DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks

Jaron Mink, Licheng Luo, and Natã M. Barbosa, *University of Illinois at Urbana-Champaign;* Olivia Figueira, *Santa Clara University;* Yang Wang and Gang Wang, *University of Illinois at Urbana-Champaign*

https://www.usenix.org/conference/usenixsecurity22/presentation/mink

This paper is included in the Proceedings of the 31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

978-1-939133-31-1

DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks

Jaron Mink[*], Licheng Luo[*], Natã M. Barbosa[*], Olivia Figueira[†], Yang Wang[*], Gang Wang[*]
[*]University of Illinois at Urbana-Champaign [†]Santa Clara University
{jaronmm2, ll6, natamb2}@illinois.edu, ofigueira@alumni.scu.edu, {yvw, gangw}@illinois.edu

Abstract

Fabricated media from deep learning models, or *deepfakes*, have been recently applied to facilitate social engineering efforts by constructing a trusted social persona. While existing works are primarily focused on deepfake detection, little is done to understand how users perceive and interact with deepfake persona (e.g., profiles) in a social engineering context. In this paper, we conduct a user study ($n = 286$) to quantitatively evaluate how deepfake artifacts affect the perceived trustworthiness of a social media profile and the profile's likelihood to connect with users. Our study investigates artifacts isolated within a single media field (images or text) as well as mismatched relations between multiple fields. We also evaluate whether user prompting (or training) benefits users in this process. We find that artifacts and prompting significantly decrease the trustworthiness and request acceptance of deepfake profiles. Even so, users still appear vulnerable with 43% of them connecting to a deepfake profile under the best-case conditions. Through qualitative data, we find numerous reasons why this task is challenging for users, such as the difficulty of distinguishing text artifacts from honest mistakes and the social pressures entailed in the connection decisions. We conclude by discussing the implications of our results for content moderators, social media platforms, and future defenses.

1 Introduction

The recent progress of deep learning models has significantly improved our ability to synthesize media content such as image, video, audio, and text [21,34,50,93]. This leads to a rising concern that "deepfake" techniques can be used to generate abusive content to manipulate public opinions [11]. More importantly, deepfakes can also be used in *social engineering* attacks [66] where attackers construct a trusted persona (e.g., a profile in online social networks) to interact with victims. Traditionally, fake profiles are constructed with hard-coded templates and stock photos [6, 35, 81], which are easier to defend against using existing techniques (e.g., reverse image search [86] and similarity-based methods [89]). Deepfake techniques, however, have the potential to circumvent these methods by generating diverse and unique text and images to build credible personas, and truly scale-up the deception efforts (without sacrificing quality).

Deepfake-based social engineering is not only a theoretically possible threat but is also starting to be applied in practice. In 2019, a deepfake LinkedIn profile was found to have successfully infiltrated Washington's political circle, connecting with government officials including the former Deputy Director of the President [13,77]. In the same year, a deepfake voice clone was used to scam a CEO of an energy firm out of 250K U.S. dollars [17]. In addition to these targeted attacks, recent investigations also revealed that deepfake profiles were actively used in state-sponsored campaigns [5, 37, 71, 72].

Prior research has explored the detection methods of artificially generated content [8, 74, 87, 91, 94] with a focus on a single media type (i.e., either image or text) by searching for algorithm-generated artifacts (e.g., unnatural facial features). These artifacts are useful for detection but can be mitigated by more advanced models; thus, this turns into a cat-and-mouse game [66]. More importantly, these works do not address the specific contexts of social engineering attack where human users are in the loop. Users view artificial profiles holistically to foster trust and *semantic inconsistencies* across different media modalities (e.g., text, image) can make a difference.

In this paper, we fill in the gaps by studying user perceptions towards deepfake-generated social persona. We seek to understand whether (and how) deepfake artifacts affect users' trust and their decisions to accept or ignore a connection request. We also explore whether priming or training affects user perceptions of deepfake profiles, which could inform intervention strategies within online social networks. Finally, we examine the common strategies users employ to assess profile trustworthiness. Unlike existing efforts that study deepfake content in isolation (e.g., text only) [23, 25], we seek to answer these questions within full social network profiles in a social engineering context.

We design a user study where users are instructed to review a series of social network profiles (in the context of LinkedIn). We control two key variables. First, we control *artifact conditions* where we introduce different types of deepfake artifacts into profile photos, biographies, and relational artifacts across profile fields. Second, we control *prompt levels* where we vary the information provided to participants before the study, ranging from not informing participants about the existence of deepfake profiles to training participants to look for artifacts. Upon each encounter of a profile, participants are instructed to rate the perceived trustworthiness of the profile, describe the reasons for their ratings, and decide whether they would accept or ignore a connection request from this profile. At the end, participants describe their overall strategies for profile assessment.

We collect results from $n = 286$ participants from Amazon Mechanical Turk. We discover that users are largely vulnerable to deception from deepfake profiles. Although generative artifacts such as grammar errors, image errors, and inconsistencies between fields result in significant decreases in trustworthiness and connection request acceptance, these reductions appear insufficient to protect users. Under conditions that result in the *lowest* average trust and acceptance rates, participants are still neutrally trusting of the profile and 43% of them accept the connection request. Under other artifact-laden conditions, participants are positively trusting with connection acceptance ranging from 56%–85%.

Additionally, by coding and analyzing the open-form responses from participants, we observe interesting behaviors as they assess profiles. To summarize a few:

First, searching for inconsistencies in a profile is already a strategy of some participants even without any prompting. However, participants are primarily searching for inconsistencies to assess an individual's honesty rather than looking for signs of algorithm-generated profiles. We also observe that *unprompted* participants often focus their attention on the wrong profile sections (i.e., sections where deepfake algorithms are least likely to introduce artifacts).

Second, participants have expressed different degrees of difficulty in assessing different media types. For example, after some training, users can easily attribute the artifact in photos as a clear indication of deepfake profiles. Conversely, even after training, users are less certain about text artifacts (such as incoherent writings) due to alternative explanations such as poor writing/communication skills of the profile owners. Text artifacts afford more plausible deniability.

Third, we also discover that prompting (or training) participants to detect these artifacts may incite unintentional behaviors, causing false accusations against otherwise authentic profiles or relying on stereotypes to make judgments.

Based on our findings, we end the paper by discussing implications for social network platforms, social network moderators, and future defenses.

2 Related Works and Research Questions

2.1 Related Work

Deepfakes for Abuse. Deepfakes (combining the words "deep learning" and "fake") [66] generally refer to synthetic media generated by deep learning models such as autoencoders [9, 52] and generative adversarial networks (GANs) [36]. Deepfake models have been used for abusive purposes such as modifying (or swapping) human faces in images/videos [55, 69], generating fake social media comments [29] or online reviews [47, 93], and synthesizing human voices to impersonate target users [34, 73].

Deepfake Artifacts and Detection. While the quality of synthetic content is improving, deepfake models still produce artifacts, including human perceivable artifacts (e.g., unnatural hairs and accessories in face images) and invisible ones (e.g., statistical patterns introduced by generative models). The detection of deepfake content has turned into a "cat-and-mouse" game. While researchers constantly develop techniques to detect artifacts in deepfake images/videos [7, 74, 87, 91] and text [8, 94], future models are subsequently proposed to remove such artifacts [66] or fool deepfake detectors using adversarial examples [44].

Understanding User Perception. Recent works have examined users' ability to distinguish human-created media from machine-generated content, including text [23, 25], images/videos [39, 53], and audio [68, 70, 80]. However, these studies only focus on a single media modality (e.g., only text or only image); they do not investigate how users may utilize multiple modalities within a user profile in a social engineering context. In addition, they do not investigate the varying effects of training/prompting on user reactions.

Related works also study the perceived trustworthiness of user profiles online [24, 28, 61, 62]. Most focus on *real profiles* and study the impact of topic choices [61], linguistic styles [62], and profile images [28] on the profiles' perceived trustworthiness. A recent work [45] also uses *real profiles* from Airbnb (text portion only) but primes users by telling them that some profiles are AI-generated. They find that such priming can significantly decrease users' trust towards these real profiles. Different from existing studies, we use profiles generated by deepfake models (including both images and text) to holistically examine their impact on user trust.

2.2 Our Motivations

Why Study Deepfake Profiles. Our study is motivated by the following considerations.

First, deepfake profiles, when used for social engineering, have a natural advantage against existing defenses. Prior works (including those from LinkedIn [89]) show that existing Sybil (fake) profiles are usually generated using hard-coded templates and stock photos [35, 81, 86], which can be easily spotted (by human labelers [86]). Stock images can be detected via Google's reverse image search [6] and text templates can cause high similarities among Sybil profiles [89]. Deepfakes can circumvent these methods as they generate original text/images without reusing similar text or stock photos.

Second, deepfake profiles have the potential to effectively deceive users. While comparing deepfake profiles with existing Sybils is not a focus of our study, we believe such a comparison is useful. As such, in Appendix B, we present a secondary user study we conducted which shows that deepfake profiles (even with artifacts) have more success in gaining users' trust than real-world Sybil profiles.

Third, deepfake profiles have started to be used in real-world social engineering attacks and deception campaigns. For instance, a deepfake profile has successfully infiltrated Washington's political circle, connecting with politicians and government officials [13, 77]. In other examples [37, 71], investigators find that deepfake profiles were used in Russian-operated campaigns that target U.S. users. While this is not a typical "targeted" social-engineering attack, it represents a dedicated attempt for deception and opinion manipulation.

Our Research Questions. We seek to understand user perception of deepfake profiles and their affect upon social engineering attacks in online social networks. We ask the following research questions:

RQ1 Do deepfake artifacts or priming (training) influence the perceived trustworthiness of a profile?

RQ2 Do deepfake artifacts or priming (training) influence user decisions on accepting a connection request[1]?

RQ3 What common strategies do users employ to assess the trustworthiness of a profile?

3 Methodology

We conduct an online user study where participants are instructed to examine a series of social network profiles. In this experiment, we consider 5 profile conditions (within-subjects) where we either present a consistent profile or a profile with one of four deepfake artifacts (i.e., *artifact variable*). Participants are also exposed to one of three different prompt levels (between-subjects) where they receive different information about deepfake profiles and artifacts (i.e., *prompt variable*). In the following section, we describe our experimental scenario, our profile generation method, and our user study procedure[2].

3.1 Experimental Scenario

We design our study around the LinkedIn social network for two reasons. First, fake LinkedIn profiles are commonly used in social engineering [4] and deepfakes are found to facilitate such efforts [13, 77]. Second, LinkedIn users expect unsolicited connection requests (e.g., from recruiters).

