It's Trying Too Hard to Look Real: Deepfake Moderation Mistakes and Identity-Based Bias

JARON MINK, MIRANDA WEI, COLLINS W. MUNYENDO, KURT HUGENBERG, TADAYOSHI KOHNO, ELISSA M. REDMILES, GANG WANG

Online platforms employ manual human moderation to distinguish human-created social media profiles from deepfake-generated ones. Biased misclassification of real profiles as artificial can harm general users as well as specific identity groups; however, no work has yet systematically investigated such mistakes and biases. We conducted a user study (n=695) that investigates how 1) the identity of the profile, 2) whether the moderator shares that identity, and 3) components of a profile shown affect the perceived artificiality of the profile. We find statistically significant biases in people's moderation of LinkedIn profiles based on all three factors. Further, upon examining how moderators make decisions, we find they rely on mental models of AI and attackers, as well as typicality expectations (how they think the world works). The latter includes reliance on race/gender stereotypes. Based on our findings, we synthesize recommendations for the design of moderation interfaces, moderation teams, and security training.

Content Warning: This paper discusses stereotypes based on gender, race, and age that are offensive and upsetting.

1 INTRODUCTION

The ability to create deceptive personas on social media has become a pressing societal concern as such artificial personas are increasingly used in disinformation and social engineering campaigns [6, 19, 100]. Breakthroughs in artificial intelligence (AI), particularly deep learning, now allow for photorealistic images to be created with a single text prompt [90, 96, 108], and human-indistinguishable text to be automatically generated [28]. In response, social media platforms have largely banned artificially-generated content (or "deepfakes") [69, 75, 112, 114] and enforce this by attempting to distinguish artificial and real content through *content moderation.*[1]

Content moderation typically consists of two detection techniques: *automatic* and *manual* detection [49, 97]. *Automatic* detection consists of classifiers that analyze account information (e.g., geolocation or sentiment of posts [122]) to provide scalable, inexpensive moderation. *Manual* detection consists of human moderators that evaluate profiles and content based on a platform's policies. This allows for a more holistic, contextualized decision than automatic detection, but at a greater cost [49]. Given these tradeoffs, many companies employ both techniques, often using manual defenses to provide training data and verification of automatic defenses [49]. Once detected, actions such as content removal and deactivation of accounts, among others, are employed to mitigate the perceived content violation.

Unfortunately, content moderation can lead to misclassifications, such as classifying a *real profile as artificial.* These errors can result in multiple harms. Economic harms may directly result for those who use social media platforms to promote or expand their professional services [5, 11, 12, 80, 106]. Indirect economic harms may be incurred for general users from the loss of social capital (or the economic benefits that result from social relations [3]) achieved through social media [34, 80]. Emotional harms may also result from users being separated from social app-provided communities [45, 80], resulting in feelings of invisibility and oppression [58]. Lastly, the platform itself may be harmed as these incorrect decisions can lead to reduced trust and de-valuation of the platform as a whole [80].

To avoid these harms, automatic defenses are expected to make as few incorrect classifications of real users as artificial as possible [39]. Furthermore, automatic defenses are also now being evaluated for algorithmic bias across

[1]While our study scope is limited to artificial content, content moderation is related to the enforcement of all terms of service beyond artificial content (e.g., hate and harassment).

sociodemographic factors such as race [107] and gender [10], to investigate if there is disproportionate harm to any specific community. However, manual content moderation has not received the same level of scrutiny. Our work aims to fill this gap by experimentally evaluating the efficacy and bias in human content moderation decisions. In particular, this work investigates whether *real profiles* are disproportionately misclassified as artificial across gender or racial identities.

Prior work finds that human moderators rely on an array of heuristics to determine whether a profile is real or artificial. For instance, text-based heuristics may include grammar errors or the perceived intentions behind the text [28, 76] while image-based heuristics may include clothing, facial, or body features that appear malformed [76]. Unsurprisingly, these heuristics also often lead moderators to incorrectly conclude that real content is artificial. Mink et. al [76] found anecdotal evidence that heuristics may incorporate patterns found along racial or gender communities, while Nightingale and Farid found that when presented with a dataset of AI-generated faces, people trusted white-presenting faces the most [87]. Prior work on the impacts of such potential biases finds race-related differences in online content moderation experiences [55] and that people associate particular gender and racial attributes with AI systems [14, 25, 92, 109].

Building on these findings, we empirically evaluate the impact of gender and race— which prior work identifies as prominent sources of bias [21, 56, 70, 87]— on errors in human moderation. Specifically, we aim to answer the following research questions:

RQ1 How do specific factors — (a) the identity of the profile, (b) whether the moderator holds the same identity as the profile, and (c) which components of the profile are shown — influence moderation error rates among people?
RQ2 How do people reason about profile moderation decisions?

To answer these questions we conducted a survey experiment (n=695) in which moderators engaged in moderation tasks on real human-made profiles and explained their decisions. Drawing on descriptive, statistical, and qualitative analysis of our data, we synthesize two key findings.

First, we find statistical evidence that all three factors examined in RQ1 influence moderation of real profiles. In particular, we find that when shown only the image and name of a profile, both the identity of the profile and whether the moderator shares that identity (in-group vs. out-group) influence the moderation decision. When either the full profile (including the text content) or only the text content of the profile is shown, biases in moderation decrease.

Second, we find that participants' decisions about profile artificiality depend on three primary perspectives: their worldview of real profiles (and human behaviors), AI functionality, and attacker strategies on online platforms. Importantly, many participants relied on identity-based stereotypes to reason about artificiality, likely explaining the impact of identity-related information on moderation decisions we observe in our statistical analysis.

Taking these findings together, we synthesize a set of recommendations to minimize bias in content moderation, including suggestions for improved design of platform content moderation interfaces and security training, as well as implications of our results on the hiring of moderator teams.

2 BACKGROUND AND RELATED WORK

Background on artificially-generated content. Artificially-generated content refers to text, images, videos, or other designs created by computer programs that could be perceived as being created by humans [77]. Deepfakes are one example of artificially-generated content. Advances in AI [90, 96, 108], and particularly deep learning, have made it easy to generate high-quality deepfake videos and images [59, 67], and even social personas/profiles. In a recent user study (n=286), Mink et al. [76] found that many participants trusted and ultimately chose to connect with deepfake profiles. Most closely related to our work is a study by Nightingale and Farid [87] that performed a series of user studies exploring the effect of race and gender of synthetic faces on their perceived artificiality. The study finds that synthetically-generated white faces, and particularly male white faces, were the least accurately classified by participants. Our work builds on these studies by exploring participants' mental models for perceiving the artificiality of *real* profiles, including their biases and stereotypes in doing so. While Nightengale and Farid's [87] study only focuses on perceived artificiality of synthetic faces, we take a complementary approach to focus on perceived artificiality of real *profiles* and additionally explore the impact of moderator identity and various profile components (name and text, in addition to face) on biases in moderation. We additionally qualitatively examine the factors underlying moderators' artificiality perceptions.

User perceptions of artificial content. To further understand how humans perceive and detect deepfakes, Tahir et al. [110] conducted a user study (n = 95) with deepfake videos generated from three different algorithms, finding that participants' detection accuracy of deepfakes was less than 26%. In another study focused on low-resourced users, Shahid et al. [104] found that most participants from their study in India (n = 36) were unaware of deepfake videos, and only perceived videos to be fake if they contained inaccurate information or artifacts. Our study (with participants from the United States), in contrast, finds that such features (e.g., picture quality, grammar issues, typos) are perceived as indicators of both artificially-generated profiles and real profiles by different participants.

Moderation of malicious and artificial content. To detect malicious and artificially-generated content, particularly deepfakes, two broad techniques have been used: automated methods and manual methods. Automated methods primarily rely on machine learning [86, 119] to detect deepfake-related features, while manual or moderator-based approaches rely on humans to determine if the content is malicious and artificially-generated [53, 71, 113]. Automated methods require large data sets of malicious and benign content for training [86]. Due to the dependency on training data, machine learning-based detectors [48, 93, 103] often face the challenge to generalize to new data [86]. To make the detection more robust, platforms such as LinkedIn leverage moderator-based methods in addition to automated techniques [18] if their budgets allow [15]. Despite the joint efforts, researchers have shown that both manual and automated detection techniques of deepfakes are subject to misclassification and errors [65, 66].

Biases and harms from moderation. As prior work [104, 110] has noted, human-moderators are prone to make mistakes when evaluating whether content or videos are artificially-generated. This can harm users, especially if the moderation decisions result in users getting removed from platforms such as LinkedIn [15]. Moderation errors can result from racial stereotypes and bias: AI is often implicitly associated with white people [25, 92] and Black people are significantly more likely to report being incorrectly moderated on social media compared to white users [55]. Historically, nation-states have politically exploited Black identities to generate artificial profiles in order to sow disinformation [47]. In the context of natural language detection, annotator demographic identity and political beliefs influenced ratings of toxicity [99]—more racist annotators were more likely to rate African-American English as toxic. Previous work in robot-human interaction has also found gender-modulated evaluations of artificiality, with robots that have feminine features viewed as warmer and more human than those with masculine features [14, 109]. Haut et al. [56] found that changing the race of a person in a static image had negligible impact on the image's perceived credibility, but the impact is significant in videos. Videos of white people are seen as more likely to be telling the truth compared to videos of those perceived as Black. These results motivate us to investigate whether certain communities may be disproportionately harmed by manual profile moderation on social media platforms.

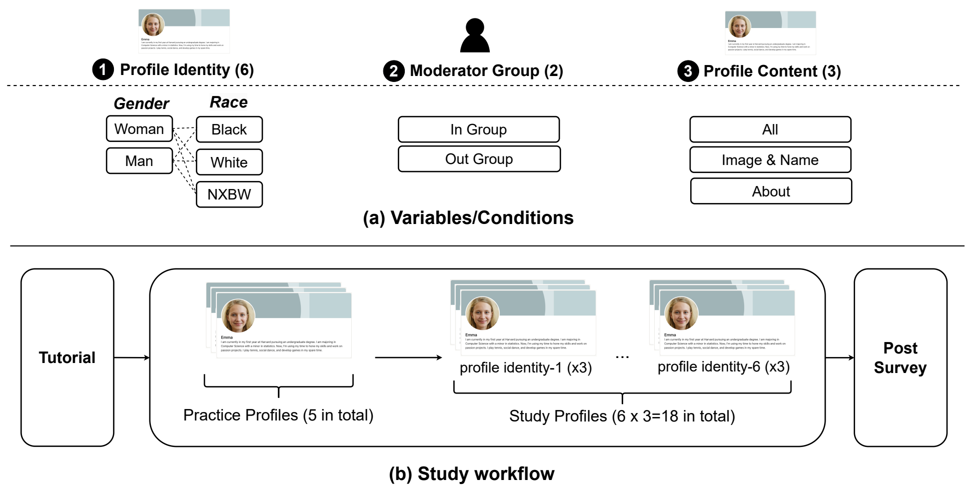

Fig. 1. **Study Overview** – We list the considered variables and conditions in the study in (a), and show the study workflow in (b). The “practice profiles” are randomly selected from the profile pool and the results are disregarded from analysis. The “study profiles” cover all 6 profile identities, presented in randomized order.

3 METHODOLOGY

To investigate whether gender and racial bias influence misclassification of real profiles (**RQ1**) and to understand why these choices are made (**RQ2**), we conducted an online survey experiment ($n = 695$). In this survey, participants acted as moderators to determine whether several LinkedIn profiles are real or artificial. Specifically, participants were asked to identify which of a subset of actual LinkedIn profiles (drawn from a population of 160 user-provided profiles varying in race and gender) were computer-generated (i.e., artificial), and their reasons for believing so. Given the sensitivity of the topic, our IRB-approved study made several design decisions to protect the privacy of our participants (see Section 3.5).

3.1 Experiment Design

Our experiment involved showing each participant LinkedIn profiles for them to evaluate as artificial or real. An overview of the study is shown in Fig. 1.

