Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

# Does It Matter Who Said It? Exploring the Impact of Deepfake-Enabled Profiles On User Perception towards Disinformation

**Margie Ruffin**[1], **Haeseung Seo**[2], **Aiping Xiong**[2], **Gang Wang**[1]
[1] University of Illinois Urbana-Champaign
[2] Pennsylvania State University
mruffin2@illinois.edu, hxs378@psu.edu, axx29@psu.edu, gangw@illinois.edu

## Abstract

Recently, deepfake techniques have been adopted by real-world adversaries to fabricate believable personas (posing as experts or insiders) in disinformation campaigns to promote false narratives and deceive the public. In this paper, we investigate how fake personas influence the user perception of the disinformation shared by such accounts. Using Twitter as an exemplary platform, we conduct a user study (N=417) where participants read tweets of fake news with (and without) the presence of the tweet authors' profiles. Our study examines and compares three types of fake profiles: deepfake-enabled profiles, profiles of relevant organizations, and simple bot profiles. Our results highlight the significant impact of deepfake-enabled profiles and organization profiles on increasing the perceived information accuracy of and engagement with fake news. Moreover, deepfake-enabled profiles are rated as significantly more real than other profile types. Finally, we observe that users may like/reply/share a tweet even though they believe it was inaccurate (e.g., for fun or truth-seeking), which could further disseminate false information. We then discuss the implications of our findings and directions for future research.

## Introduction

Recently, abundant evidence from FBI and research groups shows that *deepfake profiles* are on the rise in social media platforms, engaging disinformation campaigns (Anderson 2019; Bond 2022; Krebs 2022; FBI 2021). Deepfake refers to deep learning models that can synthesize high-quality media content such as images, text, audio, and video (Mirsky and Lee 2021). Such synthetic content is now used to create fake social media profiles on platforms such as Twitter, Facebook, and LinkedIn (Anderson 2019). For example, deepfake-enabled profiles were created, posing as journalists or military personnel during the recent Ukraine-Russian War, to spread false narratives (Banerjea 2022). Similarly, deepfake-enabled profiles were also used in a series of campaigns aiming to manipulate the U.S. election (FBI 2021; Nimmo et al. 2019) and connect with U.S. government officials (Satter 2019).

There are key advantages for adversaries to use deepfake content to create fake personas. First, deepfake photos/text are highly unique, meaning they cannot be easily reverse-searched by image search engines such as Google and TinEye (Adikari and Dutta 2014) or detected by the conventional similarity-based detectors (Xiao, Freeman, and Hwa 2015). Second, it is easy to automatically customize the persona using deepfake algorithms with respect to the fake person's gender, age, ethnicity, and professional background. Third, the quality of deepfake-generated images/text can be high, making them hard to identify by humans (Nightingale and Farid 2022).

As deepfake profiles are becoming more prevalent, in this study, we seek to understand how deepfake can be used to create social media personas (that impersonate insiders/experts) and their impact on disinformation campaigns. We noted that recent works have started to investigate how humans perceive deepfake content (e.g., photos, text) (Fosco et al. 2022; Groh et al. 2022; Ternovski, Kalla, and Aronow 2021). A key difference is that prior works mainly study deepfake content *in isolation*—our study focuses on their impact on *downstream tasks* (i.e., disinformation campaign) where deepfake content (i.e., images) is carefully integrated into a social media persona which is then used to post false information. The most closely related work is a recent study from Mink et al. (Mink et al. 2022), which integrates deepfake photos and text into LinkedIn profiles to evaluate users' trust towards them. However, this study is still focusing on the *trust of the crafted profiles* themselves instead of investigating the downstream task of *disseminating false information using such profiles*.

To close these gaps, we design a study to examine whether and how a deepfake-enabled profile would change users' perceptions of a piece of disinformation posted by the profile. We focus on Twitter since it has been one of the most targeted platforms by disinformation campaigns and deepfake profiles (SSCI 2017). We have three research questions:

- **RQ1**: Do participants increase their perceived accuracy of tweets if deepfake profiles were also presented compared to showing the tweets only?
- **RQ2**: Do participants increase their engagement with the tweets if deepfake profiles were also presented compared to showing the tweets only?
- **RQ3**: Compared with other types of fake profiles, are deepfake profiles harder to detect by participants? What

are the primary factors that participants consider when assessing the profiles?

We answer these questions by conducting an online user study with N=417 participants recruited from the Prolific platform (Prolific 2022). During the study, we emulate the process of users reading a piece of information on a tweet, followed by checking the tweet author's profile. We manipulate two main variables (1) *the presence of a profile*, i.e., whether the profile is presented alongside the tweet, and (2) *the profile type*, under a within-subject design. Regarding the profile type, we construct three groups of fake profiles. The first group contains deepfake profiles posing as journalists. For comparison, the second group contains profiles of relevant organizations which are also commonly used in real-world disinformation campaigns (Wong 2020; Marr 2020). For example, given a piece of fake news on COVID-19, the profiles would take on the guise of a healthcare organization. The third comparison group contains simple bot profiles that do not contain a photo or detailed personal information (i.e., "simplefake"). Note that our goal here is not to compare deepfake content with non-deepfake content; instead, we seek to study this type of emerging deepfake-enabled profiles observed in practice, in comparison to other well-known types of fake profiles.

**Key Findings.** Our study leads to several important findings. *First*, we found that showing a deepfake-enabled or organization profile to users can significantly increase the perceived accuracy (and engagement) of the tweeted information, compared with showing the tweet alone. Simplefake, however, had the opposite effect. This indicates the effectiveness of using deepfake-enabled/organization profiles for disseminating disinformation. *Second*, interestingly, we observed that 12.9% of the participants decided to engage with a tweet (e.g., through "like", "comment") even though they rated the tweet as inaccurate. By analyzing their open responses, we found the users were engaging with the tweets in an attempt to verify or refute disinformation, to make fun of it, or to save/bookmark the tweet. However, such implicit sharing may further spread disinformation to other users. *Third*, deepfake-enabled profiles were perceived to be the most real, compared with other types of profiles (i.e., organization and simplefake profiles). Further analysis shows that the deepfake profile photos did not raise wide suspicion.

In summary, this study takes the first step towards understanding deepfake-enabled social personas and their impact on *downstream attacks*. The result shows that deepfake-enabled profiles pose a real threat in the context of disinformation dissemination. More work is needed to design and test effective countermeasures and intervention strategies.

## Background and Related Work

**Disinformation.** The term "disinformation" involves false information coupled with a deliberate intention to deceive an audience (Wu et al. 2019). Misinformation refers to information that is false or incorrect, including human errors (Wu et al. 2019). We will use *disinformation* (instead of misinformation ) in our study since we only examine fake news and fake personas with the intent to deceive the audience.