We develop a role-playing scenario for the study, which is a commonly used method to study phishing susceptibility [54, 63, 79, 88]. To construct our scenario, we reference real-world phishing incidents which happened during a company merger or acquisition [30, 42, 56], a time when employees expect (unsolicited) communications from people in different organizations and third-party services (e.g., law firms) [30]. Based on real-world incidents, we use the following scenario:

> *You are a project manager at the fictional startup company "Pear Co", which has recently undergone a merger with another company, "Bird Inc". You have received a number of connection requests on LinkedIn with profiles that say they are currently working at Bird Inc. This is your first time meeting these users.*

We choose to employ role-playing as it allows us to put participants in a scenario where the profiles they encounter are relevant to their persona (i.e., potential colleagues during a company merger). In order to study social engineering attacks without role-playing, we would need to create profiles that cater to *each individual participant*, which is infeasible.

3.2 Constructing Social Network Profiles

For our study, we methodically construct LinkedIn profiles using deepfake models. As shown in Figure 1, a LinkedIn profile contains several important fields such as a name, a profile image, a biography/summary (i.e., "About"), a job history (i.e., "Experience"), and an educational history (i.e., "Education"). To generate profiles, we first decide on a set of professions, and then train deepfake models to populate the profiles with generated images and text.

Selecting Professions. We select the profiles' professions such that the profession is relevant to the "company merger" scenario and is at an appropriate position to interact with the participant's role. Following the occupational groups defined by U.S. Bureau of Labor [2], we select three professions from three different departments: a Human Resource Manager ("HR"), a Database Administrator ("IT"), and a Billing Manager ("Fin").

Generating Deepfake Text and Images. Our goal is not necessarily to advance the state-of-the-art for deepfake generation, but rather to control the "artifacts" in the generated profiles to study user perception.

[1] We are interested in connection requests because they are usually the first step in conducting phishing and can help an attacker make lateral movements within an organization. For instance, an attacker may establish several connections with company employees before reaching a CEO.

[2] Due to space, we provide supplementary materials online [65].

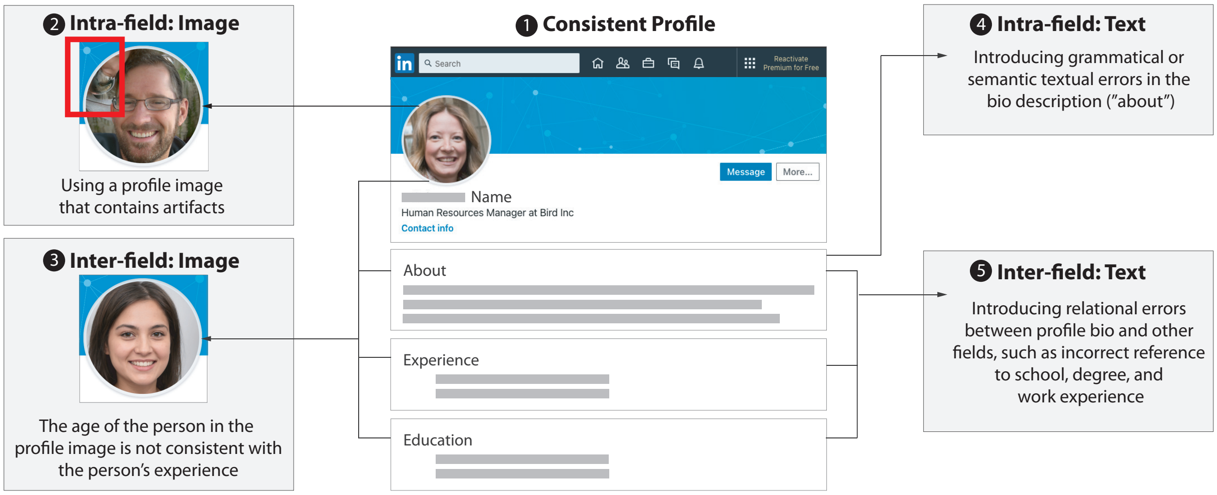

❷ Intra-field: Image
Using a profile image that contains artifacts
❸ Inter-field: Image
The age of the person in the profile image is not consistent with the person's experience
❶ Consistent Profile
in
Search
Reactivate Premium for Free
Message
More...
Name
Human Resources Manager at Bird Inc
Contact info
About
Experience
Education
❹ Intra-field: Text
Introducing grammatical or semantic textual errors in the bio description ("about")
❺ Inter-field: Text
Introducing relational errors between profile bio and other fields, such as incorrect reference to school, degree, and work experience

Figure 1: **Profile Artifact Conditions** – We have 5 conditions: 1 consistent profile condition and 4 conditions with profile artifacts.

First, to generate profile images, we use a popular generative model, StyleGAN2 [50][3]. We train the model with a public human-face dataset [49] using the recommended configurations.

Second, we generate the "About" section, which is free-form text. Although there are pre-trained text generation models, most of them are too generic to be generate a LinkedIn summary. Instead, we take the GPT-2 model [75] and perform fine-tuning[4] using the public resume database from Might Recruiter [1]. We take a standard 12-layer GPT-2 as the base and fine-tune it for 20 additional epochs. For each of the three professions, we gather 1,200 resumes where each resume includes a bio-summary, an educational history, and a job history. Using the summary text, we train a customized GPT-2 model *for each profession*. We further apply nucleus sampling ($p = 0.9$) to improve text fluency and coherence [43].

Third, we use gender-neutral names for the profiles to avoid causing inconsistencies with the profile images. We construct a pool of 7 gender-neutral English names [83], including "Alex", "Sam", "Lee", "Chris", "Terry", "Pat", and "Robin".

Finally, the rest of the data fields (e.g., educational history and job history) can be constructed by sampling the corresponding entities from Might Recruiter resume dataset.

3.3 Controlling Profile Artifacts

To study how different artifacts affect user trust, we explicitly control the artifacts in generated profiles. As shown in Figure 1, we have five profile conditions, including one consistent profile and four profiles with deepfake artifacts. We consider inconsistencies that commonly arise from the use of deepfake models. We investigate two *intra-field inconsistencies* that are the result of isolated mistakes within the free-form text summary and the profile image, and two *inter-field inconsistencies* that are the result of semantic differences between a generated data field and another field in the profile.

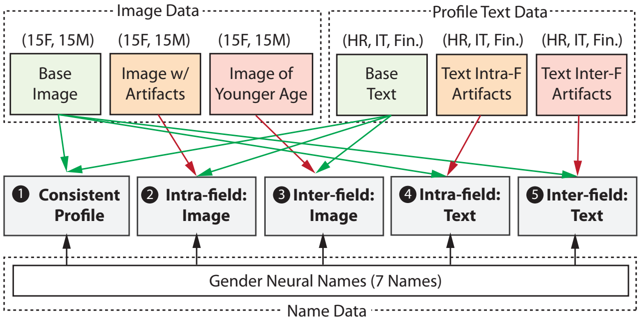


Figure 2: **Profile Construction Method** – To construct a profile, an image is randomly selected from one of the three image pools (each image pool has 15 female photos and 15 male photos), a gender-neutral name is random selected from a pool of 7 names, and a profile text template is selected from three text pools (each text pool has 3 templates, one template per profession: HR, IT, and Fin).

The Building Blocks. Figure 2 shows the basic building blocks used to construct profiles under each condition. At the high-level, we have a dataset of profile images, including three image pools (each pool has 15 female and 15 male images). Then we have a dataset of profile text templates, including three template pools (each pool has three templates for HR, IT, and Fin, respectively). Each text template contains an "About" summary and the experience and education sections. Finally, we have a set of gender-neutral names. Each time we construct a profile, we randomly select one image from a

[3] We choose StyleGAN2 (for image generation) and GPT2 (for text generation) because (1) these models are publicly available (meaning they are also available to attackers) and (2) these models have been used to generate abusive deepfakes in practice [3, 37, 72].

[4] We did not use the GPT-3 model [14] as the API was not yet open to the public (GPT-3 is also expensive to retrain). We manually confirm that the text artifacts used in our study still exist in GPT-3 outputs.

specified image pool, one text template, and one random name for the profile. In the following, we introduce each dataset and condition in detail[5].

❶ Consistent Profile. Consistent profiles are constructed *without using any deepfake content*. As shown in Figure 2, a consistent profile is constructed with the *base text* and *base image* pools that contain real photos and real resume text. More specifically, base images are real photos randomly selected from the training dataset of human faces [50], including 15 male and 15 female images. We ensure these images are (1) professional, (2) smiling, and (3) between the ages of late 20s to early 40s. Note that (1) is informed by common profile photos used in LinkedIn, (2) is informed by a recent study that shows smiling photos could improve perceived trust [28], and (3) is to make sure the person's age roughly matches the job experience. To account for potential bias towards any specific image, a random image from this base image pool will be selected to generate a "consistent profile" during our study.

For the text, we construct the "About" summary, education history, and work experience based on *real resumes* in the Might Recruiter dataset. Given a profession, we randomly select 3 resumes and manually synthesize them into a single profile (to protect the resume owner's privacy). Even though the resume data is publicly available, we did not use the original resumes directly. We ensure that no semantic inconsistencies are introduced, i.e., the summary always correctly references the items in the experience/education sections. We also replace their current company name with *Bird Inc* to match our study's scenario (Section 3.1). The summary is between 80-120 words in length [61]. The work experience is between 6-12 years to match with the age of the profile images.

Finally, as mentioned before, a gender-neutral name is randomly selected for the profile. A screenshot of an example profile page is in Figure 8 in the Appendix.

In this study, we construct "consistent profiles" as the baseline as they do not contain any deepfake image/text. We did not directly use real LinkedIn users' profiles because it would require crawling a representative set of user profiles, which is against LinkedIn's Terms of Service [59]. In our experiment (Section 4.2), we show that the acceptance rate of friend requests from our consistent profiles is comparable with those of real users reported in prior works[6].

❷ Intra-field: Image Artifacts. These profiles are constructed similarly with consistent profiles with one key difference—we use deepfake images generated by StyleGAN that contain artifacts. Figure 3 shows some example artifacts produced by StyleGAN such as malformed faces, distorted/asymmetric accessories, and blurry backgrounds. We select 30 generated images (15 female, 15 male) with noticeable artifacts to represent a "worse-case" scenario for attackers. These images follow the same selection criteria as those for the *base images* (i.e., professional, smiling, and with an age from late 20s to early 40s).