Use of Real Profiles. We use a noise-alone 2-alternative forced choice (2AFC) study design to investigate how non-signals (i.e., real aka non-artificial profiles) are misclassified [51, pg. 43-52]. Similar 2AFC designs have been used in social psychology to investigate racial biases in other domains [60, 70]. Thus, all profiles shown to survey participants were real[2] and collected from real LinkedIn users (see Section 3.2). By only using real profiles, all responses noting a profile as artificial were false positives. Using this 2AFC design allowed us to set up a “demand effect” [91] which prompted participants to mark some portion of real profiles as artificial. We intentionally embed a demand effect for two reasons. (1) We were most interested in whether profiles incorrectly identified as artificial are *disproportionately* from particular identity groups, i.e., whether these mistakes are biased across several experimentally controlled variables (see below). This setup amplified such effects to be better understood in an experimental setting, rather than to investigate bias prevalence (which we leave for future work). As we anticipated that the effects of bias would be subtle and nuanced,

[2]In our study, we define “real” profiles to be profiles that we, to the best of our ability, have verified as existing on LinkedIn, only changing content to preserve the PII of the individual and the name to the participants’ provided psuedonym (Section 3.2).

our exploratory study aims to identify which biases might influence content moderator decisions. (2) In practical content moderation, moderators do not know the actual incidence of artificial versus real content, so there may be scenarios when all content is real but moderators are nevertheless primed to look for artificial content. Our study emulates these environments.

Experimentally Controlled Variables. We determined which of the total 160 profiles we collected to show to each participant by balancing two hierarchical treatment effects[3] (within-subjects) and assigning subjects to one experimental condition (between-subjects). Specifically, each participant received a random selection of study profiles balanced across:

- ❶ "Profile Identity": the intersectional identity of the *shown profile*. Specifically, we evaluate 6 different intersectional profile identities: Black women, Black men, white women, white men, not exclusively Black or white (NXBW) women, and NXBW men. The choice of identities is directly informed via prior literature that found differences in perceived artificiality, warmness, and moderation between people who identify as Black and white, and people who identify as women and men (see Section 2). In addition to these four studied identities, we also include profiles of those who identify as not exclusively Black or white (NXBW women and NXBW men). This was done to prevent participants from realizing that they've only been shown Black and white faces, infer that this racial distinction was an important aspect of the study, and bias responses towards (inauthentically) equitable behavior (e.g., a social-desirability bias [79]). Such distractor stimuli are common in studies centering race and face perception (e.g., [31]); therefore we include NXBW women and NXBW men to mitigate such bias, but do not analyze the related responses.
- ❷ "Moderator Identity": a boolean saying whether the intersectional identity of the *participant* is the same as the profile. Each profile is an "in-group" or an "out-group" of the participant's identity. For example, when a profile whose user identifies as a Black man is evaluated by a moderator who also identifies as a Black man, that resulting moderator's identity is regarded as an "in-group;" conversely, if the moderator self-identities as anything *besides* a Black man, the moderator's identity would be regarded as an "out-group."

Additionally, participants were divided into three conditions, which determine which ❸ "Profile Content" is made visible to the participant. Based on prior work that showed that racial bias may result from identity-laden content such as online users' first name [41], we vary which content is made available to viewers to investigate whether any discovered bias is due to differences in profile content (e.g., the "about" section), or the identity-laden content (e.g., the profile image and name). Thus we have three profile content conditions for each profile: the image and name only, the text "about" section only, or all content (image, name, and text). We chose not to control for any ancillary information participants provide in this content (e.g., image quality, professional experience). This both ensures a higher degree of external validity and allows us to capture biases directly due to identity, as well as factors that correlate with identity; this aligns with modern definitions of identity-based discrimination [32, pg. 39-42].

Moderators may have access to other information to make their decision (e.g., posts, the profile's social connections); however, we focus on profile content because it is the most directly relevant to our research questions on profile identity. Furthermore, content moderators often face large workloads and must make decisions rapidly [7, 8, 97], and first impressions are frequently made based on faces [121].

Choice of LinkedIn as a Platform. While harms from content moderation are becoming an increasing concern across many platforms (e.g., Twitter [85], Instagram [15], Facebook [54, 73]), we situate our study on LinkedIn as

[3] Hierarchal meaning that treatments are non-exclusive of one another.

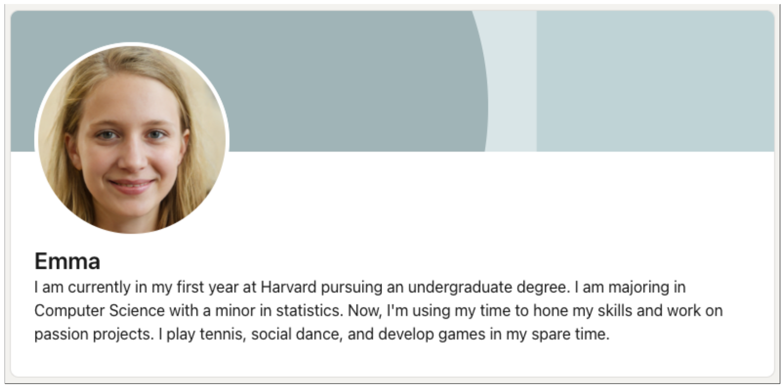

Fig. 2. **Profile Example** – During profile collection, participants submit their current public profile image, “about” text, and a chosen pseudonym. This information is then presented in the format of this example profile to participants in the content moderation survey. *Note: This example profile is composed of a deepfake image and an author-created about section. No participant data is presented.*

it is a fitting real-world setting to understand tensions in deepfake moderation. Given the professional context of LinkedIn, attackers have found value in conducting real-world deepfake campaigns [13, 16, 101]; however, incorrect moderation decisions have also resulted in real-world economic harm to users [54, 82, 85]. Furthermore, unlike *content-based* moderation common in several text-oriented or pseudonymous platforms (e.g., Reddit [105]), LinkedIn performs *identity-based* verification and moderation to ensure that profiles accurately represent a real individual [69]. While verification should only be based on objective characteristics, there is nothing to prevent the gender and racial characteristics of the investigated profile from influencing a moderator’s decision when such information is available.

3.2 Profile Dataset Construction

In order to collect authentic profiles for the user study, we conducted a separate online survey (n = 298) on Prolific [2] to obtain users’ public LinkedIn profiles and self-identified demographic information.[4] An example profile (not from an actual participant) is shown in Fig. 2.

While a similar profile dataset may be obtained via scraping of LinkedIn’s website, we were opposed to this methodology for several reasons. *First*, although public, the profile owners may not be comfortable if their persona was the subject of human studies, thus we required explicit consent that could only be obtained via a separate survey. *Second*, most profiles do not explicitly self-report race or gender. *Third*, scraping LinkedIn would be in violation of the site’s terms of service [68].

In this survey, participants began by providing basic information about their LinkedIn profile and platform usage (**Q1-Q4**). Afterward, participants chose to either upload their current LinkedIn profile image and “about” text anonymously, or provide a URL to their profile. If a URL was provided, participants needed to confirm profile ownership by following a LinkedIn page we created to verify they could take actions on behalf of the profile and thus were the owner.[5] To deanonymize their identity in the profile, we then asked each participant to create a first name pseudonym of similar

[4] All participants were directly informed that their public data (demographics were not part of this) would be shown to future study participants.
[5] Once all accounts were verified, the page was deleted.

Requirements	Description
Functional:	
Required Sections	Contains a profile image and "about" section
Represents Owner	The image and about section represents the owner
Reported Identity	The owner reported their gender/racial identity
Attentive	The owner passed both survey attention checks
Private*:	
No PII Provided	Beyond the image of the owner, no PII is provided.
Exclusively Owner	No information is provided about others
No Confounding Factor:	
US-EN Writing	The profile was reported to be written in US-EN
No Post-Processing	Images do not contain virtual effects/backgrounds

Table 1. **Profile Dataset Requirements**— All collected profiles were required to meet this criterion for use in our user study. *If possible, minor profile modifications were made to meet privacy criteria (e.g., text that contained "contact me at PII@email.com" may be changed to "contact me at my email").

gender and racial characteristics to their real name (**Q9**).[6] [7] Participants then reported their demographics including their gender identity, racial identity, English fluency, sex, age, and education level (**Q10**-**Q21**).

Profile Filtering and Extraction. To account for extraneous factors and preserve the privacy of the profile owners, we systematically filtered/modified certain profiles according to a set of requirements shown in Table 1. We continued collecting until we had at least 25 profiles within each identity group to provide a degree of generalization over that evaluated identity [95]. Overall, out of the $n = 298$ submitted surveys, $n = 160$ profiles met these requirements and were included in our final dataset. For the profile collection survey, participants spent 10.4 minutes on average, and similar to other surveys that offer differential payments for hard-to-reach populations [9, 38], participants were compensated between $2.00-$3.00 ($11.50-$17.30 per hour). The demographics of the finalized profile dataset can be found in Table 8 of Appendix D.[8]

3.3 Main Experimental Procedure

Using the collected profiles, we conducted our main experimental procedure (Fig. 1). Each participant first received a brief background defining computer-generated text and images (i.e., deepfakes) and how they can be used to create artificial profiles. Each participant was then asked to review 23 LinkedIn profiles. For the duration of the study, each participant was assigned one of the "Visible Profile" treatments (❸) and shown either just the "about" text, just the name and image, or all content (name, image, and text) for all 23 profiles.

To prevent participants from answering differently when first exposed to the task compared to when they are accustomed to the task (i.e., a learning effect [95]), the 23 profiles were divided into two phases: 5 initial "practice" profiles, and 18 "study" profiles.[9] The "practice" profiles were composed of 5 randomly chosen profiles from the full dataset and only served to acclimatize participants to the study; therefore, these responses are not analyzed.

The "study" profiles were 18 in total and equally divided among the 6 "Profile Identities" (❶) as described in Section 3.1. Participants were recruited based on whether they were one of the four identities of study (Black/white and

[6]Prior work found that perceived identity-based inconsistencies between the presented image and name are a used detection strategy [76].
[7]To avoid bias introduced by members of the research team and respect participants' lived experiences, we allowed participants to choose pseudonyms aligning with their own racial/gender identity. We only performed verification to confirm their that names were appropriate (e.g., not a reversed spelling of a common name).
[8]Due to an error in the recording of one profile, we disregard the data corresponding to this particular profile; this only affected 0.5% of our finalized data.
[9]Participants are not made aware of the different phases, and no visible differences exist.

Race	Black	white	**Total**
Gender			
Woman	172	163	**335**
Man	190	170	**360**
Age			
18-29	124	116	**240**
30-49	170	136	**306**
50-69	64	73	**137**
70+	3	8	**11**
Prefer not to say	1	0	**1**
Highest Education			
High School or Less	70	62	**132**
Some College / 2yr Degree	104	97	**201**
Bachelor's/Post-Grad	188	174	**362**
Prefer not to say	0	0	**0**
Moderator Experience			
None	290	277	**567**
Less than 6 Months	40	29	**69**
6 Months+	32	27	**59**
Total	362	333	**695**

Table 2. **Moderator Demographics** — We present the demographics and moderation of our participants. We intentionally recruited a balanced pool of the four intersection identities of interest in this study (Black/white × woman/man).

a woman/man), thus each participant viewed 3 study profiles that were an "in-group" of their own identity, and 15 study profiles that were an "out-group" (❷).

For each of the 23 profiles, the participant was asked to rate how artificial each profile appeared on a 6-point Likert scale (**Q22**). We opt for a 6-point Likert scale as it enables a more sensitive measurement of potential biases compared to a binary response. This approach still requires participants to make a decisive judgment, mirroring the dynamics of actual moderation where profiles are categorized as either "artificial" or "real". After rating all 23 profiles, participants were re-presented with 6 of their decisions from the "study profiles" (1 randomly selected from each identity) and were then asked to explain what aspects of the profile influenced their decision (**Q23**).

Lastly, participants were asked background questions about their prior experience with content moderation and artificial content (**Q24-Q34**), and their demographics (**Q10-Q21**).

3.4 Experiment Recruitment

Human content moderation is performed by a diverse group of moderators that exist along a continuum of experience [49, pg. 116-135]. Even within a single platform, these include a small group of "expert" full-time staff internal teams who handle particularly challenging/important moderation decisions, alongside a massively larger group of contracted third-party crowd workers who enforce company policy but are trained to a much lesser extent [88]. However, non-professional end-users also play into moderation by managing community groups (e.g., subreddits, Facebook groups), and flagging content for review by other groups. To represent the diversity of experiences that exist within the moderation process, in our study, we recruit moderators without controlling for specific experiences or training.