**Deepfake and Human Perception.** Deepfake is a combined word of "deep learning" and "fake" (Mirsky and Lee 2021), which refers to synthetic media (e.g., images (Wang et al. 2020), text (Fagni et al. 2020), and videos (Lyu 2020)) forged by deep learning methods. Recent studies have examined whether users can distinguish human-created media from deepfakes, including text (Everett, Nurse, and Erola 2016), images/videos (Nightingale and Farid 2022), and audio (Mukhopadhyay, Shirvanian, and Saxena 2015). The results showed participants cannot effectively differentiate them (Ternovski, Kalla, and Aronow 2021; Groh et al. 2022; Korshunov and Marcel 2020). These initial efforts provided a preliminary understanding of the human perception of deepfake, but they only focused on the isolated deepfake content (e.g., standalone images) and did not explore the downstream consequences within specific attack contexts.

**Deepfake and Social Network Profiles.** Recently, researchers further explored whether deepfake images and text can be used to create believable social media personas (Mink et al. 2022), and studied what would make a *real* user profile trustworthy (Ma, Neeraj, and Naaman 2017). Finally, researchers have mixed real and fake profiles to assess users' decision-making (Kenny et al. 2022; Jakesch et al. 2019). These studies are focused on the profile-level assessment without exploring their impact on downstream attacks (e.g., disinformation campaigns), which is the focus of our paper.

**Disinformation Judgment and Sharing.** The use of deepfake profiles in disinformation campaigns is a recent trend. To the best of our knowledge, this phenomenon is not well understood yet, and our study aims to fill in the gap.

When judging (dis)information online, a key factor is *perceived source credibility* (Vraga and Bode 2017; Seo et al. 2022). Perceived credibility can be based on declarative information (e.g., a person's expertise, institutional affiliation), or the absence of conflicting interests (Petty and Cacioppo 1986). Since it is difficult to evaluate the credibility of sources in the digital age (Marsh and Yang 2017), users are likely to use simple cues in the context (e.g., cues of expertise) to make judgments about the source's credibility. In our study, we propose to examine whether and how the presence of user profiles (which act as the information source) affects users' judgment of disinformation to advance the current understanding of this problem.

Another related direction is to understand the sharing behavior of disinformation online (Talwar et al. 2019). Existing research often focused on explicit sharing decisions on news (Epstein et al. 2022; Yaqub et al. 2020). However, in practice, online users may have *implicit* news sharing via "like" and "comment" for social purposes. With these considerations, our study will examine users' engagement with disinformation (covering "comment", "like", "share", and "retweet") and explore how the presence of deepfake profiles affects users' engagement choices. We consider these engagement factors in part because of Twitter's recommendation algorithm that determines what a user sees based on the perceived popularity of a piece of content, which further affects information dissemination. Content that receives more user engagements (e.g., likes, replies) is more likely to be shown to other users' news feeds (Twitter 2023).

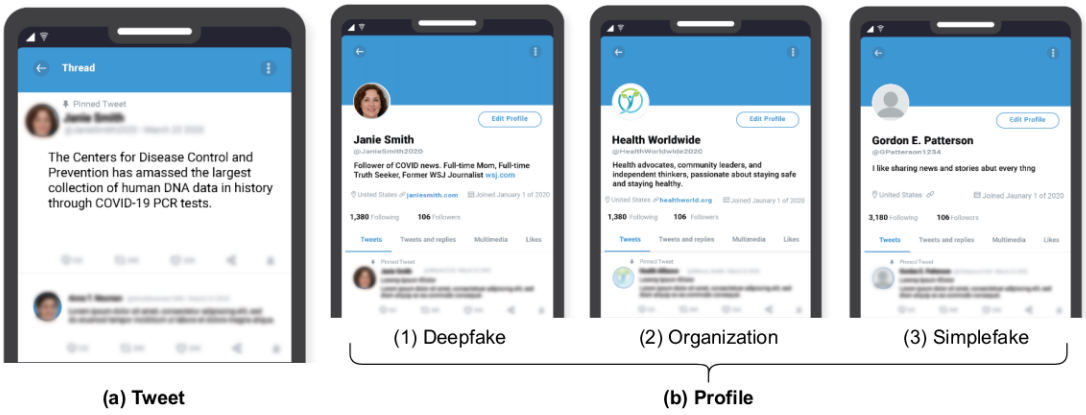


Figure 1: Example tweet and profiles.

## Methodology

To answer our research questions, we conduct an online study where participants examine three pieces of fake news, along with three different types of social media profiles (presented as the senders of the tweets), using a within-subject design. This means each participant will examine all three profile types (and three pieces of fake news), in a randomized order. [1] We do not include real news or real profiles because "using fake profiles to promote real news" or "using real profiles to promote real news" are not part of the real-world threat model. We design our study around Twitter because it has been one of the most targeted platforms by disinformation campaigns (SSCI 2017). To minimize distracting factors, we focus on fake news related to COVID because (1) it is an important global issue that affects all people around the world and (2) it has also become a polarized topic targeted by disinformation (Hart, Chinn, and Soroka 2020). During the study, participants answered questions about the accuracy of the tweeted information, their engagement with the tweets, and the perceived fakeness of the profiles. We also examine participants' political views and their prior knowledge of and experience with COVID.

### Constructing User Profiles and Tweets

On Twitter, users can read tweets from their personal or public news feeds. An example tweet is shown in Fig. 1 (a). Upon reading a tweet, users may also click on the icon of the tweet author to further examine the tweet author's profile (Fig. 1 (b)). Our study is mainly to emulate this process.

**Profile Type Stimuli.** We construct three profile types. *First*, we generate deepfake-enabled profiles posing as journalists. This is motivated by a series of real-world campaigns where journalist profiles with deepfake photos are used to spread fake news (Banerjea 2022; Gleicher and Agranovich 2022; Anderson 2019; Nimmo et al. 2020). In the case of COVID-19, it makes sense for these attackers to choose journalists as a preferred persona since journalists are commonly perceived as reliable sources of information (Hayes, Singer, and Ceppos 2007). *Second*, for comparison, we construct a known type of fake profiles posing as health organizations relevant to COVID-19 fake news. Health organizations, such as CDC or WHO, use social media to push content to online audiences. There were real-world campaigns that used similar organization profiles to spread disinformation (Wong 2020; Marr 2020). Recent works have shown that organizations (compared with individuals) are perceived as more credible sources (Vraga and Bode 2017; Vraga, Bode, and Tully 2022), and thus can serve as an up-to-date comparison baseline. *Third*, as another baseline, we construct "simplefake" profiles that mimic automatically generated bots with minimal profile information, which is also commonly observed in practice (Thomas et al. 2013).

We want to emphasize that our goal is not to compare deepfake content with non-deepfake content within a profile. Instead, we focus on such emerging deepfake-enabled personas observed in practice (as a whole), in comparison to other well-known types of fake profiles. For brevity, in the rest of the paper, we will use "deepfake profiles" to refer to deepfake-enabled social personas or profiles.

Fig. 1 (b) shows an example profile for each group. Each profile is constructed with (1) a text template (e.g., name, Twitter handle, bio, link), and (2) a profile photo. Table 1 summarizes the constructed profiles. The full set of profile screenshots and profile templates are available in our supplementary materials (Ruffin et al. 2024).