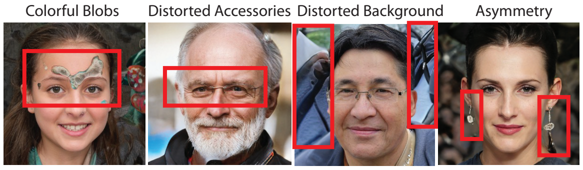


Figure 3: **Image Artifacts Examples** – We show example images produced by StyleGAN2 that contain visual artifacts.

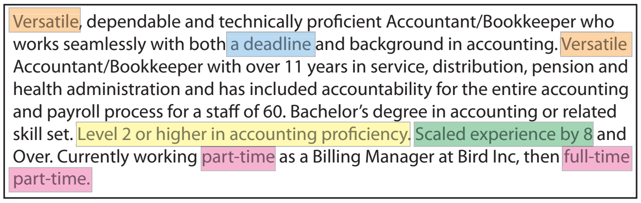

Versatile, dependable and technically proficient Accountant/Bookkeeper who works seamlessly with both a deadline and background in accounting. Versatile Accountant/Bookkeeper with over 11 years in service, distribution, pension and health administration and has included accountability for the entire accounting and payroll process for a staff of 60. Bachelor's degree in accounting or related skill set. Level 2 or higher in accounting proficiency. Scaled experience by 8 and Over. Currently working part-time as a Billing Manager at Bird Inc, then full-time part-time.

Figure 4: **Text Artifacts Example** – The example is picked from a text template used in our user study. We highlight the text that is incohesive, repetitive, or out of context.

❸ Inter-field: Image Artifacts. This condition introduces an inter-field inconsistency between the age shown in a profile image and the age implied by a profile's work experience. The intuition is that the age of a user is often correlated to their educational and work experience (e.g., it is rare for a 19-year-old to be a bar-certified lawyer). We select 30 generated images by StyleGAN (15 female, 15 male) that appear to be in their late teens or early 20s and thus unlikely to hold the positions described in our base text. We also ensure these images do not have perceptible artifacts. The rest of the selection criteria remain the same with those of the base images.

❹ Intra-field: Text Artifacts. This condition introduces artifacts to the "About" section. Instead of using text from real resumes, we use GPT-2 generated text that contains common grammatical and semantic errors (e.g., repetition of words/phrases, and incoherent/conflicting train of thought). Figure 4 shows an example summary used in our study.

❺ Inter-field: Text Artifacts. This condition introduces an inter-field inconsistency between the "About" section and the job/education history sections. These errors include incorrect references to educational degrees they received (or companies they worked at) and skills that do not match with their stated profession. To create such summary text, we feed our GPT-2 model with the original, consistent text up to a key reference (e.g., to a job) and use GPT-2 to generate new text to replace the original reference; GPT2 is then fed again with the newly generated text until we reach another reference. The process continues until we reach the end of the original summary text. This method ensures that the overall structure remains

[5] Profile screenshots can be found in the supplementary material [65].

[6] Our consistent profiles have an acceptance rate of 90% (see Figure 6), which is within range of the acceptance rate of cloned profiles of real people (60%–90%) [10] and is comparable to that of real user profiles (79%) [92].

similar to the original, consistent text while the references are replaced with authentic GPT-2 generated text. We also correct any grammar errors in the generated text.

3.4 Prompting Users

To understand the impact of user prompting/training, we divide participants into *three* groups (between-subjects). We expect the results from different prompt groups can help inform intervention strategies for online social networks.

- **No Prompt:** We do not mention that the profiles could be algorithm generated nor ask users to look for potential artifacts. The tutorial page only briefly introduces the background of LinkedIn and its profiles.
- **Soft Prompt:** We additionally inform participants during the tutorial that some profiles may be fake, i.e., generated by Artificial Intelligence (AI). However, we do not provide specific information on what fake profiles look like or how to detect them.
- **Hard Prompt:** We inform participants that some profiles may be fake and include a detailed tutorial to describe deepfake techniques and common artifacts in generated images and text.

3.5 Experimental Design

Our study controls two variables: (1) prompt level (between-subjects) and (2) artifact condition (within-subjects). First, participants are divided into three prompt groups. Then, each participant is instructed to view *three* profiles and answer a set of questions. The three profiles include one consistent profile (❶), one profile with intra-field artifacts (randomly choosing intra-field image ❷ or text ❹), and one profile with inter-field artifacts (randomly choosing inter-field image ❸ or text ❺). To make sure participants do not view any repeated profile elements, we have ensured that the three profiles (1) cover all three professions (HR, IT, and Fin), and (2) do not use the same profile image or name. We also randomize the order of the three profiles to avoid biases.

Study Procedure. First, the participant reads and signs the consent form and reads a tutorial page. The tutorial page varies based on the prompt group (see Section 3.4). Then, the participant enters the *main task* to play the role of a project manager in the company merger scenario (see Section 3.1). The participant examines three LinkedIn profiles and answers questions under each profile. Finally, we ask *follow-up questions* and collect their demographic information. The questions used in our study can be found in Appendix A.

Definition of Trust. A key component of our study is to assess trust. Theoretical literature on trust [46] typically defines trust as the relationship between two parties that denotes one party's (the "Trustor") willingness to take a risk with another party (the "Trustee") given that they cannot control the actions of the other party (e.g., loaning a book to a friend with the expectation that they will return it). Characteristics of both the trustee and the trustor are considered to impact the overall trust relationship; our study considers both.

The perceived "trustworthiness" of the Trustee (i.e., a profile) is the focus of study. We follow Mayer et al.'s integrated definition [64] which postulates that trustworthiness is a combination of three factors that the Trustor perceives in the Trustee: The Trustee's adherence to a set of principles (Integrity), the Trustee's skill and competence (Ability), and the Trustee's intent to do good to the Trustor (Benevolence). We measure these qualities when a participant encounters a profile in the *main task* of our study.

The characteristic of a Trustor (i.e., the participants in our study) is typically defined as the general propensity to trust others (or one's "generalized trust"). In other words, different people may be more or less trusting of others. We measure this participant-specific quality within the *follow-up questions*.

Main Task Questions. Upon viewing each profile, the participant will answer five questions (see Appendix A.1).

Three of these questions directly measure the participants perception of integrity, ability, and benevolence in the presented profile, which collectively measure the perceived "trustworthiness" of the profile. Like prior works [15, 45, 61], we use three scenario-specific statements to directly map to each quality and ask participants to state their agreement on a scale from [0–100]: "*This profile is an accurate depiction of the user* " [Integrity; **Q1**]; "*This user is knowledgeable in their role as a [Database Administrator / Human Resource Manager / Billing Manager]*" [Ability; **Q2**]; "*This person will make newcomers feel welcome*" [Benevolence; **Q3**]. Then, we ask participants for an open-form response regarding the profile-specific reasons behind their trust ratings [**Q4**], and a binary response on whether they would accept or ignore the connection request from this profile [**Q5**].

Follow-up Questions. After the main tasks, we ask a series of follow up questions (see Appendix A.2), including their general strategies to assess profiles [**Q6**], and their "generalized trust" [**Q7**–**Q9**] (inspired by [45]). We also collect users' background knowledge in the relevant professions, social network habits, and demographic information.

Attention Check and Action Tracking. To make sure the obtained results are reliable, we design two attention questions. On each profile page, we implement JavaScript code to record the amount of time a user hovers over notable sections on the page and their interaction events with the profile (e.g., image expansion). The results will be used to infer user actions and their center of attention in later analysis.

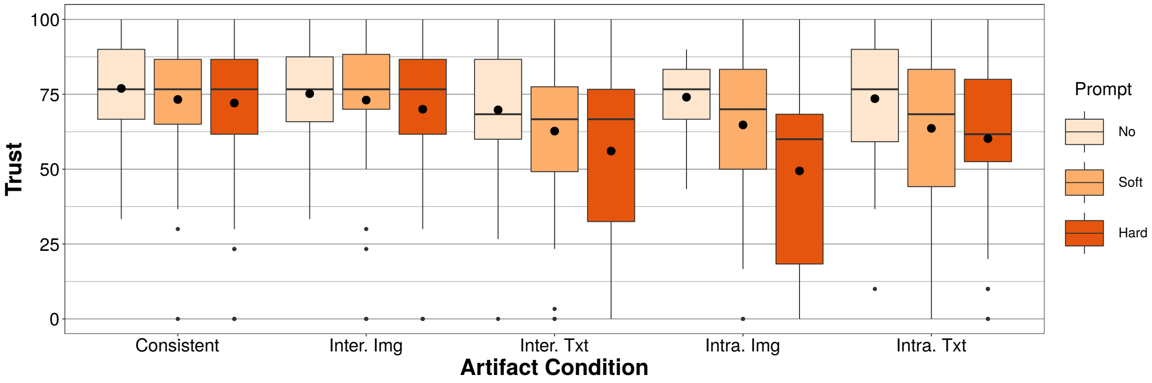


Figure 5: **Profile Trust Score** – We show the box plot of profile trust scores (based on **Q1**, **Q2**, **Q3**) under different prompt and artifact conditions. The line (and dot) inside the box represents the median (and mean) trust score, the box hinges represent the first and third quartiles, and the whiskers extend up to 1.5 times the inter-quartile range.

3.6 Recruitment and Ethics

We recruited participants from Amazon Mechanical Turk (MTurk) between March and April of 2021[7]. We did not collect any personal identifiable information (PII) from the participants. After the study, we debriefed the participants with the full details of the study. We also offered the participants the opportunity to withdraw their data after debriefing (we did not receive such requests). To get high-quality MTurk workers, we recruited U.S. workers who have completed at least 100 tasks successfully with an approval rate greater than 95%. Participants were excluded if they failed one of two attention checks and then were further verified to be dishonest[8]. In total, we had $n = 286$ qualified participants.

Participants ranged in age between 20-70+ with a median of 35-39 years old; 60% identify as male, 39% as female, and 1% as non-binary. The survey took a median of 12.95 minutes to complete, and each participant was compensated $1.75 for their time. Full demographic information of participants is presented in the supplementary material [65].

4 Impacts on Trust and Request Acceptance

We first analyze how deepfake artifacts and prompting conditions affect the perceived trustworthiness of a profile (**RQ1**) and the likelihood of friend request acceptance (**RQ2**).

In addition to the primary variables, we also analyze other potentially influential factors. We expect an individual's generalized trust (their propensity to trust others) to affect the perceived trustworthiness of a profile and thus the response to the friend request. Additionally, we evaluate how the demographic information of the individual may influence trust. We consider age and gender given they are reported to have an impact in prior phishing-related studies [20, 58, 79]. For brevity, the analysis of other factors such as education background, professional experience, and social media experience is presented in the supplementary material [65].

