p < 0.001$).

Table 4. **Participant Demographics**. An overview of the participants' demographic backgrounds in all three studies.

	Study 1	Study 2	Study 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

Received ; revised ; accepted