*(1) Deepfake Profiles:* We create three profile templates that pose as journalists. The names on the templates are randomly chosen using name generation software (Works 2022). The Twitter handle and the link in the bio are formed in relation to the names. The bio for each template is crafted to sound generic but also mimic a reputable American journalist that follows COVID-related news. We use a deepfake model (Photos 2019) to generate deepfake human faces as profile photos. To minimize the potential biases introduced by specific profile photos, we include a diverse set of photos: we use the deepfake model to randomly generate 16 im-

[1] We chose to use a within-subject design to reduce the influences/errors associated with the individual participants' differences (e.g., their knowledge of and the ability to detect fake news). We will discuss the limitation later in the paper.

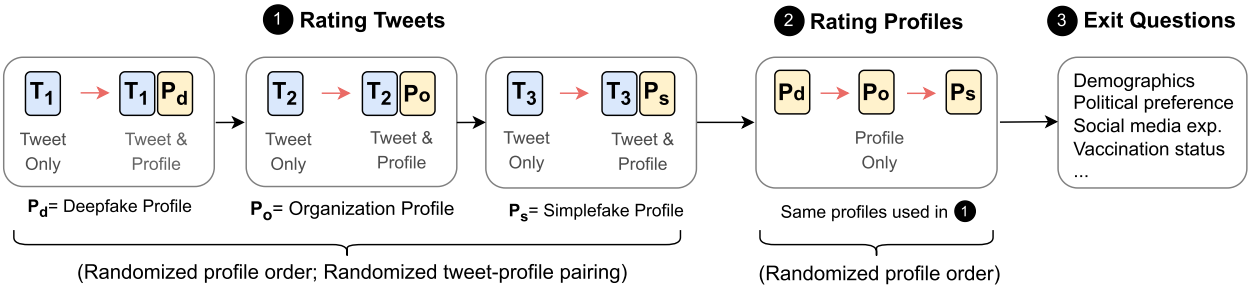


Figure 2: Study methodology (within-subject). We show the study process from *a participant's perspective*.

| Group | Text Template | Profile Photo | Profiles |
|---|---|---|---|
| Deepfake | 3 journalist | 16 human faces (2 genders × 2 ages × 4 races) | 16 |
| Organization | 3 health org. | 3 org. logos | 3 |
| Simplefake | 3 generic | 0 profile photo | 3 |

Table 1: Profile setups.

ages with a full combination of two gender groups (male and female), two age groups (old and young) and four ethnicity groups (Black, White, Asian, and Latino). The 16 photos are randomly paired with the three profile templates to generate 16 profiles. We specifically make sure the gender displayed on the photo matches that of the profile name.

*(2) Organization Profiles:* Given the context of COVID fake news, we construct three organization profiles that pose as organizations related to public health. We select three logos (Vecteezy 2022) as profile photos. The names are chosen to sound general and yet related to public health. The Twitter handles and links in the bio are formed in relation to the chosen names. The bio information is crafted to sound generic but authoritative with respect to COVID-related issues.

*(3) Simplefake Profiles:* We create three profile templates, all of which have blank profile photos. The names on these profiles are randomly selected using the same name generation software (Works 2022). The Twitter handles are formed according to the selected name. The bio information for each template is designed to mimic a generic bot profile that contains grammatical errors or typos. These profiles do not contain a link in the bio.

As shown above, for each profile condition, we create a pool of slightly varied profiles, to reduce the risk of bias toward one specific profile, in line with prior work (Kenny et al. 2022). For all three profile conditions, we keep other profile fields the same, e.g., location, and date of joining Twitter. The location is chosen based on where we recruit our participants (i.e., the United States). Finally, we set the number of "following" and "followers" based on existing measurement studies (Gurajala et al. 2016). The main consideration is that fake profiles usually have a high following-to-follower ratio. Simplefake has more "following" than other profiles, to mimic a typical bot behavior.

**Tweets: Fake News Stimuli.** We first collect a sample of COVID-related fake news from fact-checking websites including `Snopes.com`, `PolitiFact.com` and `FactCheck.org`. From a collection of 40 pieces of fake news articles, we select three stories that are recent and have been verified to be false:

*Tweet 1: "The Centers for Disease Control and Prevention has amassed the largest collection of human DNA data in history through COVID-19 PCR tests."* (Reyes 2022)

*Tweet 2: "On Dec. 28, 2021, three days before her death, Betty White said 'Eat healthy and get all your vaccines. I just got boosted today."'* (Emery 2022)

*Tweet 3: "There's a positive correlation between higher mask usage and COVID-19 deaths."* (Settles 2022)

## Experimental Design

The experiment manipulates two variables: (1) whether a profile is presented alongside the tweet; and (2) the type of profile being presented, using a within-subject design. In the following, we describe the study process *from a participants' perspective* using Fig. 2, and explain our design choices. The detailed question list is available in the supplementary materials (Ruffin et al. 2024).

**Rating Tweets** After a participant reads and signs the consent form and a brief introduction to the study, the participant starts with step ❶ of Fig. 2. The goal of this step is to understand how the manipulated factors (i.e., the presence of the profile and the type of the profile) influence users' rating of the information accuracy of and engagement with the presented news (tweets).

As shown in Fig. 2, one participant will examine three tweets (Tweet 1, 2, and 3), and the tweets are randomly paired with three profile types (deepfake, organization, and simplefake). The participant first examines a page where we only present the tweet (i.e., "without-profile" condition). When presenting the tweet alone, we have blurred other areas (e.g., the tweet author's photo) to control the influence of other stimuli. This simulates the scenario where users only browse tweets from the news feed (e.g., their personal feed) without clicking on the tweet author's profile.

After that, the participant moves to the next page, where the tweet is presented alongside the tweet author's profile (i.e., "with profile" condition). On this page, all the areas are unblurred. This process is *within-subject*, as it simulates

the process that a user first reads a tweet from the news feed and then clicks on the tweet author's icon to check out their profile. Under *both pages*, participants answer the same set of two questions (with minor wording changes, see below).

- **Q1: Information accuracy.** "On a scale from 1 to 5, after looking at this tweet [tweet and profile], how accurate do you think it is?" The answer is recorded on a five-point scale: very inaccurate (1), somewhat inaccurate (2), neither accurate nor inaccurate (3), somewhat accurate (4), and very accurate (5).
- **Q2: Engagement.** "After looking at this tweet [tweet and profile], in what way would you engage with it?" As a single-choice question, the answer is recorded from "no engagement", "like", "retweet", "reply", "external share", and "prefer not to say".

After reviewing the first tweet and profile set, participants will move on to repeat the procedure with two more tweet-profile sets. To make sure a participant does not view any repeated profiles or tweets, we have ensured that (1) the three profiles cover all three profile types and (2) the three news stimuli are randomly paired with the three profile types. To further reduce the influences of the ordering effect, we have randomized the order in which the three types of profiles appear. Given a participant and one of the three profile types, the profile is randomly selected from the profile pool for the corresponding profile type to reduce bias. In summary, in step ❶, each participant will go through six pages, i.e., with the profile being present or absent for three profile types.

**Rating Profiles** After step ❶, this participant will move to step ❷ to further examine the three profiles (the same profiles used in step ❶). The order of the profile appearance is randomized. They will answer one question for each profile:

- **Q3: Profile authenticity.** "On a scale from 1 to 5, does this profile appear fake?". We use a five-point scale: definitely fake (1), more fake than real (2), I'm not sure (3), more real than fake (4), definitely real (5).