4.1 Perceived Profile Trustworthiness

For each profile, participants rate their trustworthiness by assessing the integrity [**Q1**], ability [**Q2**], and benevolence [**Q3**] of the profile. While these factors are theoretically orthogonal [64], we find a strong positive pair-wise correlation between these measured factors ($R >= 0.71$, $p < 0.001$ for all correlations). Similar to past work [15, 45, 61], we aggregate these factors by defining a "trust score" as the mean of these variables to simplify the analysis. The trust score has a value range of $[0, 100]$. This score associates values less than 50 with a general distrust towards the profile, greater than 50 with a general trust, and equal to 50 with a lack of opinion.

Figure 5 shows a box plot of the trust score. General trends can be observed in the *descriptive statistics*. For example, "Hard"-prompted profiles are generally perceived as the least trustworthy ($\bar{x} = 63.3$, $s = 27.3$)[9] followed by "Soft" ($\bar{x} = 68.4$, $s = 22.8$) and "No" prompt ($\bar{x} = 74.4$, $s = 17.4$).

Surprisingly, in nearly all cases of prompting and artifact conditions, the average trust score of profiles is relatively positive (i.e., over the value of 50), denoting that most participants are trusting of presented profiles.

Modeling and Results. To quantify these effects, we use *linear mixed-effects regression* (or LMER) modeling. Regression is a common tool for statistical significance tests on a set of explanatory variables where we can isolate the effect of one variable while keeping other variables constant [67]. Unlike traditional linear regression, LMER can model random effects, allowing for non-independence between measured outcomes. Given the repeated-measured design of our study, this modeling is most appropriate.

[7] Our study was reviewed and approved by the IRB prior to recruitment.

[8] We manually inspect the open-form responses and exclude those with irrelevant single-word answers, answers copied from online sources, and duplicated answers across multiple participants.

[9] $\bar{x}$ denotes sample mean; s denotes sample standard deviation.

Variable	Estimate (β)	Std. Err.	p-value
Intercept	75.118	3.005	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	5.187	2.019	0.011*
Hard Prompt	-4.447	2.024	0.029*
Artifact (Reference = Consistent)			
Inter Image	-2.018	2.093	0.335
Inter Text	-10.700	2.072	<0.001***
Intra Image	-12.334	2.079	<0.001***
Intra Text	-7.207	2.089	<0.001***
Gender (Reference = Female)			
Male	-2.317	1.703	0.175
Non-Binary	-8.664	8.287	0.297
Age	0.027	0.436	0.951
Generalized Trust	0.388	0.053	<0.001***

Table 1: **Trust Rating Analysis** – Linear mixed-effects regression model. The unit for estimate and standard error is the aggregated trust score. “Age” (in units of 5 years) and “Generalized Trust” are numeric and thus do not have a reference group. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

We model the profile artifact, prompting condition, generalized trust, and demographics[10] of a participant as *fixed effects* upon the measured trust score. Additionally, to account for the fact that each participant responds to multiple profiles and thus introduces non-independence between ratings, we model each participant as a *random effect* upon the trust score.

The model is summarized in Table 1. Estimates (β) are the regression coefficients which represent the mean change in the output variable given a one-unit shift in the input variable while holding other input variables constant. A positive estimate indicates a positive correlation. Standard errors are represented in units of the aggregated trust score (ranging from 0 to 100), The p value represents the likelihood that the observed differences are caused by chance. The difference is considered statistically significant when $p < 0.05$.

We find that all prompt levels have significant effects on the trust score. The trust score under No prompt is significantly higher in comparison to that under Soft prompt ($\beta = 5.187$, $p < 0.05$). This indicates that simply making users aware that some profiles are algorithm generated (without revealing the characteristics of the fake profiles) will reduce users' trust. We also find a significant decrease of trust when comparing Soft prompt to Hard prompt ($\beta = -4.447$, $p < 0.05$). The result implies that user training will further reduce trust by showing what deepfake content and their artifacts look like.

To evaluate whether prompting causes a “spillover effect” and significantly affects the trust of *consistent profiles*, we run an ANOVA test. Over all consistent profiles, no statistically significant differences in trust are found with respect to the prompt conditions of No ($\bar{x} = 76.8$, $s = 16.0$), Soft ($\bar{x} = 73.3$, $s = 18.0$), and Hard ($\bar{x} = 72.1$, $s = 21.6$) as determined by a one-way ANOVA: $F(2,281) = 1.595$; $p = 0.205$. In other words, while we find evidence that prompting decreases the trustworthiness over all profile conditions (Table 1), we do not find evidence that prompting significantly affects the trustworthiness of *consistent* profiles.

Observation 1: User prompting can decrease participants' trust towards the profiles overall. However, we do not find evidence that such an effect is significant on *consistent* profiles.

A number of artifact conditions significantly decrease the perceived trustworthiness of a profile, including intra-field image artifacts ($\beta = -12.334$, $p < 0.001$), intra-field text artifacts ($\beta = -7.207$, $p < 0.001$), and inter-field text artifacts ($\beta = -10.700$, $p < 0.001$). We do not find evidence that inter-field image artifacts (where a photo appears younger than the age implied in the work experience) significantly affect the trust score. The results suggest that a range of artifacts should be carefully handled by the attackers if they want to generate deepfake profiles to gain users' trust.

Finally, we find that self-reported generalized trust[11] is positively associated with the instance-specific trust ratings ($\beta = 0.388, p < 0.001$), which is consistent with prior works [64]. We find no evidence of either participant gender or age significantly affecting the resulting trust score.

Observation 2: Intra-field artifacts (in image and text) and inter-field text artifacts can decrease participants' trust towards a profile.

4.2 Profile Acceptance

After viewing and rating a profile, we ask participants whether they would *accept* or *ignore* the connection request sent by this profile [**Q5**].

Figure 6 shows the acceptance rate broken down by prompt level and artifact condition. For No-prompt condition, the consistent profiles have an acceptance rate of 90%; the acceptance rates of deepfake profiles (regardless the artifact conditions) are also fairly high (79%–85%). However, for Soft- and Hard-prompt conditions, the acceptance rate seems to be lower for certain artifact conditions. In particular, profiles with intra-field image, intra-field text, and inter-field text artifacts under Soft and Hard prompt, appear to be the least accepted by participants with acceptance rates between 43%–71%.

Modeling and Results. To determine the factors that significantly affect profile acceptance, we use *mixed-effects logistic*

[10] The two participants that did not disclose their demographics were excluded from all modeling analysis.

[11] Similar to the measured trust score, the generalized trust score is a mean aggregation of the three generalized trust factors [**Q7**-**Q9**].

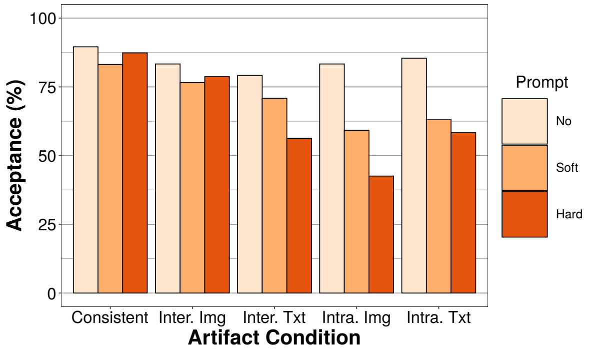


Figure 6: **Request Acceptance Rate** – Profiles' request acceptance rates (based on Q5) under different prompt/artifact conditions.

Variable	Estimate (β)	Std. Err.	p-value
Intercept	1.702	0.350	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	0.761	0.228	<0.001***
Hard Prompt	-0.163	0.201	0.418
Artifact (Reference = Consistent)			
Inter Image	-0.570	0.282	0.043*
Inter Text	-1.142	0.263	<0.001***
Intra Image	-1.525	0.264	<0.001***
Intra Text	-1.092	0.264	<0.001***
Gender (Reference = Female)			
Male	-0.026	0.180	0.887
Non-Binary	0.516	0.940	0.583
Age	0.025	0.047	0.589
Generalized Trust	0.023	0.006	<0.001***

Table 2: **Request Acceptance Analysis** – Logistic mixed effects regression model. The unit for estimate and standard error is log odds scaled. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

regression. Logistic regression is appropriate because the outcome variable is *binary* — in our case, either profile acceptance (1) or rejection (0). Under this model, we include the profile artifact, prompting condition, generalized trust, and demographics of a participant as *fixed effects* and set each participant as a *random effect* upon the acceptance outcome.

Table 2 shows the estimates and standard error in log odds scaled units to represent the probability of acceptance. We find that participants with the No prompt are associated with significantly higher log odds of profile acceptance compared to participants with the Soft prompt ($\beta = 0.761$, $p < 0.001$). However, we do not find evidence that the Hard prompt further significantly decreases the likelihood of acceptance compared to the Soft prompt ($\beta = -0.163$, $p = 0.418$). This implies that just making users aware that deepfake profiles exist can already provide some degree of protection.

Observation 3: Knowledge of the existence of deepfake profiles can already decrease participants' likelihood to accept connection requests.

We find that all artifact conditions led to a significant reduction in the likelihood of request acceptance. Interestingly, inter-field ($\beta = -0.570$, $p < 0.05$) and intra-field ($\beta = -1.525$, $p < 0.001$) *image* artifacts show the smallest and largest decrease on profile acceptance, respectively. Both the inter-field ($\beta = -1.142$, $p < 0.001$) and intra-field ($\beta = -1.092$, $p < 0.001$) *text* artifacts have a much closer effect size that lies between the range of the image artifacts.

Like before, generalized trust of a participant is found to be a significant factor that affects a participant's likelihood to accept the connection request ($\beta = 0.023$, $p < 0.001$). We find no evidence to suggest that either participant gender or age significantly affects profile acceptance.

Observation 4: Both intra-field artifacts and inter-field artifacts, alongside prompting, can decrease the likelihood of a participant accepting a friend request.

4.3 Comparison of Different Artifacts

So far, we have compared the impact of different artifacts against the "consistent" profiles. Next, we evaluate whether these artifacts are significantly different among one another. To determine differences between treatments, we run a post-hoc comparison of the estimated marginal means produced by the models. We use the Bonferroni correction [41] because it allows us to make multiple pairwise comparisons of treatments without inflating false positives. Due to space limit, we only summarize our main observations, and the detailed analysis results are presented in the supplementary material [65].