For the main study, participants were recruited from Prolific [2] and were required to be 18+, from the US, and not a participant of the profile collection survey (Section 3.2) to participate. To achieve a balanced set of identities, we utilized Prolific's gender and racial filters to balance participants in each of the four studied intersectional identities (Black/white × man/woman). Despite this, several participants did not exclusively identify as one of our studied identities during the user study; given the small number of participants and lack of insight we have into these groups, we omit this data. Any response which did not pass both of the two embedded attention checks was also omitted.

Initially, we recruited $n = 497$ participants and reached concept saturation in our qualitative data [30, chp. 7]; however, our quantitative analysis still required more responses,[10] so we recruited an additional $n = 308$ participants who completed the same moderation task but were not asked any open-response questions. Overall, participants who were asked open-response questions spent an average of 19.2 minutes and were compensated $2.80 ($8.75 per hour) and participants who were not asked open-response questions spent an average of 9.2 minutes and were compensated $2.20 ($14.30 per hour).

In total, of the 819 participants who submitted the survey, 695 participants met our identity and attention-based filtering criteria. As shown in Table 2, this resulted in an identity-balanced pool of participants that exclusively identify as Black women (24.7%), Black men (27.3%), white women (23.4%), and white men (24.5%). Furthermore, participants varied in age, ranging from 18-70+ years with a median age of 35-39, and varied in education, with about half holding at least a Bachelor's degree (52.1%). Furthermore, 18.4% of participants reported having previous moderation experience (from less than six months to over four years) on a social platform.

3.5 Participant Protection and Ethics

While all our study procedures were approved by our IRB, we carefully considered the ethical implications of our study beyond these requirements as it involves concerns related to the use of real public profiles and sensitive identities. We implemented several mitigations to prevent harm. First, we acted transparently and allowed for participant autonomy by disclosing the intentions of the study. We used simple language to inform participants that their provided profile would be used to understand biases in content moderation and that their profiles would be viewed by future participants. We also allowed participants to opt-out of the study at any time, and to skip any demographic question they desired. Furthermore, we took steps to ensure no personally identifiable information (PII) existed in the provided public profiles (see Section 3.2). The collected demographics were never released to anyone outside the immediate research team and were only used to perform the analyses presented in this paper.

3.6 Limitations

First, while we study a set of identities motivated by prior work [14, 25, 47, 55, 92, 99, 109], these represent a narrow set of identities that may be affected by such processes. We focus on this narrow set of identities to ensure enough power in our analysis, and to provide a set of findings upon which future work can build. *Second,* as our focus is on how false positives are materialized and whether they are *disproportionately assigned* to any identity groups, we intentionally cause a demand effect that encouraged participants to mark some portion of real profiles as artificial. Thus while our study gives insights into why these mistakes are made and whether they are biased, we cannot be sure that the proportion of measured false positives translates into a real-world setting. *Third*, while we attempt to keep profiles as close as possible to their online presentation, to preserve the privacy of participants, we used participant-provided pseudonyms rather than their real names, and minimally changed text content to remove PII (Section 3.2). *Fourth*, while we hold a number of necessary requirements to reduce confounding factors and ensure profile owner privacy, this also prevents us from investigating profiles that are outside of these requirements (e.g., ones that display PII, are written in a non-US-English language, or do not contain profiles images). *Fifth*, while LinkedIn profiles were chosen for their

[10] As determined a priori via a simulation-based power analysis for generalized linear mixed models [52].

general image and text-based structure found on many social media platforms, ultimately we cannot generalize beyond LinkedIn. *Sixth*, recruiting from Prolific may lead to certain biases in our presented profiles and moderators; however, these biases are also typical of those found within LinkedIn's user base [36, 37]. *Seventh*, while we recruit a diverse range of moderation experience, we do not claim how such experience may affect our results, as our related analyses are exploratory (see Section 4.2). Future work should continue assessing its impact on moderation errors and biases.

The Authors' Positionality. Throughout this research, we carefully reflected on our position as researchers and inspected how our identities, backgrounds, and perspectives may have influenced the study design and analysis of the results. As the study investigates forms of bias that some or none of us may experience, we discuss our motivations and relevant backgrounds here.

As researchers working within usable security and privacy, we are increasingly observant of the ways that gender and racial biases can have disparate impacts. Further, findings in prior work [14, 25, 47, 55, 76, 92, 99, 109] have led us to hypothesize that bias exists within human content moderation and motivated us to study the research questions identified in this study.

The authors of this paper have knowledge and prior expertise in studying user perceptions of computer-generated content. Other co-authors have knowledge and prior expertise studying security and privacy concerning historically marginalized populations. Another co-author has knowledge and prior expertise in studying the psychological dimensions of stereotyping and prejudice. Through our collaboration, we seek to provide insight into the technical and statistical aspects of content moderation and bias, as informed by our technical and statistical expertise. However, we acknowledge that we do not provide insight into the lived experience of biased content moderation, and refer readers to other work beginning to explore such topics [55]. This team includes [race and gender of research team members, anonymized for review]. Intersectionality lends valuable perspectives to research, and as our team does not include Black women, there are lived experiences that our positionality does not reflect.

4 BIAS IN THE MODERATION OF ARTIFICIAL PROFILES

To determine whether human moderators' decisions are biased by profile identity (**RQ1**), we evaluated whether the gender or race of a person in a profile changed moderators' likelihood to rate the profile as artificial (**Q22**). Specifically, we investigate how ① profile identity, ② moderator identity, and ③ profile content affect the perceived artificiality of profiles.

4.1 Summary Statistics

Participants slightly to strongly agreed that 40.8% of presented profiles were artificial (*all* of which are real). However, when we view across our variables of interest, we see that belief of profile artificiality varies. As shown in Fig. 3, when evaluating text-only profiles, participants evaluated similar percentages of profiles across identities as artificial: 44.3% of Black woman profiles, 45.2% of Black man profiles, 43.0% of white woman profiles, and 44.5% of white man profiles. However, when looking at profiles with all content (i.e., images, names, and text), we begin to see divergences in perceived artificiality; the percentage of profiles rated as artificial decreased for profiles of Black women (from 44.3% to 36.3%) and Black men (from 45.2% to 41.7%) but stayed similar for profiles of white women (41.6%) and white men (45.8%). These differences became more evident when evaluating profiles with only image and name content; the percentage of profiles rated as artificial decreased for Black women (from 44.3% to 28.4%) and Black men (from 45.2% to 34.3%) but stayed similar for profiles of white women (41.9%) and profiles of white men (43.6%).

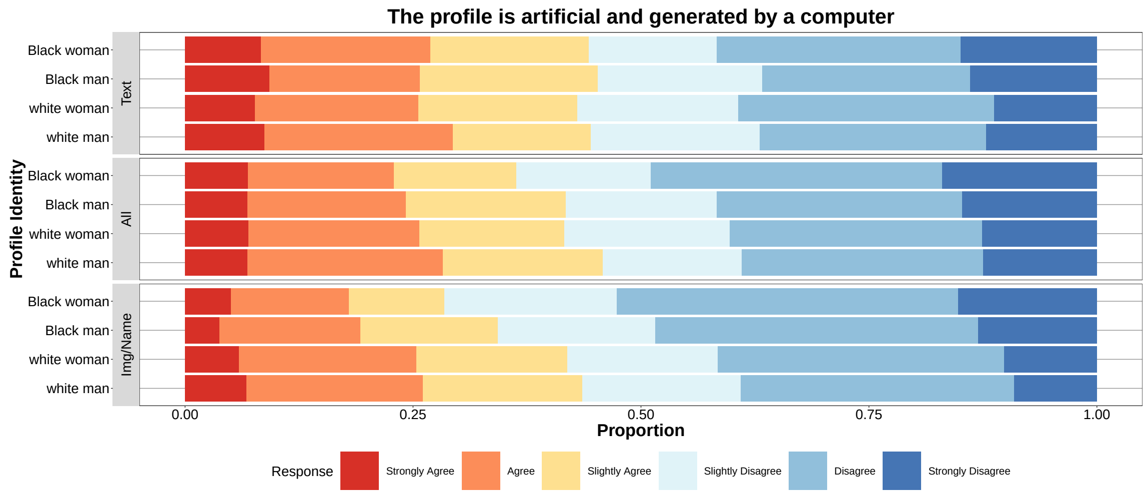

Fig. 3. **Effect of Profile Identity and Profile Content on Artificiality** – Participant's agreement with the statement "The profile is artificial and generated by a computer" (**Q22**). We partition the responses by two of our controlled treatments: ① profile identity and ③ profile content.

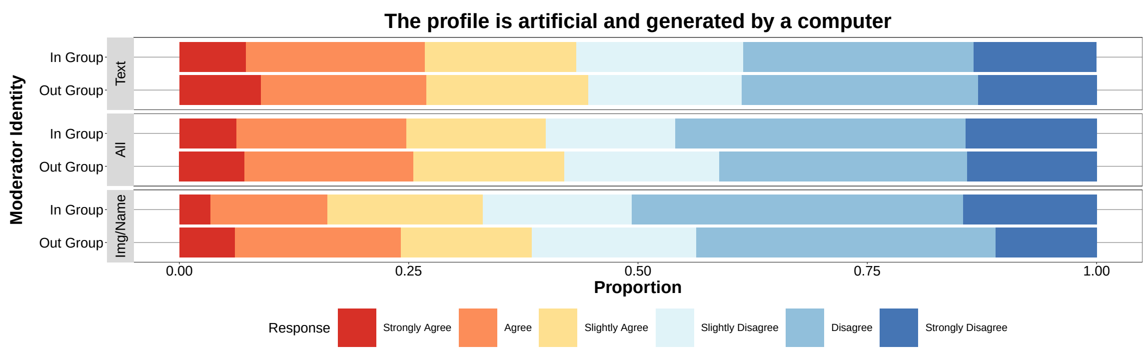

Fig. 4. **Effect of Moderator Identity and Profile Content on Artificiality** – Participant's agreement with the statement "The profile is artificial and generated by a computer" (**Q22**). We partition the responses by two of our controlled treatments: ② moderator identity and ③ profile content. "In Group" means the moderator and the shown profile self-identify as the same identity group.

Investigating the effects of moderator identity and profile identity reveals similar trends. We define a moderator's identity as being "in-group" when the moderator and the shown profile are within the same identity group, and "out-group" otherwise. As shown in Fig. 4, when shown text-only content, the moderators' identity appears to have little correlation with the perceived artificiality of the profiles: 43.3% of profiles with in-group identities were evaluated as artificial, compared to 44.6% of profiles with out-group identities. For profiles with all content (image, name, and text) these differences were also minimal: 43.3% were perceived as artificial for in-group and 44.6% for out-group. In contrast, when moderators were shown profiles with only the image and name (but not text) content, 33.1% of profiles with in-group identities and 38.4% of profiles with out-group identities were perceived as artificial.

Factor	Likelihood Ratio χ^2	P-value
Primary Factor		
Profile Identity (PI)	**48.497**	**<0.001**
Moderator Identity (MI)	**12.625**	**<0.001**
Profile Content (PC)	**9.309**	**0.010**
Two-way Interaction		
PI : PC	**16.675**	**0.011**
MI : PC	**6.658**	**0.036**
PI : MI	2.174	0.537
Three-way Interaction		
PI : MG : VP	9.2526	0.16

Table 3. **Factors' Significance on Perceived Artificiality** – Via an analysis of variance, we find significant primary effects in each factor as well as several significant two-way interactions. Rows that denote significant relations are bolded.

4.2 Statistical Analysis

To statistically evaluate our results, we modeled our data with a Cumulative Link Mixed Model (CLMM) regression [27], to see if the estimates of the fixed effects are significantly different from one another. As opposed to other forms of hypothesis testing on ordinal, non-parametric response variables (e.g., a Kruskall-Wallis H-test [33]), CLMMs allows the modeling and testing of variables of interest via *fixed effects*, while accounting for the non-independence between measured outcomes via *random effects* [111]. As our study asks each participant to make multiple decisions over multiple trials, modeling this non-independence via a mixed effects model is most appropriate.