After that, we further ask the participant to rate the important profile features (e.g., name, profile photo, bio) that have affected their determination. We ask this question for each profile *after* the participants rate the authenticity (Q3) of all three profiles to avoid priming.

- **Q4: Profile features.** "Which profile feature(s) helped you to determine whether or not the user shared accurate news? (Check all that apply)".

**Exit Questions** In the last step (step ❸), we collect demographic information of the participant (age group, gender, and education level) and ask additional questions about their frequency of listening to COVID news, political views, experience with Twitter, experience with photo editing software, and COVID vaccination status. In addition, we measure the participant's propensity of trust toward the information on Twitter. Finally, to ensure the reliability of the obtained data, we have one attention question where participants give open-ended answers describing their personal experiences with fake profiles in everyday use of social media.

**Follow-up Questions Added After Pilots** We did a small pilot run (N=20) before launching the actual study to validate the overall survey design. By analyzing the data, we observed an interesting phenomenon: some participants had rated a tweet as *very inaccurate* or *somewhat inaccurate* but still decided to engage the tweet in various ways. As such, we added two additional questions to the main study protocol to understand the reasons behind user decisions. The two questions are placed after participants complete step ❶ in Fig. 2 so that the questions would not affect their answers to Q1 and Q2 (rating tweets).

- **Q5: Recollection.** "In this survey, have you rated any tweets as Very/Somewhat Inaccurate and still decided to Like, Reply, Retweet, or Share the inaccurate tweets?". This question prompts users to recall their decisions.
- **Q6: Reason to engage.** "If you answered 'Yes, I have' to the previous question, can you explain why?" This question is open-ended.

## Recruitment

We recruited participants from Prolific between July and August 2022. We recruited participants from the U.S. who had a Minimum Approval Rating of 95%. Participants' results were excluded if they failed to give a meaningful answer to the attention check question (manually verified by the authors). Out of 424 responses, only 7 were filtered. In total, we had N=417 valid responses for the final version of the study[2]. Overall, participants' ages are between 18-50+ with a median of 30-39 years old; 48.44% identified as male, 48.44% as female, and 3.12% as non-binary or other. The survey takes a median of five minutes to complete, and each participant is compensated $1.40 for their time.

# Results

## Perceived Information Accuracy

We start with **RQ1** regarding the perceived accuracy of the tweet. Recall that we have two manipulated factors: (1) profile type and (2) whether the profile is presented alongside the tweet (referred to as "presence of profile"). As shown in Fig. 3, presenting the profile together with the tweet either increases or decreases the mean value of the perceived tweet accuracy rating, depending on the specific profile type.

To quantify the effect, we first construct an overall model to capture both factors with an ANOVA test (Norman 2010). We use a *linear mixed-effects regression* (or LMER) model with the lme4 package in R (Bates et al. 2011). We choose LMER (instead of linear regression) because LMER can model random effects, allowing for non-independence between measured outcomes[3] (Montgomery, Peck, and Vining

[2]Using G*Power (Faul et al. 2009), we run a post-hoc power analysis for small effect size (Cohen's $f$=0.1) with a 2×3 within-subject design, and an alpha of 0.05, our sample (N=417) achieved a power of 0.99.

[3]We specified the value for each option in the accuracy evaluation (and later for authenticity evaluation too) and the space between scales is the same. Thus, we chose LMER rather than Cumulative Link Mixed Models (or CLMM) for the analysis.

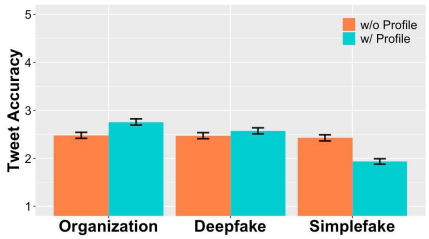


Figure 3: Perceived tweet accuracy rating: mean and standard error.

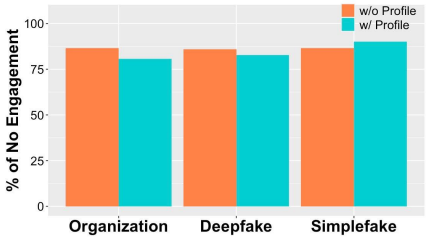


Figure 4: % of Participants who selected "no engagement" towards the tweet.

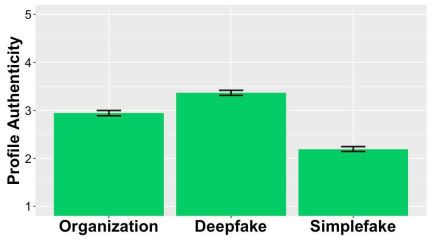


Figure 5: Perceived profile authenticity rating: mean and standard error.

| | Organization | | Deepfake | | Simplefake | |
|---|---|---|---|---|---|---|
| Variable | $\beta$ | $p$ | $\beta$ | $p$ | $\beta$ | $p$ |
| *Intercept* | -0.517 | 0.11 | -0.527 | 0.089 | -0.577 | 0.058 |
| Presence of Profile (Reference = w/o Profile) | | | | | | |
| w/ Profile | 0.278 | <0.001*** | 0.101 | 0.0175* | -0.492 | <0.001*** |

Table 2: Effect of the presence of profile on tweet accuracy. We build one LMER model for each profile type. For each LMER, we compare the effect of showing profile (i.e., w/ profile) with not showing profile (i.e., w/o profile) on the perceived tweet accuracy. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

2001). In our case, we have repeated measurements (in terms of participants and tweet messages), and thus this modeling is most appropriate. We model the tweet accuracy rating as the dependent variable and take the 2-way interaction between the two factors *profile type* and *presence of profile* as the fixed effect. To account for the within-subject design and the tweets used across conditions, we take the participant ID and tweet ID as the *random effect*,[4] which is similar to that used in prior work (Pennycook et al. 2021).

**ANOVA Results.** Using this LMER model, we first report the ANOVA results, including the $F$ values and $p$ values. We report the degree of freedom with Satterthwaite approximation. For the $\beta$ slopes of mixed-effect analyses, we will report them later during the post-hoc analysis. We assume an $\alpha = 0.05$ for significance in hypothesis testing for this and the following sections (i.e., significant if $p < 0.05$).

The modeling results confirm the observed trend. First, *profile type* has a significant effect on the perceived tweet accuracy ($F_{(2,2083)} = 47.40$, $p < 0.001$). Second, the *presence of profile* does not have a significant effect overall ($F_{(1,2083)} < 1.0$, $p = 0.34$), but the interaction result indicates that this effect is significantly different under different profile types ($F_{(2,2083)} = 35.30$, $p < 0.001$).