Overall, we do not observe a clear pattern, except that inter-image artifacts (i.e., age inconsistencies between the profile image and the rest of the profile) have a lower impact on trust and acceptance than other artifacts. A possible explanation is that it is already common for people to use a photo of younger appearance on LinkedIn for impression management or reduce age discrimination during job seeking [31, 51], and thus, age inconsistencies have a lower impact on users' trust.

5 Reasoning and Strategies of Participants

In this section, we focus on the last research question (**RQ3**) and analyze participants' open-text responses to understand the reasons for the trust ratings of individual profiles [**Q4**] and their general strategies to assess profiles [**Q6**]. Open responses are analyzed via thematic analysis [12]. For both questions, a primary coder is assigned to code the collected responses and develop a codebook. A secondary coder then independently codes 15% of previously coded responses to determine inter-coder agreement via Cohens-κ. When high agreement ($\kappa > .7$) is not reached, both coders meet up to discuss disagreements, make codebook improvements, and use the newly developed codebook to re-code each analyzed response. This process is repeated until high agreement is met. The resulting codebooks and the counts of each code are provided in the supplementary material [65]. In total, we code all $n = 858$ responses for **Q4** ($\kappa = 0.87$) and $n = 286$ for **Q6** ($\kappa = 0.72$) in our analysis[12]. Since we observe similar results under Soft- and Hard-prompt conditions, we simply refer them as "prompted" condition if not otherwise stated.

5.1 Profile Assessment Reasoning

Under each profile, we ask participants to point out aspects of the current profile that strongly influenced their rating [**Q4**]. $n = 858$ responses are thematically coded based on whether the participants referred to each user-interface (UI) element, whether they mention an inconsistency, and their overall feeling towards a profile.

Areas of Focus. Combining all conditions, most responses (n=478; 56%) mention the "Experience" section as an important consideration. Participants tend to use experience to judge the professional capabilities of a user, e.g., "*the 'Experience' section as it indicated not only that [the user] has more than 8 years of experience in human resources, but that the majority of that time has been spent in his current position as the HR manager at Bird Inc*" (P108).

The second most mentioned UI is the "About" section (n=304; 35%). Participants often use the "About" section to assess professional and personal qualities, e.g., "*well written description [not only] highlights [the users'] experience but her as a person*" (P115).

In comparison, fewer responses mention the profile images (n=153; 18%), "Education" section (n=139; 16%), the name (n=6; 1%), or connection number (n=1; <1%).

The overall results suggest that participants tend to put more attention to the "Experience" section than the "About" and profile image sections. Note that the experience section is highly structured. One possible explanation is that the structured list can be used as heuristics or mental shortcut, which reduces the cognitive load. However, this also leads to an important implication. Well-structured sections such as "Experience" are easier to generate using algorithms without introducing artifacts. In comparison, free-format text and images are harder to generate, and yet these sections do not draw as much user attention. The result is further supported by the mouse tracking data (Table 3).

Comparing prompted and non-prompted groups, we observe prompting seems to slightly change participants' area of focus. For example, more participant responses mention profile images and the "About" sections under the Hard prompt than the No prompt. This observation is also supported by the mouse tracking data (Table 3).

Metrics		No	Soft	Hard
Profile Image (expand)		12%	17%	40%
Experience (expand)		36%	28%	30%
About (hover over)	Mean	4,841	4,486	7,341
	Median	611	331	989
Experience (hover over)	Mean	7,760	6,646	6,584
	Median	3,376	1,872	1,912
Education (hover over)	Mean	1,564	1,957	1,695
	Median	125	133	128

Table 3: **Mouse Tracking Results** – The top two rows report the percentage of participants that clicked on the UIs to expand profile images and the experience items. The bottom three rows report the mouse hover-over time (median and mean) in each UI section for "About", "Experience", and "Education" in milliseconds. After the Hard prompt, users are more likely to look for artifacts in profile images and the "About" text.

Observation 5: Participants pay more attention to the well-structured lists in the "Experience" section of a profile where generative algorithms are less likely to leave artifacts.

Artifacts Noticeability. To see how often artifacts are correctly noticed by participants, we focus on different artifact conditions (over all prompt levels) to analyze their responses.

First, intra-field text artifacts (e.g., grammar errors and repetitions) are noted in some responses (n=19; 13%). However, for both unprompted and promoted participants, they often attribute the artifacts to a lack of communication/writing skills. One unprompted participant mentions: "*his About section is honest, but it contains extraneous information that make him seem less interested in putting out a career focused LinkedIn profile*" (P189). One prompted participant notes: "*the description is repetitive and doesn't seem all that intelligent or knowledgeable*" (P59).

Intra-field image artifacts are also noted in some responses ($n = 30$; 21%). For unprompted participants, while these artifacts cause confusion, they do not appear to greatly affect user decision. For example, "*The profile looks excellent but I am wondering about her photo what's going on with that?*" (P43) and "*Again, it's the photo that is bothering me. Are they wearing a hat? I cannot tell*" (P156). After prompting, participants who saw such artifacts often immediately identify them as signs of an artificial profile, e.g., "*the picture is very off - this looks like an AI-generated picture to me*" (P215) and "*the hair is very weird and seems AI like*" (P51).

We observe that the text and image artifacts (intra-field) appear to have different effects. Even after the Hard prompt, par-

[12] We have $n = 286$ participants and thus receive 286 responses for **Q6**. Each participant views 3 profiles and answers **Q4** for each profile, which produces 858 responses for **Q4**.

ticipants often have difficulties telling whether the observed text artifacts are the result of deepfake algorithms or of poor, but genuine, writing/communication skills. In contrast, after prompting, it is easier for the participants to attribute image artifacts because there are no other plausible explanations.

Observation 6: After prompting, participants can attribute intra-field image artifacts to fake profiles. However, they often find alternative explanations for intra-field text artifacts (e.g., poor writing/communication skills of the profile owner). These alternative explanations make it difficult to attribute intra-field text artifacts to fake profiles.

Inter-field text artifacts are also noted in some responses (n=27; 19%). Without prompts, participants express concerns but do not consider them as signs of fake profiles, e.g., "*I was worried that his background (legal) did not seem to match the jobs he's held. I am worried about his commitment*" (P37). After prompting, users are better at attributing such inter-field inconsistencies. For example, "*She is told about she was entrepreneur and she working different projects. i guess that this resume was fake*" (P62) and "*None of this persons about section was consistent with their role at Bird Inc. It appears to be completely fabricated or AI*" (P237).

Lastly, inter-field image artifacts are hardly noticed (echoing our findings in Section 4). Across all conditions, only one participant explicitly notes this inconsistency: "*The individual looks very young in his profile image. Doesn't really align with what I was expecting after reading their bio and looking over their experience and education*" (P93).

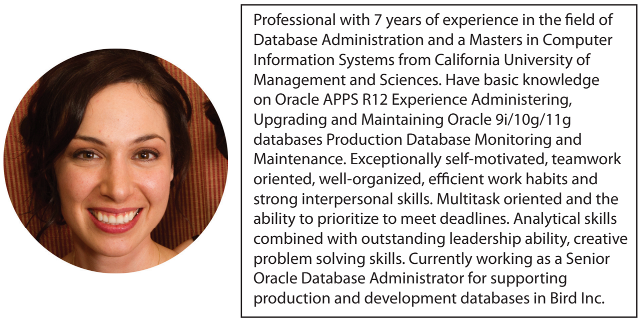
Observation 7: While participants have difficulties attributing *intra*-field text artifacts to fake profiles, they can attribute relational *inter*-field text artifacts after prompting.

5.2 Unexpected Perceptions of Artifacts

While prompting did not result in a significant difference in trust for "consistent" profiles (*Observation 1*), we still observe that prompted users occasionally perceive qualities within real images/text as signs of algorithm-generated artifacts [**Q4**].

Perception of Non-Existent Artifacts in Images. For profiles that only used *real* human images ($n = 572$), a few participants' responses (n=14; 2%) mention that AI-generated artifacts in the image affect their ratings.

When shown the image in Figure 7a), one Hard-prompted participant misinterprets the shoulder of another off-image person as algorithm-generated artifact: "*There is something wrong with the applicant's photo...the detail of one of her shoulders is impossible*" (P224). When shown the same image, another Hard-prompted participant notes a small glint on a tooth (likely a dental filling) and becomes suspicious: "*There is a little glitch in the corner of her mouth that makes me wonder if an AI made the photo"* (P111). Similar comments appear under other real images as well.



Professional with 7 years of experience in the field of Database Administration and a Masters in Computer Information Systems from California University of Management and Sciences. Have basic knowledge on Oracle APPS R12 Experience Administering, Upgrading and Maintaining Oracle 9i/10g/11g databases Production Database Monitoring and Maintenance. Exceptionally self-motivated, teamwork oriented, well-organized, efficient work habits and strong interpersonal skills. Multitask oriented and the ability to prioritize to meet deadlines. Analytical skills combined with outstanding leadership ability, creative problem solving skills. Currently working as a Senior Oracle Database Administrator for supporting production and development databases in Bird Inc.

(a) Profile Image **(b) "About" Text**

Figure 7: **Unexpected Artifacts noted in Consistent Profiles** – Examples of real image/text recognized as generated ones by some participants.

Interestingly, some perceived artifacts stemmed from racial and gender expectations of those in the images. When shown one image that depicted a woman of African descent, several participants perceive a disagreement between the name and the demographics of the subject. Two Soft-prompted participants note that "*the picture shows a Black woman but the name seems to belong to a White man"* (P68) (the name was Chris), and "*the name also doesn't suit the person in the image"* (P141) (the name was Alex). When shown a white man with the name "Lee", one Hard-prompted participant notes that "*the name is [of] Asian [descent], but the pic doesn't match"* (P202).

The result suggests that after prompting, users may overcompensate in their effort of looking for algorithm-generated artifacts and potentially begin to make judgments based on stereotypes. While the intention is likely benign, such behaviors could lead to disproportionate distrust towards real people who defy the stereotypes held by others.

Perception of Non-Existent Artifacts in Text. For profiles that were free of any major grammar errors or semantic inconsistencies ($n = 570$), some prompted participants (n=17; 3%) explicitly mention that AI-induced text artifacts impacted their ratings; this never happened under the No prompt.

For example, Figure 7 shows the "About" text of a consistent database administrator profile. Even so, some Hard-prompted participants believe that there are "*grammatical errors in his bio"* (P222) and "*there are some typos"* (P98).