To investigate how ① profile identity, ② the moderator identity, and ③ profile content affect the perceived artificiality of a profile, we modeled each treatment as a fixed effect. We theorize that perceptions may be different between identities and that these identities are more or less prominent with different profile content, thus we include interaction effects between all three primary factors. This allows us to evaluate whether a change in one factor changes the effect of another factor (e.g., the effect of profile identity may be different if only the text is displayed vs. if only the image/name is displayed). Thus in total we have three primary effects (①,②,③), three two-way interactions (①:②, ①:③, ②:③), as well as one three-way interaction (①:②:③). As each participant may have varying propensities for believing profiles are fake, we account for this non-independence by modeling each participant as a *random effect* in the model. As we are performing a confirmatory analysis of a controlled experiment, we limit our model to our controlled factors (e.g., gender-racial identity of profile/moderator, and profile content). As our experimental set-up has taken measures to reduce the differential effects of groups to their controlled treatments, we don't include other explanatory variables [84, pg. 343-346]. Thus, we follow a design-driven model specification rather than a data-driven specification (e.g., one that iteratively uses goodness of fit or information criterion as a metric for forward/backward model selection). Once the model was chosen, we performed a power analysis using a small set of pilot data to estimate our effect sizes and recruited a number of participants to try to ensure that each non-interaction factor had sufficient power (>80%) for the estimated effect size to be found [20].[11]

Profile artificiality is affected by profile identity, moderator identity, and profile content. To determine whether any of our factors significantly influenced perceptions of artificiality, we conducted an ANOVA test over our fitted model [44]. As shown in Table 3, we find a significant relation in all three of our primary effects of profile identity (PI; $p < 0.001$), moderator identity (MI; $p < 0.001$), and profile content (PC; $p < 0.01$); however, we also find that each of these factors is also part of significant two-way interactions. Specifically, we find a significant relationship between the between-subjects condition (which "Profile Content" is shown) and the within-subjects treatment effects: profile

[11]We did not consider the interaction factors since the estimated effect size was small and the found differences may not be of value.

Content	Factor Level Comparison	Est.	SE	P-value
Image & Name	Profile Identity			
	Black woman – Black man	-0.260	0.191	0.083
	Black woman – white woman	**-0.560**	**0.113**	**<0.001**
	Black woman – white man	**-0.628**	**0.113**	**<0.001**
	Black man – white woman	**-0.300**	**0.111**	**0.033**
	Black man – white man	**-0.368**	**0.119**	**0.005**
	white woman – white man	-0.068	0.114	0.934
	Moderator Identity			
	In Group - Out Group	**-0.314**	**0.076**	**<0.001**
All	Profile Identity			
	Black woman – Black man	-0.184	0.119	0.409
	Black woman – white woman	-0.237	0.122	0.208
	Black woman – white man	**-0.383**	**0.12**	**0.008**
	Black man – white woman	-0.053	0.116	0.968
	Black man – white man	-0.199	0.114	0.298
	white woman – white man	-0.146	0.116	0.593
	Moderator Identity			
	In Group - Out Group	-0.111	0.081	0.167
Text	Profile Identity			
	Black woman – Black man	0.061	0.119	0.957
	Black woman – white woman	0.153	0.12	0.575
	Black woman – white man	-0.047	0.119	0.979
	Black man – white woman	0.093	0.116	0.855
	Black man – white man	-0.108	0.115	0.783
	white woman – white man	-0.201	0.116	0.304
	Moderator Identity			
	In Group - Out Group	-0.032	0.080	0.692

Table 4. **Factor Level Comparison by Profile Content** — The post-hoc analysis for statistical variance across different profile identities and moderator groups, under different profile content conditions. When comparing two factors, if the directionality is negative, it means the first factor is perceived as less artificial compared to the second factor. E.g., "Image & Name: Black woman – white woman=−0.560" implies that the profiles of Black women are perceived as less artificial than the profiles of white women.

identity (PI:PC, $p < 0.05$) and moderator identity (MI:PC, $p < 0.05$).[12] This both provides an answer to **RQ1** – perceived artificiality of moderators varies based on profile identity, moderator identity, and which profile content is shown – and informs the rest of our analysis. To properly interpret our findings, we continue our analysis by investigating the effects and relationship between these factors via deeper post hoc analysis. Specifically, we evaluate the statistical variance across different profile identities (**RQ1**a), moderator groups (**RQ1**b), and profile content (**RQ1**c) using separate Tukey-adjusted post hoc pairwise tests over the data in each profile content level (e.g., Image & Name, Text, and All).

When shown the "image and name", biases from profile identities and moderator identities exist. As shown in Table 4, when just the image and name are shown to participants, we find several significant differences between profile identities (**RQ1**a) and whether the moderator was part of that identity group (**RQ1**b). In this condition, the profiles of Black women are perceived as significantly less artificial than the profiles of white women ($p < 0.001$) and white men ($p < 0.001$); similarly, the profiles of Black men are also perceived as significantly less artificial than the profiles of white women ($p < 0.05$) and white men ($p < 0.01$). When considering the identity of the moderator, we also find that profiles that share the same identity as the moderator (e.g., in-group) are perceived as significantly less artificial than profiles that don't share the same identity ($p < 0.001$).

Biases are lessened when "text" is included and the "image and name" are removed from profiles. From Table 4, we also find changes in the effects on bias depending on what profile content is shown (**RQ1**c). As previously noted, we find significant differences in artificiality due to profile identity (**RQ1**a) when showing image and name

[12]Due to the significance of the two-way interactions, we do not interpret our primary factors alone as doing so may result in incorrect conclusions [95].

Identity	Factor Level Comparison	Est.	SE	P-value
Black w.	Profile Content			
	Image & Name - All	-0.308	0.137	0.063
	Image & Name - Text	**-0.736**	**0.136**	**<0.001**
	All - Text	**-0.428**	**0.142**	**0.007**
Black m.	Profile Content			
	Image & Name - All	-0.233	0.129	0.170
	Image & Name - Text	**-0.416**	**0.131**	**0.004**
	All - Text	-0.183	0.134	0.357
white w.	Profile Content			
	Image & Name - All	0.015	0.135	0.994
	Image & Name - Text	-0.023	0.134	0.984
	All - Text	-0.037	0.137	0.960
white m.	Profile Content			
	Image & Name - All	-0.064	0.133	0.881
	Image & Name - Text	-0.156	0.133	0.471
	All - Text	-0.092	0.134	0.771

Table 5. **Factor Level Comparison by Identity** — The post-hoc analysis for statistical variance across different profile content, under different profile identity conditions. When comparing two factors, if the directionality is negative, it means the first factor is perceived as less artificial compared to the second factor. E.g., "Black woman: Image & Name – Text=−0.308" implies that the profiles that show only Image & Name content are perceived as less artificial than the profiles that show only Text content.

content; however, when we instead show all content, only one significant identity-based profile difference is found: the profiles of Black women are perceived as significantly less artificial than the profiles of white men; however, no significant differences are found between profiles of any other identity. Additionally, while the intersection between moderator and profile identity (**RQ1**b) is significant for image and name content, when shown all content no significant differences between in-group and out-group identities are found. Furthermore, these differences appeared to be further minimized when only showing the text of the profile; in these cases, we found no significant differences in artificiality between any profiles of any identity or whether the moderator's identity was in-grouped or out-grouped.

Certain profile identities are more affected by changes in visible profile content than others. We can also investigate how each identity is affected by the change in profile content. The corresponding post-hoc analysis is presented in Table 5. It shows profiles of Black women have perceived differences of artificiality when comparing profiles with images to those that don't (Image & Name - Text, p<0.001; All - Text, p<0.01;). We notice similar trends for Black men, however, we only find significant differences when comparing between Image & Name and Text (p<0.01). For both white women and white men, however, no differences in perceived artificiality occur when different portions of the profile content are displayed.

Moderation experience may not affect identity-based biases. It is possible that participants who have previously moderated online content may hold different biases than participants who are new to moderating. While prior moderation experience was not one of our a priori research questions or control variables, we exploratorily investigate this by re-fitting our CLMM with an additional fixed factor, "prior moderation experience" (**Q33**), and re-conducting an ANOVA test of this model (Table 9 in the Appendix).

First, by including prior moderation experience our fitted model did not have a significantly better or worse fit than our original model (as measured by Akaike Information Criteria and Bayesian Information Criteria [23]). Second, when performing an ANOVA test, we found that while all of our significant primary and secondary factors of our original model still significantly explain participant's perception of profile artificiality (e.g., $p<0.05$ for PI, MI, PC, PI:PC, and MI:PC), prior moderation experience did not significantly explain measured artificiality. However, as this was

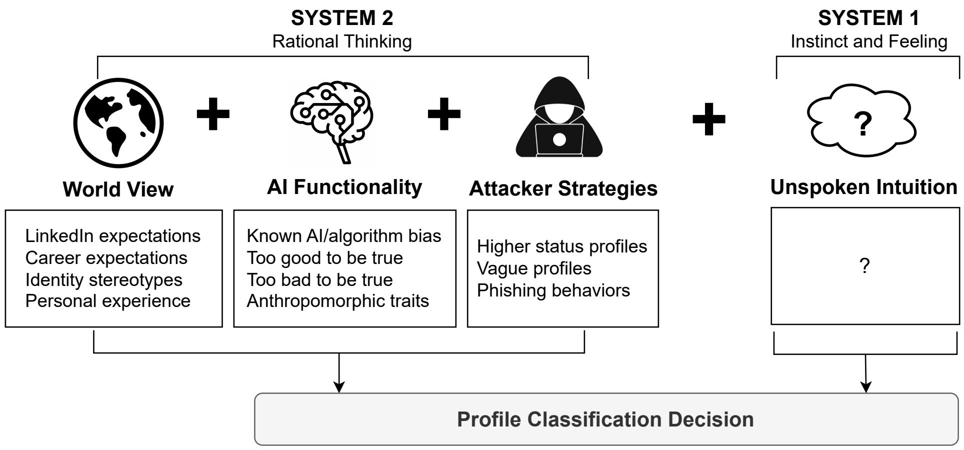

Fig. 5. **Participants' Mental Model for Classifying Deepfake Profiles**—based on our qualitative analysis. Key themes are organized under System 2 (rational thinking) and System 1 (instinct and feeling), based on the dual-process model [63].

an exploratory analysis, we do not make causal claims. Instead, future work should continue to assess the impact of moderation experience on biases.

5 MENTAL MODELS OF ARTIFICIAL PROFILES

We find that participants' incorrect evaluations of real profiles as artificial (**RQ2**) were informed by their mental model (Fig. 5): a combination of conceptions of 1) the real world, 2) AI functionality, and 3) common attacker strategies alongside 4) unspoken intuition. We identify several inaccuracies in these models and find that these inaccuracies may affect some profile identities more than others.

Qualitative Analysis. We qualitatively analyzed participants' responses (**Q23**) using an inductive thematic coding approach and then analyzed the resulting codes to form high-level themes and theories of participants' mental model. In total, we received 1,622 responses describing why participants perceived a particular profile as artificial or real from the 497 participants who were asked open-answer questions (see Section 3.4). The coding process began with one coder developing a codebook using 300 responses (18%). A second coder used this initial codebook to code 60 responses and calculated the resulting inter-rater reliability (IRR) of the codes using Cohen's-κ [29]. If $\kappa \leq 0.7$ for any one code, the coders met to resolve disagreements via relabeling and codebook changes[13] and then repeated this process for an additional 60 responses. On average, it took 4 rounds of coding to reach an agreement for each code. Once IRR was reached for all codes ($\kappa \geq 0.7$), one coder then independently coded until 50% of the total data was analyzed; by this point no new ideas were emerging and concept saturation was believed to be reached [30, chp. 7]. In total 811 responses were coded; the full codebook and counts are in Appendix C. These codes were analyzed via reflexive thematic analysis [17] to generate the high-level themes and theories of mental models presented here.

5.1 Participants' Mental Model of Profiles in the Real World

Participants often compared shown profiles to their "typicality expectations" [116] of profiles in the "real world" to determine whether the shown profile was artificial. Similar to prior work on phishing detection [116], some participants

[13]If codebook changes occurred, all previous data was re-coded.

believed that deviations from their expectations meant the profile was more likely to be artificial. However, we also observed the opposite: some participants believed that if a profile followed expectations too closely, they were artificially crafted to align with expectations. Broadly, participants' pre-existing beliefs included expectations related to 1) LinkedIn profiles, 2) personas and careers, and 3) identity-based stereotypes.