**Post-hoc Analysis.** The ANOVA result motivates us to run a post-hoc analysis to quantify the exact effect of the *presence of profile* under different profile types. The goal is to reveal whether the presence of a profile has a significantly positive or negative effect, for each of the profile types. Since we are not comparing the relative differences between profile types, for this analysis, we construct three LMER models, one for each profile type. The dependent variable is the tweet accuracy rating, and the *presence of profile* is the fixed effect. Participant ID and tweet ID are still random effects. In Table 2, we present the estimates $\beta$ (regression coefficients) and $p$ values for each model. First, for "organization" profiles, showing the profile alongside the tweet (compared with showing the tweet alone) has significantly improved the perceived tweet accuracy ($\beta = 0.278$, $p < 0.001$). As shown in Fig. 3, the mean accuracy is increased from 2.48 to 2.76 with the profile present. Second, the presence of "deepfake" profiles also has a positive and significant effect ($\beta = 0.101$, $p < 0.05$). The mean increases from 2.47 to 2.57 when the profile is present. Third, "simplefake" profiles have the opposite (negative) effect as they significantly decrease the perceived tweet accuracy ($\beta = -0.492$, $p < 0.001$). The mean accuracy decreases from 2.43 to 1.94. This also explains why the presence of profile does not have a significant *overall* effect, since the effects under each profile type may cancel each other.

The result answers **RQ1**: the perceived tweet accuracy is increased when a relevant/reputable profile (organization or deepfake) is presented compared with showing the tweet alone. However, the perceived accuracy decreases when the presented profile is a simple bot profile. Such results are in line with prior research showing a positive influence of source credibility on persuasion (Petty and Cacioppo 1986).

## User Engagement with Tweets

To investigate **RQ2**, we analyze user responses to the question regarding their engagement method with the presented tweet. Participants can choose one of the six options: "like", "reply", "external share", "retweet", "no engagement", and "prefer not to say." We conservatively coded "prefer not to say (NTS)" to the negative category (i.e., no-engagement).

[4]Each participant has a unique user ID, and each tweet is assigned to a unique tweet ID. They were part of the repeated measurements and thus treated as the random effect.

| | Organization | | Deepfake | | Simplefake | |
|---|---|---|---|---|---|---|
| Variable | $\beta$ | $p$ | $\beta$ | $p$ | $\beta$ | $p$ |
| *Intercept* | -14.287 | <0.001*** | -9.155 | <0.001*** | -9.055 | <0.001*** |
| Presence of Profile (Reference = w/o Profile) | | | | | | |
| w/ Profile | 4.763 | 0.001*** | 1.146 | 0.011* | -2.404 | <0.001*** |

Table 3: Effect of the presence of profile on engagement.

As shown in Fig. 4, over 80% of the participants across all conditions chose "no engagement." Comparing different conditions, Fig. 4 shows that presenting profiles (compared with showing tweets alone) either increases or decreases the rate of no-engagement, depending on the profile type.

To quantify the effect, we run a similar model as before. For simplicity, we present users' decisions as a binary variable with two possible values "engagement" (coded as 1) and "no-engagement" (coded as 0). We evaluate this as a binary variable because all engagement types (regardless of intention) can potentially increase the popularity of the tweet on the platform (Twitter 2023). We construct a *logistic mixed-effects regression* model using the user engagement decision as the dependent variable. We choose logic regression considering the binary nature of the dependent variable. Similar to before, we take the 2-way interaction between the two factors *profile type* and *presence of profile* as the fixed effect and participant ID as the random effect to account for the multiple measurements from a participant. We have simplified the random effect structure (by excluding Tweet ID) since the more complex model would not converge well even after we optimized the parameters using the all_fit function (Singmann et al. 2015). This may be a less accurate model, but it gives reliable results (Brown 2021). We find the main effects of profile type ($\chi^2_{(2)} = 10.68, p = 0.005$) and its interaction with the presence of profile ($\chi^2_{(2)} = 10.66, p = 0.005$) are significant.

We perform the same post-hoc analysis using *presence of profile* as the fixed effect. As shown in Table 3, the effect is similar to that of the perceived tweet accuracy. The likelihood of users deciding to engage with the tweet is significantly higher when presenting organization profiles ($\beta = 4.763$, $p = 0.001$) compared with presenting the tweet alone. This corresponds to the result in Fig. 4, which shows the percentage of "engaging" participants increases from 11.43% to 19.18%. The same observation applies to deepfake profiles, which also increase the likelihood of engagement ($\beta = 1.146$, $p < 0.05$). The percentage of participants who chose to engage after seeing the profile increases from 14.15% to 17.27%. The effect is negative for simplefake profiles as they decrease the likelihood of user engagement ($\beta = -2.404$, $p < 0.05$). The percentage of engaging participants decreases from 13.43% to 9.83%.

While "no engagement" is the most selected option, we further analyze *the other choices* made by participants when they *decided to engage* with the tweet. We find that "like" was chosen the most across conditions (see the figures in the supplementary materials for a detailed breakdown (Ruffin et al. 2024)). For organization/deepfake profiles, when a profile is presented, the percentage of participants that chose "like" is also increasing, going from 5.76% to 9.59%, and 5.04% to 6.71%, respectively). This is not true for simplefake, which decreases from 5.28% to 3.80%. We also compare the time that participants spent on answering Q1 and Q2. No statistical differences are obtained across conditions except that participants spent a shorter time when they viewed the tweet again with profile ($\beta = -3.11, p = 0.005$).

Overall, the results answer **RQ2**: the engagement of tweet increases when a profile of authority (organization and deepfake) is presented, compared with showing the tweet alone. However, engagement decreases when the presented profile is a simple bot profile.

## Why Do Users Engage Inaccurate Tweets?

During the engagement analysis, we observe an interesting phenomenon: users may decide to like/reply/share/retweet a tweet even if they believe the tweet is inaccurate. More specifically, in step ❶, out of 417 participants, 54 (12.9%) of them at least once decided to engage with the tweet that they rated as "very inaccurate" or "somewhat inaccurate." In the previous analysis, we code user engagement decisions into "engagement" and "no-engagement" for ease of analysis. Here, we provide a more in-depth analysis of different engagement types and study the user motivations behind their engagement decisions.

To understand *why*, in our survey, we first ask participants to recall their accuracy decisions and engagement choice selections (**Q5**) and then ask those who recalled their engaging with disinformation to explain their reasons (**Q6**). Among these 54 participants who engaged with inaccurate tweets in step ❶, 37 (68.5%) correctly recalled their accuracy and engagement decisions and answered **Q6** to explain their reasons. The remaining 17 (31.5%) couldn't recall or believed they didn't do so and thus did not answer **Q6**. Surprisingly, another 17 participants answered **Q6** even though they did not engage with inaccurate tweets (the number is also 17, by incident), including 8 people who did not engage with any tweet and 9 people who only engaged with tweets rated "accurate." In total, we got 54 responses from **Q6**. As a comparison, for participants who did not engage with inaccurate tweet, most of them correctly recalled their decisions from step ❶ (242/251=96.4%).

We do a thematic analysis (Braun and Clarke 2012) to all the 54 responses but mark the results from the participants who made recall errors on engagement (8) or accuracy (9). To develop a code book, two authors independently code the responses (Strauss and Corbin 1997) in the opposite order.

| | Profile Authenticity Rating | | Tweet Accuracy Rating | |
|---|---|---|---|---|
| Variable | $\beta$ | $p$ | $\beta$ | $p$ |
| Intercept | 0.365 | <0.001*** | -0.427 | 0.101 |
| Profile Type (Reference = Deepfake) | | | | |
| Organization | -0.422 | <0.001*** | 0.188 | 0.009** |
| Simplefake | -1.170 | <0.001*** | -0.641 | <0.001*** |

Table 4: Effect of profile type. We construct a LMER model to compare the effect of different profile types on profile authenticity (left). Then, we construct a similar LMER model to compare the effect of profile types on tweet accuracy rating (right).