Observation 8: Prompted participants sometimes overcompensate in the process of searching for deepfake artifacts, which leads to mistakes such as perceiving artifacts that do not exist or making judgments based on stereotypes.

5.3 Profile Assessment Strategies

At the end of the study, participants are asked to describe their profile assessment strategies [**Q6**]. $n = 286$ responses are thematically coded according to the referenced UI sections, personal qualities inferred from the profile, inconsistencies noted, and the reasons for their ultimate decision.

Noted UIs. A majority of participants (*n*=218; 76%) mention specific UI sections in their response such as "Experience" (*n*=159; 56%), "About" (*n*=127; 44%), profile images (*n*=86; 30%), and "Education" (*n*=74; 26%). This result is consistent with the *instance-specific* decisions noted in Section 5.1 and gives more credence to *Observation 5* which notes that the well-structured experience section catches more user attention than free-form text and images. This behavior gives deepfake models room to make mistakes in text/image generation.

Search for Personal Qualities. Another common strategy mentioned is to examine the personal qualities of the individual in the profile (*n*=79; 28%). Most of these participants (*n*=64/79) focus on the profile's aptitude or ability, e.g., "*their abilities, qualifications and achievements played major roles in my assessment"* (P73). Some attempt to infer the underlying personality (*n*=28/79) to see if they were confident, professional, or easygoing. To do so, one participant uses profile text to note "*the tone of how they wrote about themselves"* (P164) while another uses their profile image to make this deduction: "*by looking at the picture of them you could see if they were friendly or if they showed signs of stress on their faces"* (P191).

Interestingly, the prompt level appears to affect this strategy. While commonly used among No-prompt participants (*n*=42; 44%), searching for personal traits is less frequently observed among Soft-prompted (*n*=22; 23%) and Hard-prompted (*n*=15; 16%) participants.

Search for Inconsistencies. Another common strategy is to search for inconsistencies in profiles (*n*=89; 31%). For example, when searching for intra-field text inconsistencies, some participants note specific qualities such as repetitions, grammar mistakes, and contradictions. Others rely on their own interpretation of fake profiles. One participant believes fake profiles were more likely to use generic descriptions, e.g., "*I also trusted the profile that spoke 'like a person' and not just a generic description*" (P194).

Again, the prompt level appears to affect the use of this strategy with respect to frequency and purpose. In terms of frequency, this strategy is more commonly used among Hard-prompted (*n*=41; 43%) and Soft-prompted participants (*n*=33; 35%) than unprompted participants (*n*=15; 16%). In terms of purpose, prompted participants typically look at a profile "*for 'red flags' that it could be an AI created profile"* (P221); interestingly, unprompted participants are not necessarily searching for signs of a fake profile, but rather to determine "*how honest they were"* (P266) and whether "*they are willing to 'skillfully' stretch the truth"* (P187). For this reason, under the unprompted condition, participants mention more about searching for disagreements *between* fields rather than focusing on grammar errors or picture anomalies.

Observation 9:
Prompting affect participants' profile assessment strategy. These warnings promote strategies that look for intra-field artifacts and demote strategies that infer the personal qualities of the individual in the profile.

Reasons for Actions. While no participants explicitly note why they would decline a request, several provided reasons why they would be inclined to accept and even hesitant to decline a request (*n*=10; 4%). One reason mentioned is the potential benefits a good connection could provide, e.g., "*how [the user's] experience may help me to better communicate and work with the Bird Inc team"* (P169).

Another reason is the lack of perceived risks associated with the action: "*I feel like there would be nothing to lose from accepting any and all invitations from Bird Inc. employees"* (P41).

Lastly, a group of participants noted that they feel *obliged* to accept such requests from fellow employees: "*These were all future coworkers, as they are all Bird employees, so I feel it's incumbent upon me to accept their connection requests*" (P21) and "*Since they're co-workers, even ones I haven't met to date, I feel predisposed to accepting their requests by default*" (P171). Some even note how not accepting a request could harm them: "*If they had a senior position then I would almost surely accept... It would be bad form to decline someone as that would only hurt and not help me*" (P184).

These results help to illuminate why professional social networks are attractive for social engineering attacks. When faced with a connection request, users not only weigh the opportunity cost of ignoring a valuable connection, but also the potential negative risks such actions may cause. Furthermore, when faced with these decisions, some participants are not aware of how a malicious user could negatively impact them.

Observation 10: Participants make decisions by weighing the benefits and costs of a connection. The trustworthiness of a profile is only one of many impacting factors in accepting or ignoring a connection request.

6 Discussion

In this section, we discuss countermeasures against deepfakes, platform moderation, and user intervention strategies. Then, we discuss the limitations of our study and future works.

6.1 Countermeasures for Deepfake Profiles

A key takeaway from our study is that average users (unprompted) are overly trusting of deepfake profiles (Figure 5) and are highly likely to accept their connection requests (with an acceptance rate of 79%–85%). To this end, we argue that individual users should *not* be the front line of defense. Instead, automated detection methods and community-based moderation can play a bigger role.

Deepfake Profile Detection. Existing deepfake detection methods often only focus on a specific media type [66]; there is an opportunity to build detectors based on inter-field inconsistencies across media modalities. From the attackers' perspective, fully addressing inter-field inconsistencies using generative models alone is challenging. It would require them to develop more advanced knowledge representation methods to handle heterogeneous media types and even tackle the open challenges in common-sense reasoning [19, 57]. These challenges for attackers mean opportunities for defenders.

In addition, detection methods can inspect *other metadata* beyond profile content. For example, by looking into IP addresses, account registration information, social connections, and on-site activities [18, 33, 85, 90], detection methods may detect abusive accounts even if the profiles appear authentic. Meanwhile, intra-field deepfake image/text detection methods [7, 87, 94] can be used to analyze profile data to produce useful features. These defenses are relatively orthogonal and can be applied jointly for a more comprehensive defense.

Community-based Moderation. Community-based moderation plays a large role in platforms such as Twitter and Facebook to combat misinformation [26, 27, 82]. However, our results point out nuanced challenges when using this idea to deal with deepfake profiles. For instance, while we show that *some users* are able to reason about profile consistency at higher-level semantics (*Observation 7*) and that inconsistencies actively affect users' trust and actions (*Observations 2* and *4*), average users are likely to fail at detection. This echoes the result from a prior work [32]. In particular, we show that it is difficult for average users to disentangle honest mistakes from generative errors when dealing with texts (*Observation 6*), a major component of most platforms' content. To enable effective moderation, one possible direction is for social media platforms to identify capable users and appoint them as community moderators (e.g., the Reddit model). This can support expert-led crowdsourcing moderation. Recent work shows that knowledge moderators can effectively guide the crowd to conduct investigation tasks [84].

Empirical Measurements of Deepfake Profiles. Finally, developing effective countermeasures requires understanding how deepfake profiles are (and will be) used *in practice*. So far, empirical investigations [5, 13, 37, 71, 72, 77] suggest that real-world deepfake profiles are not (yet) used for large-scale attacks but are often used for targeted purposes. This means detection methods that focus on clustering large groups of similar accounts/profiles (e.g., [89]) are likely to be ineffective. However, these investigations are focused on specific campaigns and may not be representative. More systematic measurements on deepfake profiles are needed in future works.

6.2 Intervention Strategies for Users

Our study shows a *mixed result* with respect to potential intervention strategies for social media users. *For deepfake profiles*, we show that even a Soft prompt (i.e., informing the existence of deepfake profiles) reduces users' trust towards such profiles (*Observations 1* and *3*) and encourages users to focus on distinguishing features of deepfakes (*Observation 9*).

For legitimate profiles, however, the impact of prompting is inconclusive. In Section 4, we find no *statistically significant* evidence that prompting affects users' trust in consistent profiles (*Observation 1*), but our *qualitative* results in Section 5.2 show multiple cases where participants discredit legitimate users based on their fixation with artifact detection after reading deepfake tutorials (*Observation 8*). These mistakes could be due to the priming effect of the tutorials or a belief in stereotypes (e.g., expecting certain profile names given the inferred race and gender from profile images) which may disproportionately affect certain user populations. As such, we *do not* currently recommend social media platforms adopt user-oriented training or warning. Instead, platforms should focus on improving automated defenses and moderation strategies to detect and remove deepfake profiles before they can reach lay users.

6.3 Limitations

Our methodology still has a few limitations. *First*, recruiting participants from MTurk may lead to certain biases in user demographics [22] and privacy attitudes [48, 78]; additionally, we cannot guarantee participant attention (even with attention checks). These are inherent limitations of using MTurk. Nevertheless, MTurk responses have been shown to generalize well. Recent studies show that the reported security behaviors from MTurk generalize as well as census-representative panels [76], and MTurk workers are at least as attentive as subject pool participants [40]. *Second*, the role-playing methodology might not always reflect real-world behaviors. To the best of our ability, we follow existing guidelines [38] designed to improve the generalizability of results using role playing methods. The alternative (e.g., real-world phishing experiments), however, is challenging to execute given ethical considerations. *Third*, our study is conducted around LinkedIn profiles. It is possible that certain findings may not generalize to other social networks such as Twitter and Facebook. Future work is needed to extend the analysis to other social media platforms. *Fourth*, our study focuses on the perceived profile trustworthiness and likelihood to accept connection requests, which are related to the initial stages of social engineering. Future works may look into the effectiveness of using deepfakes to facilitate further actions (e.g., extracting sensitive information from target users). *Finally*, our study does not consider other profile factors such as the number of mutual connections or location of residence. Future work may look into comparing the effects and interactions of these factors.

7 Conclusion

In this paper, we quantitatively evaluate how deepfake artifacts impact the perceived trustworthiness of a social network profile and users' willingness to connect with it. We also explore the effects of user prompting. We find evidence that various artifacts and all prompts significantly decrease the trustworthiness and acceptance rate of profiles; however, users still appear to largely fall victim to deception. We also qualitatively analyze users' reasoning and strategies when assessing a profile. We find users typically focus on well-structured sections instead of free-form areas (that are likely to have artifacts) and have difficulty in distinguishing text artifacts from honest mistakes. Prompting participants may help to recognize certain artifacts (e.g., those in images), but can lead to negative effect such as discrediting authentic profiles. Overall, the results suggest the need for future research to study defense mechanisms to protect users from deepfake profiles.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and suggestions. This work was supported in part by NSF grants 2030521 and 2055233, and the Graduate Research Fellowship Program under Grant No. DGE-1746047.

References

[1] MightyRecruiter. https://www.mightyrecruiter.com/.

[2] Occupational Outlook Handbook. U.S. Bureau of Labor Statistics. https://www.bls.gov/ooh/.