LinkedIn Expectations. Given that the profiles in our study were shown in the context of LinkedIn, participants' evaluations depended on their expectations of how people would act on the site and in a professional manner; we do however stress that what constitutes a "professional expectation" has been known to be influenced by culture and stereotypes [72]. Often, if a participant perceived a profile as deviating from their expectations of a LinkedIn, the profile was considered artificial. For example, P268 perceived a profile as artificial because it did not align with their expectations of LinkedIn's purpose: "*To me just writing about your blog on a site like LinkedIn is weird and does not exactly fit the purpose I believe LinkedIn is there for. LinkedIn is usually a way to connect with others for job positions or to grow a network, but this writer just talks about their blog.*" Another participant believed a profile was artificial because of the photo composition: "*The background of the photo is unusual, and the shot is a bit unusual for LinkedIn*" (P218).

Career Expectations. Participants also expect the content of the profiles to match the stated career type. For example, participants often note that certain careers should hold particular skills, and that the lack of required skills or inclusion of unnecessary skills are considered artificial: "*It simply doesn't make sense to me to have all of those specialties*" (P248). Participants also considered skills demonstrated in the profile itself, for instance, if the author is organized, professional, or a good writer: "*This writing style also doesn't match what I would expect from someone education in Communications*" (P388). Other participants noted that profiles with only expected skills appeared mass-produced and thus artificial: "*It felt like all of the things they were skilled in were just "keywords" that were being used to show up in more searches. I think all nurses would be skilled in most of those things like CPR. A nurse shouldn't need to advertise that they are skilled in CPR because it's pretty much a given*" (P202).

Participants also considered a person's appearance in the profile image, compared to their expectations of people in that career: "*He looks like someone who would be a licensed CPA*" (P277); "*I can see him being a medical tech*" (P385). These participants did not give further detail of what made them perceive particular images as representing someone who exemplified a particular career; however, expectations of people in certain careers are often dependent on identity-based stereotypes and thus may hold biases for different identities [40, 57, 98].

Identity-based Stereotypes. Participants also evaluated profiles based on identity-based stereotypes, e.g., stereotypes related to the race/gender of the person in the profile. P376 for instance, rated a profile as artificial because of a mismatch between the perceived racial identity of the name compared to the image: "*I did feel like maybe the name was a little off. I haven't met many Black men named Adam.*" P255 also notes a similar relationship; however, they felt the identity of the owner too closely followed stereotypes by having the same name as another famous Black African-American: "*It's a random Black man with the title of 'Barack' under him.*" Similarly, P159 notes that a profile appears real as a perceived hairstyle makes sense given the current cultural context of an identity: "*Her hair resembles styles that are current for African Americans.*"

Participants also used the perceived age of the people in profiles to make assumptions about career progression: "*I just have a hard time believing that this very young person is so accomplished*" (P250). Age was also used to determine ostensibly age-appropriate profile behaviors and appearances. P243 believed a profile is real, not only because "*he's talking about being passionate*" but importantly because "*[this] is kind of a thing with the kids these days.*" The alignment with real-world expectations can also help explain other perceived discontinuities. For instance, P297 notes that although the shown profile has an awkward image one might not typically see, it is aligned with their expectation of what an older user might do, and outside of what generated content can simulate: "*This picture screams old guy that doesn't quite know how to take proper selfies and I don't think AI can capture that aura.*"

Personal Experiences. In justifying why certain profiles are real, participants also noted how their personal life experiences have constructed and altered their worldview, and ultimately informed their perspective of what is typical. For instance, P277 noted that while a profile of someone from Louisana did not align with any common stereotypical expectations of a Lousianan, it did align with a personal experience that shaped their expectations, and thus appeared real: "*The woman in the profile pictures looks like someone I know who goes to [a university in] Louisiana, which is not the most objective judgment but that is the honest truth.*" In addition to identity, these experiences also influence participants' perception of career-related personas. For instance, while P58 felt that the shown profile was plain, they also believed the profile was real: "*It also kind of reads like a college student who just doesn't have much to put on their profile yet. My own profile looked a lot like this when I was in school*". Speaking more generally to the sense of familiarity, P142 noted that "*The person in the profile has an almost familiar feeling to them. I mainly based my answer on that.*" This may imply that participants who hold more diverse experiences with a range of people may have different perceptions of "what personas are real", compared to participants who have limited experiences with different people. This is similar to work in psychology that finds that social contact between identity groups reduces prejudice between those groups [94].

5.2 Participants' View of AI Functionality

When determining if a profile is artificial, participants draw on their knowledge of what deepfake algorithms, or more generally AI, tend to produce. While some of these perspectives are informed by academic and industry news, they are also informed by "folk theories" on what algorithms can do and how they behave [35, 43]. Generally, we find 1) that participants' understandings of common biases in AI systems lead to bias in their content moderation behaviors, 2) that participants hold conflicting views about AI algorithm performance—whether AI tends to produce outputs that are perfect or error-ridden—-that can lead to false detection, and 3) that anthropomorphized views of AI—as cold, narcissistic, or bland—led participants to consider profiles that exemplify those traits as artificial.

Known Algorithm Bias Influences Perceived Artificiality. It is becoming increasingly established that machine learning algorithms struggle with the representation of certain identities (commonly, Black people [74]), and several participants use their knowledge of such biases to inform their decision of whether a profile of a given identity is from an AI model. In particular, participants noted that algorithms may be biased toward representing specific identities poorly; as such, a high-quality profile whose identity is perceived as "being poorly represented by deep learning algorithms" is regarded as less likely to have been AI-generated. For example, P55 noted that: "*AI as it stands right now has a hard time with Black faces and hair. Her hair is in braids and that would be hard for AI to do.*" Similarly, P406 implied that the shown profile of a Black man was unlikely to be created by an algorithm due to prevalent bias within them: "*Truthfully, I listened to a news story recently that AI is 'taught' to be racist since it's fed white-biased information.*" Conversely, it may also be the case that identities perceived as being served by this algorithmic bias may be perceived as being more likely to be AI.

Too Good to be True. As several participants believed that AI could produce high-quality text and images, participants often commented on the quality of a profile, noting that artificially-generated content often appears more pristine than human-generated content. These participants tended to believe that perfect grammar indicated algorithmically-generated content: "*No grammatical errors, good vocabulary and structured perfectly. No human mistakes*" (P194). Similarly, highly structured content was perceived as templated and AI-generated: "*This is presented in a very organized manner which makes me think [it] could be AI*" (P214). Overall, several participants felt that text mistakes were telling of human fault and thus an indication of realness: "*'A graduated integrated studies major' is not proper English - a computer would not make this mistake*" (P275).

With respect to image content, participants expected AI-generated images to appear symmetrical, have perfect lighting, or have high-resolution: "*This picture look[s] too perfect and professional for someone to have created it;*" (P170); in contrast real images held natural errors: "*The photo looks like it was taken with a typical mobile phone. I think if it were generated by a computer, the lighting would likely be cleaned up a bit*" (P218). However, some participants believe AI could easily fake this as well, intentionally adding blemishes to compensate for their perfection: "*The glare could have easily been placed there by AI to 'trick' the human mind into thinking the pic must be real*" (P321).

Too Bad to be True. In contrast to believing that AI produces high-quality text and images, several other participants also believed that AI outputs may contain errors/artifacts. Importantly, certain profile signals that were regarded as signs of *authenticity* (see the above subsection) were believed to be signals of *artificiality* of these other participants.

Participants used narrative structure and detail depth in the profile text to distinguish real and artificial content. For example, P421 believed AI content would not have a cohesive story: "*This just reads like a jumble of buzzwords with no point or detail. This reads like it was written by an AI with poor direction rather than a person who has actual experience to share.*" Vague writing was also attributed to algorithms that may not understand the semantic meaning of the information they were producing, e.g., "*the writing is pretty vague which could be mindless computer*" (P241) and "*Way too much detail. I can't imagine how difficult it would be to feed a prompt that led to this output*" (P109).

For image content, participants often focused on artifacts perceived to be common in AI-generated images, e.g., blurry/plain backgrounds and distorted facial features. Several participants focus on the person in the image themselves to make the decision, noting whether they have an awkward facial expression or a disproportionate body: "*The photo looks a little off. His eyebrows are weird and there is some distortion around the right side of his face. It just doesn't look real*" (P55). Some participants also focus on constraints that are hard to replicate by AI in generated images, for instance, appropriate shadows for an object and proper reflection of light: "*Mari's picture looks like a selfie taken in front of a window and you can see the phone reflection in the glasses. I do not think that can be replicated with AI*" (P16).

Finally, some participants held bimodal perceptions of AI-generated content: "*From my experience, AI tends to have an extremely polished feel to it or is a complete disaster. This felt really human because it isn't shockingly perfect and the choice of words aren't extremely encyclopedia-like*" (P130).

Anthropomorphized Views of AI. Participants often use personability as a proxy for real content. Profiles that incorporated personal details in the text or image were often seen as more real. Not only did this depict a life outside of LinkedIn, e.g., "*the description is too personal and speaks of a real human experience with evidence that more than likely shows that this person is human*" (P125), but also was shown to establish the profile as holding human interests and emotions. Several participants associated humanness with warmth, e.g.,"*this profile's sense of humanness and passion influenced my answer the most. It talks about how they love spending time with family and other enjoyable pass-time activities. That makes me feel as if it was written by a real person,*" as opposed to "*The bland and empty feeling this profile provides is screaming of a computer-generated profile*" (P260). Other qualities were also noted to appear to be related to AI-generation, including the use of third-person (instead of first-person) pronouns when describing themselves, the use of formal (instead of colloquial) language, and the display of callousness. Interestingly, some participants viewed negative qualities such as narcissism and egotism as being related to AI: "*It is written in a very narcissistic way which* *makes me think it is AI generated*" (P355). Some participants also noted this belief around personality could be abused to make profiles even more deceptive: "*I would expect AI to try to seem more thorough and personable*" (P72). As a whole, participants tended to view bland, fact-driven profiles as being related to AI content, and thus more likely to be artificial.

5.3 Attacker Strategies on Online Platforms

In determining artificiality, participants also draw on their knowledge of what an attacker may be trying to accomplish and how the attacker might present themselves to best achieve their goals. Similar to prior work investigating users' mental models of phishing emails [22], we find that participants are more suspicious of profiles that appear 1) to be high status, 2) intentionally vague, or 3) appear to follow known phishing and scam-related behaviors.

Higher Status Profiles are Suspicious. Participants believed that attackers would more likely use a persona they perceive as influential. As such, how influential a profile appeared to be also impacted the perceived artificiality. For example, participants noted that profiles that appeared as if they were trying to be more qualified or successful than they actually were struck them as artificial: "*[it sounds] like it was trying to sound more accomplished than it actually was*" (P8). Similarly, participants noted that personas they perceived as visually attractive were more likely to be used by attackers: "*While the features are pleasant looking, I would think that an image generated by a computer would be much more glamorous/striking in the features*" (P52) and "*Not sure someone would want to create this as a profile as he is not very handsome in my opinion*" (P225). Participants also considered that an interesting profile may be more likely to be artificial than one perceived as common or boring: "*Honestly [the profile] is so plain... It would be weird for an AI to make this profile to trick someone, I am not sure what it would ever accomplish*" (P18).

Vague Profiles are Suspicious. While some participants believed that vague profiles may be due to limitations in AI algorithms (Section 5.2), others believed they were artificial due to intentional attacker strategies. Participants perceived a lack of information to be intentional, leaving fewer possibilities for mistakes. For instance, P106 noted, "*I'm suspicious of any profile that can not be fact-checked and verified. I would have liked more info.*" Though there may be valid reasons for omitting information (e.g., privacy), participants struggled to determine whether the reasons were benign or malicious: "*While the profile uses passable wording it lacks depth. It almost seems created by a non-native speaker... or a robot*" (P318). Generally, participants assumed real people would make detailed profiles on LinkedIn: "*A lot of these entries definitely beg for explanation or more detail, as well. It's hard to imagine a real person editing this and not changing things around, adding more context, etc*" (P352).

Phishy or Scammy Behaviors are Suspicious. Participants also relied on their existing phishing heuristics when trying to identify artificial profiles. The actions suggested in some profiles' "about" sections were perceived as an initial step in a malicious interaction, such as embedded links: "*It screams 'follow this link to this corporate service scam'*" (P318). Others were distrustful of profiles that "*sounds like its trying to convince me*" (P21) or "*feels too compelling*" (P155).