The initial inter-coder agreement via Cohen's Kappa is $0.86$, indicating a high agreement. Two coders and the remaining authors resolve the four disagreed items to improve the code book. Then the two coders use the updated code book to analyze responses in opposite order again. We identify two major themes and two extra themes, each of the themes involves different user engagement actions.

**Seek More Information.** Twenty participants ($37.0\%$, 1 made recall error on engagement and 2 on accuracy) explained that they seek further information, such as sources from where the tweeted message comes, to verify the information veracity. For example, **P44** responded, "*I replied to one of them to ask why.*" Moreover, **P62** explained that "*I have liked some of the tweets I thought were inaccurate to 'save them' and go back to the tweet after doing my own research/fact-checking.*"

**Refute Disinformation.** 13 participants ($24.1\%$, 1 made recall error on accuracy) revealed that they actively refute the false information either to let the poster know the truth (e.g., "*Sometimes you need to speak some sense into people when they are incredibly wrong* (**P72**).") or to avoid disseminating false information among other people (e.g., "*Because if I see something so blatantly false I feel like I have to reply a response that sows seed of doubt and hope that people would think twice about false information* (**P323**).")

We also observe two extra themes in a small percentage (about 8.4% on average) related to reasons for sharing inaccurate tweets. First, five participants chose to share in order to *having fun*. For example, **P199** said, "*I would share them to Reddit to make fun of the inaccuracies*" and **P173** replied, "*I thought the 'CDC collecting human DNA' was very 'conspiracy theory' heavy so I marked that I would externally share it to my sister who always likes a laugh at conspiracies, but I would not have interacted with it at all in the app in hopes to not boost [its] popularity.*" Second, four participants chose to share because they thought the tweets were *consistent with their beliefs* (e.g., "*Their idea[l] goes along with my beliefs* (**P388**)." And "*There are certain topics that I believe (they) are real and true.* (**P398**)"). Such results are not surprising since these four participants believed the tweets were "accurate" initially.

Overall, the result suggests that users have diverse reasons for liking, replying, or retweeting/sharing a tweet. Such actions should not be automatically treated as a signal of the user agreeing with (or supporting) the message in the tweet.

## Perceived Profile Authenticity

Next, we explore **RQ3** regarding the perceived authenticity (fake or real) of a profile.

**Authenticity Rating.** Fig. 5 shows that profiles of authority (organization and deepfake) have a higher rating compared with simplefake. To quantify the effect, we again construct a LMER model where we take the authenticity rating as the dependent variable and take profile type as the fixed effect. To account for the within-subject design, we take participant ID as the *random effect*. We use the "deepfake" condition as the intercept (reference) because we are interested in comparing deepfake with other profile types.

The model confirms that the main effect of profile type is significant ($F_{(2,834)} = 154.42, p < 0.001$). As shown in Table 4 (left), simplefake (2.19) is considered significantly less real compared with deepfake (3.36, $\beta = -1.170$, $p < 0.001$). Surprisingly, organization profiles (2.94) are also considered to be less real compared with deepfake ($\beta = -0.422$, $p < 0.001$).

This observation is slightly different from the profile effect on *perceived tweet accuracy* for **RQ1**. As shown in Fig. 3, the mean value of the tweet accuracy rating for organization profiles (2.76, when the profile is presented) is higher than that of the deepfake (2.57). To confirm the significance, we again run LMER as a post-hoc analysis. We take the tweet accuracy rating as the dependent variable and the profile type as the fixed effect (tweet ID and participant ID as the random effect like before). The results are shown in Table 4 (right). To make the result comparable, we also use the "deepfake" condition as the reference. The results confirm that when presenting organization profiles with the tweet, the perceived tweet accuracy rating is significantly higher than the reference when presenting deepfake profiles ($\beta = 0.188$, $p < 0.01$). We also observe that when presenting the simplefake profiles with the tweet, the perceived tweet accuracy rating is significantly lower than the reference ($\beta = -0.641$, $p < 0.001$).

These analyses return two main findings. First, it answers **RQ3** that participants consider deepfake profiles to be more real than other profile types. Second, we observe an interesting difference between deepfake and organization profiles: deepfake is considered more real than organization profiles, but organization profiles are more effective in increasing the perceived tweet accuracy compared with deepfake profiles. We will further discuss this at the end of the paper.

| Org. | Count, % | Deepfake | Count, % | Simplefake | Count, % |
|---|---|---|---|---|---|
| Bio | 272, 25% | Bio | 322, 32% | Bio | 307, 33% |
| Link | 247, 23% | Link | 231, 23% | Photo | 233, 25% |
| Name | 182, 17% | Photo | 168, 16% | Tw Handle | 114, 12% |
| Tw Handle | 182, 17% | Name | 123,12% | Name | 96, 10% |
| Photo | 133, 12% | Tw Handle | 123, 12% | Link | 83, 9% |
| others | 81, 7% | others | 54, 5% | others | 92, 10% |
| Total | 1097, 100% | Total | 1021, 100% | Total | 925, 100% |

Table 5: Profile features that influence the information accuracy. Features are sorted based on participants' selection.

Additionally, we noticed participants spent different time evaluating fake profiles ($F_{(2,2085)} = 14.49, p < 0.001$). Compared to deepfake profiles, they spent a shorter time on the simplefake profiles ($\beta = -1.26$, $p < 0.001$) and the organization profiles ($\beta = -0.84$, $p < 0.001$), respectively.

**Influencing Factors.** To address the second part of **RQ3**, we asked participants to specify the factors they considered when evaluating the profiles. Among the six options, more than half of them selected "bio," "links in profile," and "profile photo" across conditions (see Table 5). Although participants examined "bio" the most regardless of profiles, Chi-squared test showed that the selection ratios varied ($\chi^2_{(2)} = 15.67$, $p < 0.001$). Specifically, "organization" presented a significant difference from "deepfake" ($p < 0.001$), and "simplefake" ($p < 0.05$), respectively. While most participants also looked into "links in profile" for the organization and deepfake profiles, they relied more on "profile photo" (or the lack of photo) when evaluating simplefake profiles.

We also noticed that, when evaluating organization profiles, participants considered more features (a total of 1097) compared to the other profiles, especially simplefake (925). Also, all the features were more evenly selected for organization profiles (i.e., participants relied on relatively fewer features for examining the deepfake and simplefake profiles). Combining this result with the profile authenticity rating and time spent, the implication is that it is difficult for participants to evaluate deepfake profiles, and they have failed to detect the critical feature (e.g., profile photo, 16%) for the deepfake evaluation.

## Exploratory Analyses

We conducted an exploratory analysis to understand whether individual differences (including political views, *photo editing software experience*, *propensity to trust the information on Twitter*, and *prior knowledge of and experience with COVID*) had an impact on the obtained results. We did not obtain any extra statistically significant results due to the added factors except the main effects of political views and trust propensity. Due to the space limit, we put the details in our supplementary materials (Ruffin et al. 2024).