[3] Amazon employees, bots or trolls? New Twitter bios emerge to speak out mostly against union effort. GeekWire, 2021. https://www.geekwire.com/2021/amazon-employees-bots-trolls-new-twitter-bios-emerge-speak-mostly-union-effort/.

[4] Hackers Spearphish Professionals on LinkedIn with Fake Job Offers, infecting them with malware, warns esentire. eSentire, 2021. https://www.esentire.com/security-advisories/hackers-spearphish-professionals-on-linkedin-with-fake-job-offers-infecting-them-with-malware-warns-esentire.

[5] Private industry notification (210310-001). FBI, 2021. https://www.ic3.gov/Media/News/2021/210310-2.pdf.

[6] Shalinda Adikari and Kaushik Dutta. Identifying fake profiles in linkedin. In *Proc. of PACIS*, 2014.

[7] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting Deep-Fake Videos From Phoneme-Viseme Mismatches. In *Proc. of CVPR Workshops*, 2020.

[8] Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. Detecting Deceptive Reviews Using Generative Adversarial Networks. In *Proc. of IEEE SP Workshops*, 2018.

[9] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proc. of ICML Workshop on Unsupervised and Transfer Learning*, 2012.

[10] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. All your contacts are belong to us: Automated identity theft attacks on social networks. In *Proc. of WWW*, 2009.

[11] Dan Boneh, Andrew J. Grotto, Patrick McDaniel, and Nicolas Papernot. How Relevant is the Turing Test in the Age of Sophisbots? *CoRR*, 2019.

[12] Richard E. Boyatzis. *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE, April 1998.

[13] 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/.

[14] Tom Brown et al. Language models are few-shot learners. In *Proc. of NeurIPS*, 2020.

[15] Tommy Bruzzese, Irena Gao, Griffin Dietz, Christina Ding, and Alyssa Romanos. Effect of Confidence Indicators on Trust in AI-Generated Profiles. In *Proc. of CHI EA*, 2020.

[16] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Proc. of NSDI*, 2012.

[17] Jesse Damiani. DA Voice Deepfake Was Used To Scam A CEO Out Of $243,000. MalwareBytes, 2019. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

[18] George Danezis and Prateek Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *Proc. of NDSS*, 2009.

[19] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 2015.

[20] Daniela A. S. de Oliveira et al. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proc. of CHI*, 2017.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

[22] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proc. of WSDM*, 2018.

[23] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text. In *Proc. of EMNLP*, 2020.

[24] Chad Edwards, Autumn Edwards, Patric R. Spence, and Ashleigh K. Shelton. Is that a bot running the social media feed? testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. *Computers in Human Behavior*, 2014.

[25] Richard M. Everett, Jason R. C. Nurse, and Arnau Erola. The anatomy of online deception: what makes automated text convincing? In *Proc. of SAC*, 2016.

[26] Facebook. Facebook community standards on manipulated media, 2021. https://www.facebook.com/communitystandards/manipulated_media.

[27] Facebook. Fact-checking on facebook, 2021. https://www.facebook.com/business/help/2593586717571940.

[28] Asle Fagerstrøm, Sanchit Pawar, Valdimar Sigurdsson, Gordon R. Foxall, and Mirella Yani-de Soriano. That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on Airbnb™. *Computers in Human Behavior*, 2017.

[29] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: about detecting deepfake tweets. *CoRR*, 2020.

[30] FireEye. Fireeye m-trends: Hidden phishing risks during mergers and acquisitions. https://content.fireeye.com/m-trends/rpt-m-trends-2019, 2020.

[31] Krings Franciska, Gioaba Irina, Kaufmann Michele, Sczesny Sabine, and Zebrowitz Leslie. Older and younger job seekers' impression management on linkedin: Similar strategies, different outcomes. *Journal of Personnel Psychology*, 2020.

[32] David Mandell Freeman. Can you spot the fakes?: On the limitations of user feedback in online social networks. In *Proc. of WWW*, 2017.

[33] Peng Gao, Binghui Wang, Neil Zhenqiang Gong, Sanjeev R. Kulkarni, Kurt Thomas, and Prateek Mittal. SYBILFUSE: combining local attributes with global structure to perform robust sybil detection. In *Proc. of CNS*, 2018.

[34] Yang Gao, Rita Singh, and Bhiksha Raj. Voice Impersonation Using Generative Adversarial Networks. In *Proc. of ICASSP*, 2018.

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

[36] Ian J. Goodfellow et al. Generative adversarial nets. In *Proc. of NeurIPS*, 2014.

[37] Graphika. Step into my parler. https://public-assets.graphika.com/reports/graphika_report_step_into_my_parler.pdf, 2020.

[38] Jerald Greenberg and Don E Eskew. The role of role playing in organizational research. *Journal of management*, 1993.

[39] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. In *Proc. of PNAS*, 2022.

[40] Schwarz N. Hauser DJ. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods.*, 2016.

[41] Nathanael Heckert and James Filliben. Nist/sematech e-handbook of statistical methods; bonferroni's method (7.4.7.3). https://www.itl.nist.gov/div898/handbook/prc/section4/prc473.htm, 2003.

[42] Michael Hill. Mergers and acquisitions put orgs at greater risk of attack. https://www.infosecurity-magazine.com/news/mergers-and-acquisitions-put-orgs/, 2016.

[43] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proc. of ICLR*, 2020.

[44] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *Proc. of WACV*, 2021.

[45] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proc. of CHI*, 2019.

[46] Cynthia Johnson-George and Walter C Swap. Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of personality and social psychology*, 1982.

[47] Mika Juuti, Bo Sun, Tatsuya Mori, and N. Asokan. Stay On-Topic: Generating Context-Specific Fake Restaurant Reviews. In *Proc. of ESORICS*, 2018.

[48] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of mechanical turk workers and the U.S. public. In *Proc. of SOUPS*, 2014.

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

[50] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. of CVPR*, 2020.

[51] Michèle C. Kaufmann, Franciska Krings, Leslie A. Zebrowitz, and Sabine Sczesny. Age bias in selection decisions: The role of facial appearance and fitness impressions. *Front. Psychol.*, 2017.

[52] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 2020.

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

[54] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *Proc. of eCrime*, 2007.

[55] Mu Li, Wangmeng Zuo, and David Zhang. Deep Identity-aware Transfer of Facial Attributes. *CoRR*, 2018.

[56] Raymond Lim. M&as put your company at risk for bec losses and data breach liability, 2019. https://www.agari.com/email-security-blog/mergers-acquisitions-losses-liability/.

[57] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Proc. of EMNLP*, 2020.

[58] Tian Lin et al. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Trans. Comput. Hum. Interact.*, 2019.

[59] LinkedIn. User agreement. https://www.linkedin.com/legal/user-agreement, 2020.

[60] LinkedIn. Professional community policies. https://www.linkedin.com/legal/professional-community-policies, 2021.

[61] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proc. of CSCW*, 2017.

[62] Xiao Ma, T. Neeraj, and M. Naaman. A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles. In *Proc. of ICWSM*, 2017.

[63] Claudio Marforio, Ramya Jayaram Masti, Claudio Soriente, Kari Kostiainen, and Srdjan Čapkun. Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In *Proc. of CHI*, 2016.

[64] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 1995.

[65] Jaron Mink et al. Supplementary materials. https://github.com/JaronMink/DeepPhish, 2022.

[66] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 2021.

[67] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. Wiley, 2001.

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

[69] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. RSGAN: Face Swapping and Editing using Face and Hair Representation in Latent Spaces. In *Proc. of SIGGRAPH*, 2018.

[70] Ajaya Neupane, Nitesh Saxena, Leanne M Hirshfield, and Sarah E Bratt. The crux of voice (in) security: A brain study of speaker legitimacy detection. In *Proc of NDSS*, 2019.

[71] Ben Nimmo et al. Operationffs: Fake face swarm, 2019. https://public-assets.graphika.com/reports/graphika_report_operation_ffs_fake_face_storm.pdf.

[72] Ben Nimmo, Camille François, C Shawn Eib, and Léa Ronzaud. Facebook takes down small, recently created network linked to internet research agency, 2020. https://public-assets.graphika.com/reports/graphika_report_ira_again_unlucky_thirteen.pdf.

[73] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In *Proc. of ICLR*, 2018.

[74] Jiameng Pu, Neal Mangaokar, Bolun Wang, Chandan K. Reddy, and Bimal Viswanath. Noisescope: Detecting deepfake images in a blind setting. In *Proc. of ACSAC*, 2020.

[75] Alec Radford et al. Language models are unsupervised multitask learners. *Technical report, OpenAI blog*, 2019.

[76] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *Proc. of IEEE SP*, 2019.

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

[78] Sebastian Schnorf, Aaron Sedley, Martin Ortlieb, and Allison Woodruff. A comparison of six sample providers regarding online privacy benchmarks. In *Proc. of SOUPS*, 2014.

[79] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proc. of CHI*, 2010.

[80] Maliheh Shirvanian and Nitesh Saxena. Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones. In *Proc. of CCS*, 2014.

[81] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *Proc. of USENIX Security*, 2013.

[82] Twitter. Birdwatch on twitter guide, 2021. https://twitter.github.io/birdwatch/.

[83] David D. Van Fleet and Leanne Atwater. Gender Neutral Names: Don't Be So Sure! *Sex Roles*, 1997.

[84] Sukrit Venkatagiri, Aakash Gautam, and Kurt Luther. Crowdsolve: Managing tensions in an expert-led crowdsourced investigation. In *Proc. of CSCW*, 2021.

[85] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y. Zhao. You are how you click: Clickstream analysis for sybil detection. In *Proc. of USENIX Security*, 2013.

[86] Gang Wang, Manish Mohanlal, Christo Wilson, Xiao Wang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. Social turing tests: Crowdsourcing sybil detection. In *Proc. of NDSS*, 2013.

[87] Run Wang, Felix Juefei-Xu, Jian Wang, Yihao Huang, Xiaofei Xie, Lei Ma, and Yang Liu. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In *Proc. of IJCAI*, 2020.

[88] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.hack: Engaging anti-phishing training through a role-playing phishing simulation game. In *Proc. of CHI*, 2019.

[89] Cao Xiao, David Mandell Freeman, and Theodore Hwa. Detecting clusters of fake accounts in online social networks. In *Proc. of AISec*, 2015.

[90] Teng Xu, Gerard Goossen, Huseyin Kerem Cevahir, Sara Khodeir, Yingyezhe Jin, Frank Li, Shawn Shan, Sagar Patel, David Freeman, and Paul Pearce. Deep entity classification: Abusive account detection for online social networks. In *Proc. of USENIX Security*, 2021.