5.4 Unsubstantiated Intuition

Several participants did not provide direct reasoning for their actions when making a decision. Instead, they described generally feeling that something about a profile was off and that their "*initial gut feeling*" (P324) was that the profile was artificial, but could not describe why: "*I believe Valerie is not a real person. There is something off about the picture*" (P418) and "*I felt like the language being used in the description didn't sound very natural*" (P216). Intuition-based responses were also more common when participants saw profiles of an image and name without profile text.

6 DISCUSSION

In summary, we find statistical evidence that human moderation of potential deepfake profiles results in biased misclassification of real profiles for certain identity groups (**RQ1**); in particular, we find that the profiles of Black women and Black men are subject to changes in perceived artificiality depending on which profile content is shown at the time of moderation. These profiles *decrease* in perceived artificiality when the profile image and name are shown, while the perceived artificiality of white women and white men's profiles does not change when the shown profile content is varied. We then investigate how human moderators justify these choices (**RQ2**), finding that participants' mental model of profile artificiality depends on their worldview of authentic profiles (and human behaviors), AI functionality, and attacker strategies on online platforms. From these results, we now consider how this bias comes about during moderation, what practical steps can be performed to ameliorate this situation, and what standards of moderation and practical tradeoffs are reasonable for content moderation.

Through our qualitative results, we find several explanations for the identity-based biases that emerge in our analysis. In particular, we find that participants may base their belief of artificiality on stereotypes and expectations about gender, race, or people in certain careers (Section 5.1), perspective on what identities can be faithfully generated by algorithms (Section 5.2), and the perception that attackers are more likely to create personas of high social status identities (Section 5.3), which may all encourage differential treatment of identities. Based on our results, we provide a set of recommendations to minimize bias during moderation of LinkedIn profiles and other similar digital profiles.

Debias and Stop Anthropomorphizing AI. While many of the beliefs on which our participants relied in making their authenticity judgments are based on gender or racial stereotypes, several are based upon perceptions of AI and cybercriminals that academics and organizations help construct. In particular, we find that the *very real* identity-based biases of machine learning systems not only result in harm due to direct biases from system [21], but also result in downstream effects on people's mental models of AI: in the case of our study, their assumptions regarding which profiles are easier to synthetically generate. Thus, we join the call of many before us [1, 89] to debias AI systems, and remove biased systems from deployment, to avoid direct and downstream harms.

We also find that participants rely on anthropomorphic beliefs about AI – as narcissistic, cold, bland – in judging profiles as deepfake or not. Thus, we offer concrete empirical evidence that anthropomorphizing AI is indeed harmful [42, 118]. We must take care to avoid the personification of AI systems in media and when teaching concepts about AI and machine learning.

Adapt Phishing Training to Accommodate a Fast-Expanding Threat Landscape. We also find that participants adopt knowledge about digital attacks from traditional security domains like phishing. For example, they look for calls to action like URLs and make judgments based on "typicality violations," the presence of content that "violates the person's expectations for what is typically present in similar situations" [116]. However, these phishing-related cues vary in their relevance to the detection of deepfakes. For instance, some attackers may impersonate personas of power, as in spearphishing [24], while prior work suggests that misinformation attackers may also do the opposite [46]. In addition, unlike those used in phishing, links or calls to action are not by themselves suspicious in a LinkedIn profile. Further, attempting to apply "typicality" cues to people's profiles opens the door for reliance on stereotypes in reasoning about typicality. There is a clear opportunity for future work to extend existing phishing training and security education more broadly to prepare end users for the complexity of adapting these cues when faced with a rapidly changing ecosystem of malicious content. Traditional phishing training often focuses on teaching people "conclusive distinguishers/cues" [81, 116], e.g., cues that clearly indicate an email is a phishing threat. However, such cues may not generalize across contexts and are likely to change in the context of generative AI. Attackers will respond to the community's perception of them, and these beliefs of expected personas can and will be undercut to an attacker's advantage. Thus, educators should be cautious about teaching strict rules. Instead, it may be necessary to teach *adversarial thinking* to end users [50, 62, 64, 102] and integrate an emphasis on the triangulation techniques identified by work on misinformation [78, 120] rather than offering quick-changing and context-specific cues. Alternately, it may be increasingly necessary to make it clear what "facts and advice" [117] on one security issue (e.g., phishing emails) do and do not apply to other security problems (e.g., deepfake profiles).

Update Platform Design to Reduce Emphasis on Identity-related Profile Components. Our results suggest that while including the image and name in profile *decreases perceptions of artificiality* for some identity groups, these changes vary depending on the gender and racial identity of the profile owner and thus including the image and name *increases disparities between identities*. While one may argue that adding identity-laden information such as image and name still provides a net benefit – since it decreases perceived artificiality of the profiles of some groups – we caution against designs that emphasize identity-based factors that increase overall disparities. It is not guaranteed that including identity-based information to inform content moderation will not later be used to harm.

However, in certain cases, the name and image can be useful in identifying actual deepfake profiles. In algorithmic-based classification, state-of-the-art techniques do make effective use of image-based analysis [26, 115, 123]; however, history shows that they are then optimized against by future deepfakes [61, 83, 86]. Thus, algorithms may obtain real, but temporary, benefits from analyzing identity-laden fields. On the other hand, manual classification does not meaningfully benefit from identity-laden fields. Prior work finds that users are poor at using profile names/images to detect deepfake profiles and remain vulnerable to social engineering [76, 87].

Thus, to minimize bias while retaining proven protections during content moderation, our results suggest caution around moderation user interfaces (UIs) that focus on image and name alone. Instead, we suggest evaluating the efficacy of systems in which e.g., images are evaluated only by processes that have both proven efficacy and routine bias evaluations, such as automated bias-minimized analyses — e.g., a reverse image search that reports the number of matching results, or deepfake detection algorithm evaluated for bias. Additional areas for future evaluation include developing specialized anti-bias training for human moderators.

Beyond the context of explicit content moderation, platform users also perform similar implicit moderation when deciding who to accept connections from [76]; however, the current UI of many platforms, including LinkedIn, only shows the profile's image and name[14] in a request. To prevent biases when users connect to one another, we recommend that when requesting a connection, the UI diminishes the role of identity-embedded information and provides text-based content alongside the request. Future work could evaluate the effect of more intensive measures to reduce focus on identity-based information during first judgments by, for example, only showing the profile image once the full profile is clicked.

Focus on Intra- vs. Inter-group Differences When Composing Moderation Teams. We find that moderators who share the same identity as a profile are significantly less likely to misclassify them as artificial (Section 4). This is consistent with multiple models in social psychology. First, it is consistent with a tendency toward ingroup bias, such that people favor ingroups over outgroups [94]. It is also consistent with research indicating that perceivers are more sensitive to the signaling cues sent by ingroup members [70]. Further, and again consistent with social psychology research on intergroup contact's effects on reducing stereotypes and bias [4], we observe a qualitative trend (Section 5)

[14]Alongside a short headline text, if provided.

such that moderators who hold diverse lived experiences may also rely less on stereotypes during their moderation. Thus we recommend that platforms consider such effects when building their moderation team and when assigning profiles for review. In particular, we recommend that platforms consider testing the impact of assigning profile reviews to moderators of the same identity and prioritize diversity of lived experiences in moderator hiring.

Explore Other Sources of Bias. Finally, while our focus is on gender and race, our qualitative results also suggest the potential for other identity-based biases in moderation of deepfakes. For instance, age was commonly referenced when participants attempted to reconcile career progression with the perceived age of the user. Also, stereotypes related to specific career paths and fields, English literacy, and non-native speakers were noted when discussing grammar/spelling errors and uncommon flow in a piece of text. Furthermore, more private users were mentioned when personality traits such as being expressive, emotional, and revealing specific information were used as signals of artificiality. Future work may evaluate whether our findings – that providing text-based information reduces biases – hold for these other biases, especially those related more specifically to text (e.g., literacy, fluency).

7 CONCLUSION

To investigate whether deepfake content moderation errors hold identity-based biases, we conducted a user study (n=695) asking participants to rate the artificiality of real profiles. We find statistical evidence that real profiles differ in moderator-perceived artificiality based on the identity of the profile, whether the moderator belonged to that same identity, and what profile content was shown. In describing their decisions, we discover that participants' mental models for identifying artificial profiles may use inaccurate identity-based reasoning in their expectations of typicality in the real world, AI functionality, and common attacker strategies. Based on these findings we provide recommendations to minimize bias during the moderation of digital profiles.

REFERENCES

[1] Algorithmic justice league. https://www.ajl.org/about.
[2] Prolific. https://www.prolific.co/.
[3] Paul S Adler and Seok-Woo Kwon. Social capital: Prospects for a new concept. Academy of management review, 27(1):17–40, 2002.
[4] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. The nature of prejudice. 1954.
[5] Carolina Are and Pam Briggs. The emotional and financial impact of de-platforming on creators at the margins. Social Media+ Society, 9(1):20563051231155103, 2023.
[6] Michael Atleson. Chatbots, deepfakes, and voice clones: Ai deception for sale. Federal Trade Commission, March 2023. https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale.
[7] Paul M Barrett. Who moderates the social media giants. Center for Business, 2020.
[8] Emily Bazelon. How to stop the bullies. The Atlantic, 2013. https://www.theatlantic.com/magazine/archive/2013/03/how-to-stop-bullies/309217/.
[9] James Bell, Jacob Poushter, Moira Fagan, and Christine Huang. In response to climate change, citizens in advanced economies are willing to alter how they live and work. Pew Research Center, 2850, 2021.
[10] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9, pages 405–415. Springer, 2017.
[11] Danielle Blunt and Zahra Stardust. Automating whorephobia: sex, technology and the violence of deplatforming: An interview with hacking//hustling. Porn Studies, 8(4):350–366, 2021.
[12] Danielle Blunt and Ariel Wolf. Erased: The impact of fosta-sesta and the removal of backpage on sex workers. Anti-trafficking review, (14):117–121, 2020.
[13] Shannon Bond. That smiling linkedin profile face might be a computer-generated fake. National Public Radio, March 2022. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles.
[14] Sylvie Borau, Tobias Otterbring, Sandra Laporte, and Samuel Fosso Wamba. The most human bot: Female gendering increases humanness perceptions of bots and acceptance of ai. Psychology & Marketing, 38(7):1052–1068, 2021.