# Discussion and Conclusion

In this paper, we conduct an online study examining human perceived information accuracy in fake news (**RQ1**) and engagement with fake news (**RQ2**) as a function of the presence of a fake profile and the profile type (**RQ3**). Our results show that the effect of a fake profile depends on the profile type. When the tweets were presented along with the fake profiles, participants decreased the information accuracy rating for the simplefake but increased their accuracy rating for the organization and deepfake profiles. A similar effect is observed for user engagement with the tweets. Comparing deepfake with organization profiles, we found that deepfake profiles were perceived to be more real than organization profiles, while organization profiles have a higher impact on increasing the perceived accuracy of their tweets. Below, we discuss the possible reasons behind our observations and their implications.

**Fake Profiles and Fake News Are Associated but Different.** In practice, deepfake profiles have been used in the context of social engineering (Satter 2019) and disinformation campaign (Banerjea 2022; FBI 2021; Nimmo et al. 2019, 2020). Notably, our study provides empirical evidence showing the significant impact of displaying such profiles (i.e., combining deepfake images with an insider/expert persona) on participants' accuracy rating of and engagement with disinformation. Since we focus on such deepfake-enabled profiles as a whole, we do not intend to attribute this impact solely to the use of deepfake images. We believe that our work is just a preliminary step to exploring the impact of deepfake-enabled personas in the context of downstream attacks. Other elements of the profile can also be generated using deepfake techniques. For instance, deepfake videos can be used as "evidence" to be shared together with the fake news by the social media persona (Köbis, Doležalová, and Soraperra 2021). In addition, models such as ChatGTP (OpenAI 2022) and Bard (Pichai 2023) can generate high-quality texts, which can be used to generate profile bios and tweets, and even allow fake personas to have a live chat with victim users. Future work may examine a deepfake orchestration combining different techniques, which could result in more advanced attacks.

Another interesting observation is that deepfake-enabled profiles' impact (compared with organization profiles) varied under different tasks. While participants' perceived accuracy rating of disinformation was more influenced by the organization profile, they rated the deepfake profiles as more real. One possible explanation is that human information processing and the consequent performance are *task dependent* (Monsell 1996). When the fake profile itself was presented, participants carefully examined the features embedded within the profile (i.e., system 1, analytic) (Kahneman

2011) to assess profile authenticity. However, when the profile was presented along with the tweet message, the participants' primary goal was to evaluate the veracity of the tweet message. Thus, the fake profile was secondary to their evaluation (i.e., system 2, heuristic). They might have taken some embedded features (e.g., relevant organizations or reputable individuals) as heuristic cues (Gigerenzer and Gaissmaier 2011) to make the assessment, as suggested by the shorter time when viewing tweets again with a profile. Therefore, to have a comprehensive understanding of humans' susceptibility to deepfake and disinformation, we must consider ecologically representative tasks during the study design.

**User-initiative Correction Can Lead to Implicit Information Dissemination.** Some participants decided to engage with tweets even that they rated as very/somewhat inaccurate in an attempt to verify or refute the disinformation. Such results are in line with recent work that shows user-initiated correction of COVID-19 false information played a large role in social media platform (Bode and Vraga 2021). Also, our results point out the nuanced challenges of such user-initiative correction, that is, users may have unintentionally disseminated the information when interacting with or attempting to correct the false information.

For example, one participant "liked" a tweet to save the tweet to their profile (so that they can easily find it after seeking external verification of the information). However, currently liking a tweet would also increase the perceived popularity of the tweet, contributing to its dissemination. Nevertheless, the correction initiators (e.g., users) might not have been aware of such unexpected consequences. To facilitate more effective user-initiated correction, lightweight interventions on social media platforms can be considered. For example, when online users attempt to reply to a piece of a false story, the platforms could present a pop-up screen to inform the possible propagation of disinformation through such engagement. Meanwhile, the platforms could recommend alternative actions (e.g., private communication) to mitigate the risks. However, the proposed solutions require more research to examine before potential implementation.

**Countermeasures: Combining Deepfake Detection with User Intervention** To defend against this threat, we believe that technical solutions such as automated detection methods should be the front line of deepfake defense. Social media platforms should try to detect deepfake-enabled profiles to prevent them from reaching online users. For example, recognizing this threat, LinkedIn recently announced that they are using deepfake detection models to screen the profile photos of newly registered accounts (Rodriguez 2022). However, deepfakes are constantly and rapidly evolving (Hussain et al. 2021; Mirsky and Lee 2021), and existing automated detection methods may not be able to prevent all deepfake-based profiles from reaching end users. This is also evidenced by the fact that many fake accounts have bypassed such detection (Krebs 2022).

As such, it is also necessary to equip online users with the appropriate knowledge and skills to recognize deepfake profiles. Possible interventions include user training or displaying warnings on suspicious content. Previous studies have trained participants to examine different deepfake artifacts and found that the training improved participants' detection rate of deepfake photos/videos (Mink et al. 2022; Tahir et al. 2021). Yet both studies also reported an "implied fake" effect for the real photos or videos (e.g., people are more likely to accuse real photos as fake after training), raising the challenges of designing effective training strategies. Recently, researchers also revealed different susceptibilities of algorithm-based deepfake detection (e.g., informed by perceptual properties) and human-based detection (e.g., informed by contextual information) (Groh et al. 2022; Korshunov and Marcel 2021). Thus, future work could consider the integration of technical- and human-aspect solutions for deepfake detection and mitigation.

**Choice of Fake News Topic.** A limitation of our study is that we only examined fake news related to COVID-19. We chose this topic because it is an important global issue that has affected people around the world and it has been a polarized topic targeted by disinformation (Hart, Chinn, and Soroka 2020). Also, deepfake-enabled profiles were found in COVID-19-related disinformation campaigns. For example, Facebook identified deepfake-enabled accounts used to promote false narratives about the origins of COVID-19 (Lyons 2021). Another related concern is that participants' prior knowledge of COVID-19 and related news may have influenced our results; however, we did not find statistically significant influences from these factors (see supplementary materials (Ruffin et al. 2024)). Future work can extend our study methodology to study the impact of fake personas in spreading disinformation on other news topics.

**Other Limitations.** *First*, interval validity. To control the stimuli during the study, we only presented one tweet for each profile, and the tweets were focused on three pieces of fake news (without including truthful news). Future work may consider including more historical tweets in a profile to study their impact. Moreover, to control for participants' individual differences, we chose a within-subject design where each participant reviewed all three pieces of news and three types of profiles. It's possible participants may have anticipated the second/third rounds of stimuli viewing after the first round and thus paid more/less attention. To counter this ordering effect, we have randomized the order in which participants viewed the tweets and profiles to reduce the bias. Furthermore, in our study, we asked users questions about the information accuracy of the news, which may have primed users to check the news more carefully than they would do otherwise (Pennycook et al. 2021). This effect is applied to all experiment conditions and thus is unlikely to affect our conclusion.