[91] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *Proc. of ICASSP*, 2019.

[92] Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. Uncovering social network sybils in the wild. *ACM TKDD*, 2014.

[93] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. Automated Crowdturfing Attacks and Defenses in Online Review Systems. In *Proc. of CCS*, 2017.

[94] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending Against Neural Fake News. In *Proc. of NeurIPS*, 2017.

A Survey Questions

Here, we only show the most important questions, due to space limit. The complete list of survey questions is presented in the supplementary material [65].

A.1 Main Tasks.

Participants are instructed to review three profiles (Database Administrator, Human Resource Manager, and Billing Manager). Under each profile, the participants answer 5 questions ($5 \times 3 = 15$ questions in total).

Q1–Q3 Based on the profile, please rate to what extent you agree with each of these statements:

Q1 This profile is an accurate depiction of the user

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q2 This user is knowledgeable in their role as a [Database Administrator / Human Resource Manager / Billing Manager]
<Same scale as **Q1**>

Q3 This person will make newcomers feel welcome
<Same scale as **Q1**>

Q4 What aspect of this profile most influenced your ratings?

Q5 Based on the profile above, would you accept or ignore the connection request?

- Accept
- Ignore

A.2 Follow-up Questions

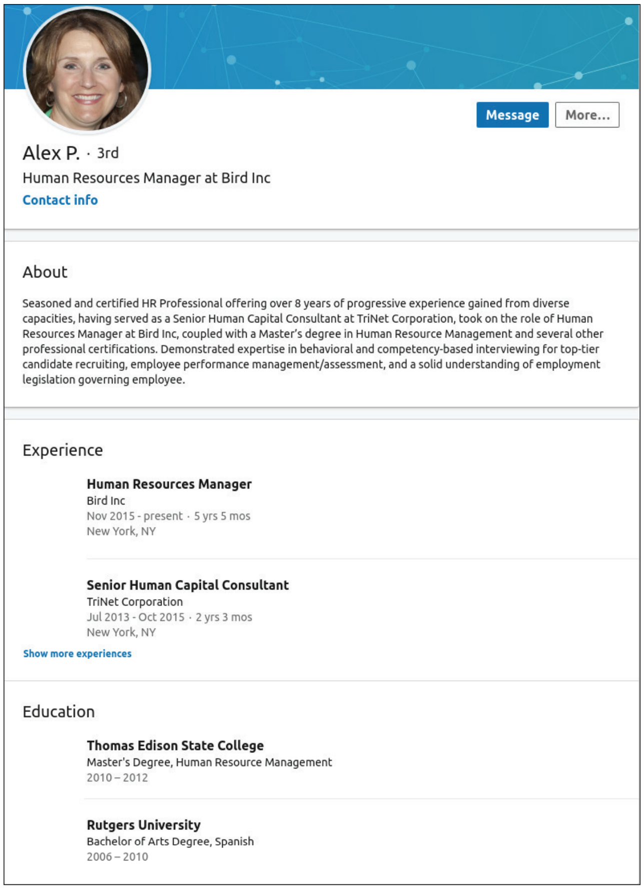
Q6 Please describe the strategies you used to assess profiles.

Q7-Q9 Please rate to what extend you agree with each of these statements:

Q7 In general, most people are honest
<Same scale as **Q1**>

Q8 In general, most people are qualified for their job
<Same scale as **Q1**>

Q9 In general, most people are good and kind
<Same scale as **Q1**>



Alex P. · 3rd
Human Resources Manager at Bird Inc
Contact info

About

Seasoned and certified HR Professional offering over 8 years of progressive experience gained from diverse capacities, having served as a Senior Human Capital Consultant at TriNet Corporation, took on the role of Human Resources Manager at Bird Inc, coupled with a Master's degree in Human Resource Management and several other professional certifications. Demonstrated expertise in behavioral and competency-based interviewing for top-tier candidate recruiting, employee performance management/assessment, and a solid understanding of employment legislation governing employee.

Experience

Human Resources Manager
Bird Inc
Nov 2015 - present · 5 yrs 5 mos
New York, NY

Senior Human Capital Consultant
TriNet Corporation
Jul 2013 - Oct 2015 · 2 yrs 3 mos
New York, NY

Show more experiences

Education

Thomas Edison State College
Master's Degree, Human Resource Management
2010 – 2012

Rutgers University
Bachelor of Arts Degree, Spanish
2006 – 2010

Figure 8: **Screenshot of a Profile** – This example shows a "consistent" profile for a Human Resources Manager used in study.

B Deepfake Profiles vs. Real-World Sybils

In our main study, all of the deepfake profiles are constructed by us so that we can better control the experimental conditions. A natural follow-up question is how they compare with real-world Sybil/fake profiles in terms of the ability to gain users' trust. In this section, we present an additional user experiment ($n = 101$) we conduct to make such comparisons.

For this user study, we manually collect a set of Sybil LinkedIn profiles. We then re-run the main user study (Section 3) under the "No-prompt" condition where we replace the *consistent profile* group with the newly collected Sybil profiles. In this way, we can compare users' trust and request acceptance of Sybil profiles with our deepfake profiles.

Sybil Profile Collection. Collecting real Sybil profiles is challenging. We are not aware of any public dataset of ground-truth Sybil profiles. In addition, LinkedIn's terms of use [59] forbids (large-scale) profile crawling. To collect a set of Sybil profiles, we take a similar approach of a prior work [6] – we obtain Sybil profiles reported in research papers, security blogs, and articles as well as perform a manual search and

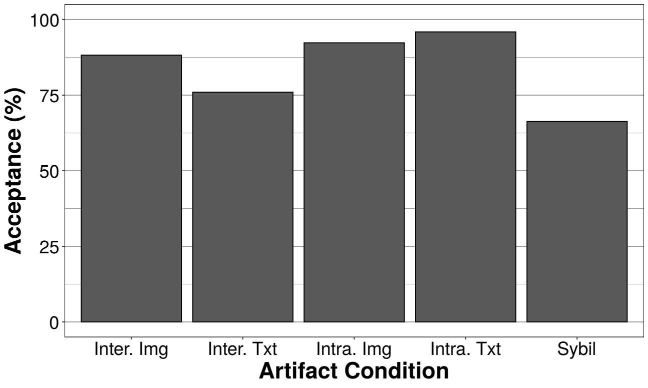


Figure 9: **Request Acceptance Rate** – We show the request acceptance rates for our evaluated artifacts and Sybil profiles.

inspection. Using this method, we obtain 30 LinkedIn Sybil profiles for this user study.

When collecting Sybil profiles, we manually validate their inauthenticity using LinkedIn's community policies as our guideline [60]. We consider a profile as a Sybil/fake if they meet at least one of two criteria. First, their profile photo does not represent themselves. This can be verified using Google reverse image search to find the sources of the photos (e.g., stock photos). Second, their profile contains a work history that is verifiably false (e.g., worked as a Dropbox sales executive before the company was founded). In addition to these determination criteria, other information can add to the confidence of labeling. For instance, since Sybils tend to connect with other Sybils [16], once we locate one fake profile, we can usually locate others using the "people also viewed" feature in LinkedIn. These Sybils often share highly similar headers, summaries, and job experience formatting.

We also have additional considerations when selecting Sybil profiles. First, we make sure they do not appear to use any deepfake images/text (to the best our ability)[13]. Second, we try to diversify the Sybil profiles — when we find a cluster of similar-looking Sybil profiles, we only include one of them in our user study set. Third, we format the Sybil profiles using the same style as other profiles used in our main study (similar to the one in Figure 8) to ensure formatting is not a factor that contributes to any differences we observe between deepfake and Sybil profiles[14].

User Study and Results. We re-run the main user study (Section 3) under the "No-prompt" condition. All configurations remain the same, *except we replaced the "consistent profile" with a random Sybil profile for each participant*. We collect data from $n = 101$ new participants via MTurk (participants of this study cannot participate in the main study and vice versa). We perform similar analyses as those in Section 4. The high-level takeaway is that in general *deepfake profiles are more successful than Sybils in gaining user trust and getting a connection request accepted.*

Variable	Estimate (β)	Std. Err.	p-value
Intercept	1.176	0.583	0.044*
Profile Type (Reference = Sybil)			
Inter Image	1.456	0.524	0.005**
Inter Text	0.449	0.420	0.285
Intra Image	1.976	0.615	0.001**
Intra Text	2.519	0.775	0.001**
Gender (Reference = Female)			
Male	0.240	0.344	0.485
Non-Binary	–	–	–
Age	-0.107	0.090	0.235
Generalized Trust	0.033	0.010	0.002**

Table 4: **Request Acceptance Analysis** – Logistic mixed effects regression model. The unit for estimate and standard error is log odds scaled. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

Figure 9 shows the acceptance rate. Sybil profiles yield a lower acceptance rate compared to different types of deepfake profiles. For example, deepfake profiles with intra-field text artifacts have the highest acceptance rate of 96%. The least accepted deepfake profiles (with inter-text artifacts) still have an acceptance rate of 76%. In comparison, Sybil profiles only have an acceptance rate of 66%.

To examine whether such observed differences are significant, we run the same statistical modeling as the main study[15]. As shown in Table 4, we now use "Sybil" as the reference group to compare with other deepfake profile groups. We find significant differences between three types of deepfake profiles and Sybil profiles. Compared to Sybil profiles, deepfake profiles with inter-field image artifacts, intra-field image artifacts and intra-text artifacts have significantly higher likelihood of acceptance. We find no evidence that inter-field text artifacts differ in this metric.

The trust score analysis returns similar conclusions and is omitted for brevity.

Recall that these deepfake profiles were intentionally constructed to include noticeable artifacts (see Section 3.3) and represent a worst-case scenario for attackers. In practice, attackers may further reduce these artifacts via post-processing and more careful configurations; however, even under this pessimistic condition, our results show that in general deepfakes are still more likely to gain users' trust and acceptance.

[13] We use reverse-image-search to confirm the sources of the profile photos and make sure the photos do not exhibit deepfake characteristics such as blurry backgrounds. We also check the profile text to make sure they do not contain any known deepfake artifacts.

[14] We change the profiles' *most recent company name* to "Bird Inc" (to fit within the role-playing scenario) and we omit profiles details that were unrelated to the research goals of our study (e.g., number of connections, skills). Studying the effect of such details can be future work.

[15] The only non-binary participant is removed to avoid model overfitting.