[15] Bridget Botelho. Linkedin scams, fake instagram accounts hit businesses, execs. TechTarget, October 2022. https://www.techtarget.com/searchsecurity/feature/LinkedIn-scams-fake-Instagram-accounts-hit-businesses-execs.
[16] Christopher Boyd. Deepfakes and LinkedIn: malign interference campaigns. MalwareBytes, November 2019. https://blog.malwarebytes.com/social-engineering/2019/11/deepfakes-and-linkedin-malign-interference-campaigns/.
[17] Virginia Braun and Victoria Clarke. Thematic analysis., volume 2. American Psychological Association, 2012.
[18] Jenelle Bray. Automated fake account detection at linkedin. LinkedIn, September 2018. https://engineering.linkedin.com/blog/2018/09/automated-fake-account-detection-at-linkedin.
[19] Tina Brooks et al. Increasing threat of deepfake identities. Department of Homeland Security, 2021. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.
[20] Marc Brysbaert and Michaël Stevens. Power analysis and effect size in mixed effects models: A tutorial. Journal of cognition, 1(1), 2018.
[21] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018.
[22] Pavlo Burda, Luca Allodi, and Nicola Zannone. Don't forget the human: a crowdsourced approach to automate response and containment against spear phishing attacks. In 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pages 471–476, 2020.
[23] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. Sociological methods & research, 33(2):261–304, 2004.
[24] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. IEEE security & privacy, 12(1):28–38, 2013.
[25] Stephen Cave and Kanta Dihal. The whiteness of ai. Philosophy & Technology, 33(4):685–703, 2020.
[26] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18710–18719, 2022.
[27] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. Submitted in J. Stat. Software, 35, 2018.
[28] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. arXiv preprint arXiv:2107.00061, 2021.
[29] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46, 1960.
[30] Juliet Corbin and Anselm Strauss. Basics of qualitative research: Techniques and procedures for developing grounded theory. 2008.
[31] O. Corneille, K. Hugenberg, and J. Potter. Applying the attractor field model to social cognition: Perceptual discrimination is facilitated but memory is impaired for faces displaying evaluatively-congruent expressions. Journal of Personality and Social Psychology, 93:335–352, 2007.
[32] National Research Council et al. Measuring racial discrimination. National Academies Press, 2004.
[33] Wayne W Daniel. Applied nonparametric statistics. Houghton Mifflin, 1978.
[34] Homero Gil de Zúñiga, Matthew Barnidge, and Andrés Scherman. Social media social capital, offline social capital, and citizenship: Exploring asymmetrical social capital effects. Political Communication, 34(1):44–68, 2017.
[35] Michael A. DeVito, Jeffrey T. Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. The algorithm and the user: How can hci use lay understandings of algorithmic systems? In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery.
[36] Stacy Jo Dixon. Distribution of linkedin users worldwide as of january 2023, by age group, Feb 2023. https://www.statista.com/statistics/273505/global-linkedin-age-group/.
[37] Stacy Jo Dixon. Gender distribution of social media audiences worldwide as of january 2023, by platform, August 2023. https://www.statista.com/statistics/274828/gender-distribution-of-active-social-media-users-worldwide-by-platform/.
[38] Samuel Dooley, Ryan Downing, George Wei, Nathan Shankar, Bradon Thymes, Gudrun Thorkelsdottir, Tiye Kurtz-Miott, Rachel Mattson, Olufemi Obiwumi, Valeriia Cherepanova, et al. Comparing human and machine bias in face recognition. arXiv preprint arXiv:2110.08396, 2021.
[39] Natasha Duarte, Emma Llanso, and Anna Loup. Mixed messages? the limits of automated social media content analysis. 2017.
[40] Asia A Eaton, Jessica F Saunders, Ryan K Jacobson, and Keon West. How gender and race stereotypes impact the advancement of scholars in stem: Professors' biased evaluations of physics and biology post-doctoral candidates. Sex Roles, 82:127–141, 2020.
[41] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. American economic journal: applied economics, 9(2):1–22, 2017.
[42] Ziv Epstein, Sydney Levine, David G Rand, and Iyad Rahwan. Who gets credit for ai-generated art? Iscience, 23(9), 2020.
[43] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First i "like" it, then i hide it: Folk theories of social feeds. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, page 2371–2382, New York, NY, USA, 2016. Association for Computing Machinery.
[44] Julian J Faraway. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press, 2016.
[45] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. Conformity of eating disorders through content moderation. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1):1–28, 2020.
[46] Deen Freelon, Michael Bossetta, Chris Wells, Josephine Lukito, Yiping Xia, and Kirsten Adams. Black trolls matter: Racial and ideological asymmetries in social media disinformation. Social Science Computer Review, 40(3):560–578, 2022.

[47] Brian Friedberg and Joan Donovan. On the internet, nobody knows you're a bot: Pseudoanonymous influence operations and networked social movements. Journal of Design and Science, 2019.

[48] Javier Galbally and Sébastien Marcel. Face anti-spoofing based on general image quality assessment. In 2014 22nd international conference on pattern recognition, pages 1173–1178. IEEE, 2014.

[49] Tarleton Gillespie. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018.

[50] Mark Gondree, Zachary NJ Peterson, and Tamara Denning. Security through play. IEEE Security & Privacy, (3):64–67, 2013.

[51] David Marvin Green, John A Swets, et al. Signal detection theory and psychophysics, volume 1. Wiley New York, 1966.

[52] Peter Green, Catriona MacLeod, and Phillip Alday. simr: Power analysis for generalised linear mixed models by simulation. https://cran.r-project.org/package=simr.

[53] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences, 119(1):e2110013119, 2022.

[54] Jessica Guynn. Facebook while black: Users call it getting 'zucked,' say talking about racism is censored as hate speech, April 2019. https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/.

[55] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. In Proc. CSCW, 2021.

[56] Kurtis Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, Dillanie Sumanthiran, Ji-ze Jang, Md Rafayet Ali, and Ehsan Hoque. Could you become more credible by being white? assessing impact of race on credibility with deepfakes. arXiv preprint arXiv:2102.08054, 2021.

[57] Madeline E Heilman. Gender stereotypes and workplace bias. Research in organizational Behavior, 32:113–135, 2012.

[58] Julian Van Horne. Shadowbanning is a thing — and it's hurting trans and disabled advocates. Salty, March 2020. https://saltyworld.net/shadowbanning-is-a-thing-and-its-hurting-trans-and-disabled-advocates/.

[59] Nils Hulzebosch, Sarah Ibrahimi, and Marcel Worring. Detecting cnn-generated facial images in real-world scenarios. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 642–643, 2020.

[60] Matthew Hunsinger, Michael Christopher, and Andi M Schmidt. Mindfulness training, implicit bias, and force response decision-making. Mindfulness, 2019.

[61] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 3348–3357, 2021.

[62] Joint Task Force on Cybersecurity Education. Cybersecurity curricula 2017: Curriculum guidelines for post-secondary degree programs in cybersecurity. ACM/IEEE Joint Task Force Report, 2017.

[63] Daniel Kahneman. Thinking, fast and slow. Farrar, Straus and Giroux, 2011.

[64] Tadayoshi Kohno and Brian D Johnson. Science fiction prototyping and security education: cultivating contextual and societal thinking in computer security education and beyond. In Proceedings of the 42nd ACM technical symposium on Computer science education, pages 9–14. ACM, 2011.

[65] Pavel Korshunov and Sébastien Marcel. Deepfake detection: humans vs. machines. arXiv preprint arXiv:2009.03155, 2020.

[66] Pavel Korshunov and Sébastien Marcel. Subjective and objective evaluation of deepfake videos. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2510–2514. IEEE, 2021.

[67] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. IEEE Signal Processing Magazine, 39(1):109–116, 2021.

[68] LinkedIn. Linkedin: User agreement, February 2022. https://www.linkedin.com/legal/user-agreement.

[69] LinkedIn. Professional community policies, July 2023. https://www.linkedin.com/legal/professional-community-policies.

[70] E.P Lloyd, K. Hugenberg, A. R. McConnell, J. W. Kunstman, and J. Deska. Black and white lies: Race based biases in deception judgments. Psychological Science, 28(8):1125–1135, 2017.

[71] Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G Rand. Crowds can effectively identify misinformation at scale. Perspectives on Psychological Science, page 17456916231190388, 2022.

[72] Courtney L McCluney, Myles I Durkee, Richard E Smith II, Kathrina J Robotham, and Serenity Sai-Lai Lee. To be, or not to be… black: The effects of racial codeswitching on perceived professionalism in the workplace. Journal of experimental social psychology, 97:104199, 2021.

[73] Aaron McDade. 'facebook jail' rules loosened as company promises to explain why posts were removed before suspending or banning users, February 2023. https://www.businessinsider.com/meta-facebook-jail-ban-under-new-policy-2023-2.

[74] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM Comput. Surv., 54(6), July 2021.

[75] Meta. Community standards: Manipulated media, July 2023. https://transparency.fb.com/policies/community-standards/manipulated-media/.

[76] Jaron Mink, Licheng Luo, Natã M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In Proc. USENIX Security, 2022.

[77] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1):1–41, 2021.

[78] Imani Munyaka, Eszter Hargittai, and Elissa Redmiles. The misinformation paradox: Older adults are cynical about news media, but engage with it anyway. Journal of Online Trust and Safety, 1(4), 2022.

[79] Diana C Mutz. Population-based survey experiments. Princeton University Press, 2011.

[80] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11):4366–4383, 2018.

[81] Bilal Naqvi, Kseniia Perova, Ali Farooq, Imran Makhdoom, Shola Oyedeji, and Jari Porras. Mitigation strategies against the phishing attacks: A systematic literature review. Computers & Security, page 103387, 2023.

[82] Aaron Nedumparambill. Linkedin kicked me out, without really telling me why. Vice, 2021. https://www.esentire.com/security-advisories/hackers-spearphish-professionals-on-linkedin-with-fake-job-offers-infecting-them-with-malware-warns-esentire.

[83] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 923–932, 2021.

[84] John Neter, Michael H Kutner, Christopher J Nachtsheim, William Wasserman, et al. Applied linear statistical models. 1996.

[85] CBS News. Twitter suspends accounts of several journalists with no explanation, December 2022. https://www.cbsnews.com/news/twitter-suspends-accounts-several-journalists/.

[86] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223:103525, 2022.

[87] Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences, 119(8):e2120481119, 2022.

[88] Annie Njanja. Meta and sama face legal action in kenya for alleged poor work conditions, March 2022. https://techcrunch.com/2022/03/30/meta-and-sama-face-legal-action-in-kenya-for-alleged-poor-work-conditions.

[89] Future of Privacy Forum. Unfairness by algorithm: Distilling the harms of automated decision-making, December 2017. https://www.ajl.org/about.

[90] OpenAI. Dall-e 2, April 2022. https://openai.com/dall-e-2.

[91] Martin T Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In Sociological methods, pages 279–299. Routledge, 2017.

[92] Shelley Park. More than skin deep: A response to "the whiteness of ai". Philosophy & Technology, 34(4):1961–1966, 2021.

[93] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015.

[94] Thomas F Pettigrew and Linda R Tropp. A meta-analytic test of intergroup contact theory. Journal of personality and social psychology, 90(5):751, 2006.

[95] Helen C. Purchase. Experimental procedure, page 51–94. Cambridge University Press, 2012.

[96] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

[97] Sarah T Roberts. Behind the screen: The hidden digital labor of commercial content moderation. University of Illinois at Urbana-Champaign, 2014.

[98] Ashleigh Shelby Rosette, Rebecca Ponce de Leon, Christy Zhou Koval, and David A Harrison. Intersectionality: Connecting experiences of gender with race at work. Research in Organizational Behavior, 38:1–22, 2018.

[99] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5884–5906. Association for Computational Linguistics, July 2022.

[100] Adam Satariano and Paul Mozur. The people onscreen are fake. the disinformation is real., February 2023. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.

[101] Raphael Satter. Experts: Spy used AI-generated face to connect with targets. APNews, June 2019. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

[102] Fred B Schneider. Cybersecurity education in universities. IEEE Security & Privacy, 11(4):3–4, 2013.

[103] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[104] Farhana Shahid, Srujana Kamath, Annie Sidotam, Vivian Jiang, Alexa Batino, and Aditya Vashistha. "it matches my worldview": Examining perceptions and attitudes around fake videos. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2022.

[105] Spandana Singh. Everything in moderation: Case study: Reddit, July 2019. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/case-study-reddit/.

[106] Leron Solomon. Fair users or content abusers: The automatic flagging of non-infringing videos by content id on youtube. Hofstra L. Rev., 44:237, 2015.

[107] Olivia Solon. Facebook ignored racial bias research, employees say, July 2020. https://www.nbcnews.com/tech/tech-news/facebook-management-ignored-internal-research-showing-racial-bias-current-former-n1234746.

[108] StabilityAI. Stable diffusion, August 2022. https://stability.ai/stablediffusion.

[109] Steven J Stroessner and Jonathan Benitez. The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. International Journal of Social Robotics, 11:305–315, 2019.
[110] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. Seeing is believing: Exploring perceptual differences in deepfake videos. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–16, 2021.
[111] Jack E Taylor, Guillaume A Rousselet, Christoph Scheepers, and Sara C Sereno. Rating norms should be calculated from cumulative link mixed effects models. Behavior Research Methods, pages 1–22, 2022.
[112] TikTok. Community guidelines: Integrity and authenticity, March 2023. https://www.tiktok.com/community-guidelines/en/integrity-authenticity/.
[113] Eleanor Tursman. Detecting deepfakes using crowd consensus. XRDS: Crossroads, The ACM Magazine for Students, 27(1):22–25, 2020.
[114] Twitter. Synthetic and manipulated media policy, April 2023. https://help.twitter.com/en/rules-and-policies/manipulated-media.
[115] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the 2022 international conference on multimedia retrieval, pages 615–623, 2022.
[116] Rick Wash. How experts detect phishing scam emails. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–28, 2020.
[117] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In Proceedings of the 2018 chi conference on human factors in computing systems, pages 1–12, 2018.
[118] David Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. Minds and Machines, 29(3):417–440, 2019.
[119] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8261–8265. IEEE, 2019.
[120] Stella Zaryan. Truth and trust: How audiences are making sense of fake news. 2017.
[121] Leslie A. Zebrowitz. First impressions from faces. Current Directions in Psychological Science, 2017.
[122] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2):102025, 2020.
[123] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2185–2194, 2021.

A PROFILE GATHERING QUESTIONS

A.1 Screening Questions

Q1 How often do you use LinkedIn?

- Everyday
- A few times per week
- A few times per month
- A few times per year
- Less than a few times per year [*open text response*]

Q2 What language is your profile written in?

- [*droplist of languages*]

Q3 What do you have as your LinkedIn Profile Photo? (Note: only consider your profile photo, NOT your background photo).

- It is an image of myself only
- It is an image of myself with others
- I have uploaded an image of something other than myself [*open text response*]
- I have the default profile image

Q4 How long is the "about"/"summary" section of your profile?

- I don't have an "about"/"summary" section on my profile
- 1-2 sentences
- 3-4 sentences
- 5+ sentences

A.2 LinkedIn Page Collection

Q5 How do you wish to provide your LinkedIn data?

- I want to link my profile URL (note: "Profile Photo" and "Summary/About" sections must be public for this option).
- I want to upload my "Profile Photo" and "Summary/About text" data manually.

Q6 [If URL selected in **Q5**] Please enter the URL of your LinkedIn Profile:
Important: We will never contact you via LinkedIn. This will only be used to gather the data on your profile.
[*open text response*]

Q7 [If manual upload is selected in **Q5**] To upload your photo, please perform the following steps:
Step 1. Click on the red upload button below
Step 2. Find and submit the formatted photo named "<random_id>.png" or "<random_id>.jpg"
Step 3. Enter your name as "random_id"
Step 4. Enter any email address you prefer (we do not see this)
Step 5. Submit your profile photo
(Upload Button)

Q8 [If manual upload is selected in **Q5**] Please copy-and-paste your About/Summary section text as it appears in your LinkedIn profile
[*open text response*]

A.3 Psueodym Selection

To protect your privacy, we will not use your real first name, but instead, we ask you to provide a first name that you feel is similar to yours.
Please provide a first name that is consistent with your own in terms of the gender and ethnic or cultural components.
Examples:

- "Sarah" may provide the name "Mary" or "Rebecca"
- "Syed" may provide the name "Muhammad" or "Ali"
- "Bon-Hwa" may provide the name "Ye-jun" or "Sung-ho"

Q9 Please enter a first name that is similar to yours.
[*open text response*]

A.4 Demographics

As the this study is intended to identify if there is any bias towards underrepresented or marginalized communities, please note how the following demographics apply to you.

This identifiable information will never be shown to anyone but the research team and will be stored on confidential, encrypted, and password-protected servers. To prevent re-identification of any answers, these results will only be published in an anonymous and aggregated format.

Q10 How would you rate your fluency in reading English?
- Beginner
- Intermediate
- Proficient
- Fully Fluent
- Prefer not to say

Q11 How would you rate your fluency in writing English?
- Beginner
- Intermediate
- Proficient
- Fully Fluent
- Prefer not to say

Q12 Rate how much you agree with the statement: "My "Summary/About" text is fluent."
- Strongly disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Strongly agree

Q13 What sex were you assigned at birth, on your original birth certificate?
- Male
- Female
- Prefer not to say

Q14 What is your current gender identity? (Check all that apply)
- Man
- Woman
- Indigenous or other cultural gender minority identity (e.g., two-spirit)

- Genderqueer/Gender non-conforming
- Prefer to self-describe (please state) [*open text response*]
- Prefer not to say

Q15 [If **Q13** and **Q14** combination is non-traditional] What gender do you current live as in your day-to-day life?

- Man
- Woman
- Indigenous or other cultural gender minority identity (e.g., two-spirit)
- Genderqueer/Gender non-conforming
- Prefer to self-describe (please state) [*open text response*]
- Prefer not to say

Q16 A person’s appearance, style, dress, or the way they walk or talk may affect how people describe them. How do you think other people may describe you? (we recognize this is distinct from identity and focuses on your presentation to others)

- Very/mostly feminine
- Somewhat feminine
- Equally feminine/masculine
- Somewhat masculine
- Very/mostly masculine
- I do not display a gender on this spectrum
- Prefer not to say

Q17 Are you of Hispanic, Latino, or Spanish Origin?

- No, not of Hispanic, Latino, or Spanish origin
- Yes, Mexican, Mexican Am., Chicano
- Yes, Puerto Rican
- Yes, Cuban
- Yes, another Hispanic, Latino, or Spanish origin - Type, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc. [*open text response*]
- Prefer not to say

Q18 What is your race? - Mark one or more answers and type origins.

- White - Type, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc [*open text response*]
- Black or African Am. - Type, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc [*open text response*]
- American Indian or Alaska Native - Type name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc. [*open text response*]
- Chinese
- Vietnamese
- Native Hawaiian
- Filipino
- Korean
- Samoan
- Asian Indian
- Japanese
- Chamorror
- Other Asian - Type, for example, Pakistani, Cambodian, Hmong, etc [*open text response*]
- Other Pacific Islander - Type, for example, Tongan, Fijian, Marshallese, etc [*open text response*]
- Some other race - Type race or origin [*open text response*]
- Prefer not to say

Q19 What is your age?

- 18-19
- 20-24
- 25-29
- 30-34
- 35-39
- 40-44

- 45-49
- 50-54
- 55-59
- 60-64
- 65-69
- 70+
- Prefer not to say

Q20 What is the highest degree or level of school you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received)
- Some high school, no diploma, or equivalent
- High school graduate, diploma, or equivalent
- Trade, technical or vocational training
- Some college/university study or Associate degree (A.A., A.S., etc.)
- Bachelor's degree (B.A., B.S., B.Eng., etc.)
- Post-Graduate degree (Masters, Ph.D, Ed.D, JD, etc)
- Prefer not to say

Q21 Are you physically located in an European Economic Area (EEA) or mainland China?
- Yes
- No

B PROFILE MODERATION QUESTIONS

B.1 Background and Task

Moderating Computer-Generated Profiles on LinkedIn Background: Computer software is capable of generated human-like images and text, often known as "deepfakes". These images and/or text can be used to create artificial profiles on a social platforms. These profiles may then be used to scam other users, gather information about others, or provide false information.

We have collected a set of profiles from LinkedIn. These profiles each contain a name, a profile image, and a self-summary. Some of these may be "**computer-generated**" and contain artificial images or text, and some may be "**human-created**" and written by a real human LinkedIn user.

Your Task: You will be shown 23 LinkedIn profiles. Your job is to determine whether each profile is "**computer-generated**" or "**human-created**".

<u>Please do not use any external resources or tools while performing this task!</u>

B.2 Profile Rating

Q22 Please select the option below that best represents how you feel about the following statement:

The profile is artificial and generated by a computer
- Strongly Disagree
- Disagree
- Slightly Disagree
- Slightly Agree
- Agree
- Strongly Agree

B.3 Profile Rating Explanation

Q23 In response to the statement "The profile is artificial and generated by a computer", you answered: [**Q22's response**].
Please explain your reasoning for your answer to the previous question. What aspects of the profile most influenced your answer and how did they affect your decision?
[*open text response*]

B.4 Post-Task Questions

Q24 **Before this task**, have you ever heard of any of these terms?
- Deepfakes
- Logarithmic Coding

- Homomorphic Encryption
- Neural Networks
- Retro Encabulator
- Javascript
- Peer-to-Peer Connections
- None of the above

Q25 **Before this task**, have you ever seen examples of computer-generated images or text (otherwise known as deepfakes)?
- Yes
- No
- I don't know

Q26 **Before this task**, have you ever had to figure out whether an image or text was computer-generated (otherwise known a deepfake)?
- Yes
- No

Q27 **In this task**, what was your primary strategy in determining if a profile was computer-generated or human-created?
[*open text response*]

Q28 **In this task**, what additional information would have helped you determine if a profile was computer-generated or human-created?
[*open text response*]

People often use a wide variety of profile characteristics to help them decide whether a profile is computer-generated or human-created.

Q29 What, if any, **text characteristics** helped you determine if a profile was computer-generated or human-created?
[*open text response*]

Q30 What, if any, **image characteristics** helped you determine if a profile was computer-generated or human-created?
[*open text response*]

Q31 If you had the option to, would you have used a tool to assist you with your decision?
- Yes
- No

Q32 **[If Q31 == Yes]** Which tools would you use and why?
[*open text response*]

Q33 Have you ever reviewed/moderated content for a social platform?
- Yes
- No

Q34 **[If Q33 == Yes]** How much experience do you have reviewing/moderating content for a social platform?
- less than 6 months
- 6 months - 1 year
- 2 years - 3 years
- more than 4 years

B.5 Demographics

[Same as A.4]

C CODEBOOK

Primary Code	Subcode	Freq.	Description
Perception	**Authentic** (κ=0.77)	441	Supports the belief that the profile is real or authentic.
	Fake (κ=0.73)	330	Supports the belief that the profile is fake or inauthentic.
	Uncertain (κ=0.88)	49	Supports valid reasonings for the profile being both real and fake.
Reasoning	**Intuition** (κ=0.87)	159	Due to inherent feeling or unexplained beliefs about the profile.
	Inter-field Relation (κ=1.00)	47	Due to an inconsistency between profile fields, or the profile and the context of the platform.
	Name (κ=1.00)	13	Due to a name-related phenomena.
	Image (κ=0.75)	309	Due to an image-related phenomena.
	About(κ=0.76)	458	Due to the about section-related phenomena.

Table 6. **Profile Reasoning Codebook: Primary Codes** — We show the code frequencies and related descriptions for our primary codes.

Primary Code	Subcode	Freq.	Description
Reasoning: Name	**Last Name** (κ=1.00)	3	The lack of a last name.
Reasoning: Image	**Photo Quality** (κ=0.86)	78	A meta-quality such as sharpness, resolution, lighting, or focus of the photo.
	Photo Type (κ=0.82)	78	How to how the photo was taken, e.g., whether it was professional, a selfie, or any photo structure.
	Background-Related (κ=0.90)	67	A background phenomena (e.g., blurriness, specific objects, transition to foreground)
	Person-Related (κ=0.89)	105	A person phenomena (e.g., identity, facial expressions, facial symmetry, clothing, personality)
Reasoning: About	**Personality** (κ=0.86)	117	A personality trait that is apparent due to their writing (e.g., professional, direct, personable).
	Quality (κ=0.83)	235	A writing-specific trait (e.g., complexity, structure, logic, length, repetition, specificity)
	Choice of Words (κ=0.83)	87	What diction the writing contains (e.g., buzzwords, pronouns, symbols, strange/common words).
	Type of Info (κ=0.87)	121	What topics the writing contains (career, education, experiences, personal life).

Table 7. **Profile Reasoning Codebook: Secondary Codes** — We show the coded and related descriptions for our secondary codes.

D PROFILE DEMOGRAPHICS

Race	Black	white	NXBW	**Total**
Gender:*				
Woman	28	29	25	**82**
Man	25	27	26	**78**
Age:				
18-29	22	21	25	**82**
30-49	29	23	12	**29**
50-59	1	11	2	**14**
70+	1	1	0	**2**
Prefer not to say	0	0	0	**0**
Highest Education:				
High School or Less High School	3	2	5	**10**
Some College / 2yr Degree	17	6	2	**25**
Bachelor's/Post-Grad	33	48	44	**125**
Prefer not to say	0	0	0	**0**
Total	53	56	51	**160**

Table 8. **Profile Demographics**— We present the self-reported demographics of the valid profiles used as experimental stimuli for the moderation study.*No participants identified as transgender.

E EXPLORATORY MODELING OF MODERATION EXPERIENCE

Factor	Likelihood Ratio χ^2	P-value
Primary Factor		
Profile Identity (PI)	**48.552**	**<0.001**
Moderator Identity (MI)	**12.649**	**<0.001**
Profile Content (PC)	**9.386**	**0.009**
Prior Moderation Experience	2.392	0.122
Two-way Interaction		
PI : PC	**16.680**	**0.011**
MI : PC	**6.643**	**0.036**
PI : MI	2.189	0.534
Three-way Interaction		
PI : MG : VP	9.218	0.162

Table 9. **Factors' & Experience Significance on Perceived Artificiality** – Via an analysis of variance, we find that our extended model that includes "Prior Moderation Experience" does not result in any new significant effects, but still retains all the significant primary effects and two-way interactions of the original model (Table 3). Rows that denote significant relations are bolded.