*Second*, external validity. Our study is focused on the Twitter platform and COVID-19. Further research is needed to explore whether the findings generalize to other social media platforms and other new topics. *Third*, ecological validity. We presented participants with mock-up interfaces. Although we tried to make the stimuli as tangible as possible, they might have offered different experiences to the participants. *Fourth*, our study is focused on users from the United States (so that they are familiar with the subjects/topics of the selected fake news). Further studies are needed to explore how the results generalize to news and

participants from other regions of the world. *Fifth*, our participants were recruited from Prolific online (Tang, Birrell, and Lerner 2022). Users recruited online may not be representative of the general U.S. population (e.g., skewed to younger and more educated populations) (Redmiles, Kross, and Mazurek 2019; Kang et al. 2014).

**Broader Perspective, Ethics and Competing Interests.** Our study was reviewed and approved by our Institutional Review Board (IRB). We asked for informed consent from participants at the beginning of the study. We did not collect personally identifiable information (PII) from the participants. Participants can withdraw their data at any time after completing the survey. While our study involves showing fake news to users, we believe the potential risk is minimal, as a recent study shows that misinformation research studies in general do not significantly increase participants' long-term susceptibility to misinformation used in the experiments (Murphy et al. 2020). In the meantime, the benefit of the study is to provide a deeper understanding of the impact of fake social personas in disinformation campaigns, and results can help to inform effective countermeasures. The benefit outweighs the potential risk.

## Acknowledgments

This work was supported in part by NSF grants 2121097, 2030521, 2055233, and 2229876, the Graduate Research Fellowship Program under Grant No 21-46756, and an Amazon Research Award.

## References

Adikari, S.; and Dutta, K. 2014. Identifying Fake Profiles in LinkedIn. In *Proc. of PACIS*.

Anderson, M. 2019. Twitter, Facebook ban fake users; some had AI-created photos. TechXplore.

Banerjea, A. 2022. Digital war: How Russia is using deep fakes in Ukraine for propaganda. Business Today.

Bates, D.; et al. 2011. Package lme4: Linear mixed-effects models using S4 classes (Version 1.1-27.1).

Bode, L.; and Vraga, E. K. 2021. Correction experiences on social media during COVID-19. *Social Media+ Society*, 7(2).

Bond, S. 2022. AI-generated fake faces have become a hallmark of online influence operations. NPR.

Braun, V.; and Clarke, V. 2012. *Thematic analysis.* American Psychological Association.

Brown, V. A. 2021. An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1).

Emery, D. 2022. Did Betty White say she got Covid Booster 3 days before she died? Snopes.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proc. of ICWSM*.

Everett, R. M.; Nurse, J. R. C.; and Erola, A. 2016. The anatomy of online deception: what makes automated text convincing? In *Proc. of SAC*.

Fagni, T.; Falchi, F.; Gambini, M.; Martella, A.; and Tesconi, M. 2020. TweepFake: about Detecting Deepfake Tweets. *CoRR*.

Faul, F.; Erdfelder, E.; Buchner, A.; and Lang, A.-G. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4): 1149–1160.

FBI. 2021. Private Industry Notification (210310-001). FBI. https://www.ic3.gov/Media/News/2021/210310-2.pdf.

Fosco, C.; Josephs, E.; Andonian, A.; Lee, A.; Wang, X.; and Oliva, A. 2022. Deepfake Caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines. *arXiv preprint arXiv:2206.00535*.

Gigerenzer, G.; and Gaissmaier, W. 2011. Heuristic decision making. *Annual Review of Psychology*, 62(1): 451–482.

Gleicher, N.; and Agranovich, D. 2022. Updates on our security work in Ukraine. https://about.fb.com/news/2022/02/security-updates-ukraine/.

Groh, M.; Epstein, Z.; Firestone, C.; and Picard, R. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *National Academy of Sciences*, 119(1): e2110013119.

Gurajala, S.; White, J. S.; Hudson, B.; Voter, B. R.; and Matthews, J. N. 2016. Profile characteristics of fake Twitter accounts. *Big Data and Society*, 3(2).

Hart, P. S.; Chinn, S.; and Soroka, S. 2020. Politicization and polarization in COVID-19 news coverage. *Science Communication*, 42(5): 679–697.

Hayes, A. S.; Singer, J. B.; and Ceppos, J. 2007. Shifting roles, enduring values: The credible journalist in a digital age. *Journal of Mass Media Ethics*, 22(4): 262–279.

Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; and McAuley, J. 2021. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *Proc. of WACV*.

Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naaman, M. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proc. of CHI*.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kang, R.; Brown, S.; Dabbish, L.; and Kiesler, S. 2014. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *Proc. of SOUPS*.

Kenny, R.; Fischhoff, B.; Davis, A.; Carley, K. M.; and Canfield, C. 2022. Duped by bots: why some are better than others at detecting fake social media personas. *Human Factors*.

Köbis, N. C.; Doležalová, B.; and Soraperra, I. 2021. Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11): 103364.

Korshunov, P.; and Marcel, S. 2020. Deepfake detection: humans vs. machines. *arXiv preprint arXiv:2009.03155*.

Korshunov, P.; and Marcel, S. 2021. Subjective and objective evaluation of deepfake videos. In *Proc. of ICASSP*.

Krebs, B. 2022. Glut of Fake LinkedIn Profiles Pits HR Against the Bots. Krebs on Security.

Lyons, K. 2021. Facebook took down a fake Swiss scientist account that was part of an international misinfo campaign. https://www.theverge.com/2021/12/3/22815906/facebook-meta-instagram-fake-swiss-scientist-account-china-misinformation.

Lyu, S. 2020. Deepfake detection: Current challenges and next steps. In *Proc. of ICMEW*.

Ma, X.; Neeraj, T.; and Naaman, M. 2017. A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles. In *Proc. of ICWSM*.

Marr, B. 2020. Coronavirus fake news: How Facebook, Twitter, and Instagram are tackling the problem. Forbes.

Marsh, E. J.; and Yang, B. W. 2017. A call to think broadly about information literacy. https://psycnet.apa.org/doi/10.1016/j.jarmac.2017.09.012.

Mink, J.; Luo, L.; Barbosa, N. M.; Figueira, O.; Wang, Y.; and Wang, G. 2022. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *Proc. of USENIX Security*.

Mirsky, Y.; and Lee, W. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1): 1–41.

Monsell, S. 1996. Control of mental processes. In Bruce, V., ed., *Unsolved mysteries of the mind: Tutorial essays in cognition*, 93–148. Lawrence Erlbaum.

Montgomery, D.; Peck, E.; and Vining, G. 2001. *Introduction to linear regression analysis*. Wiley.

Mukhopadhyay, D.; Shirvanian, M.; and Saxena, N. 2015. All your voices are belong to us: Stealing voices to fool humans and machines. In *Proc. of ESORICS*.

Murphy, G.; Loftus, E.; Grady, R. H.; Levine, L. J.; and Greene, C. M. 2020. Fool me twice: How effective is debriefing in false memory studies? *Memory*, 28(7): 938–949.

Nightingale, S. J.; and Farid, H. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. In *Proc. of PNAS*.

Nimmo, B.; Eib, C. S.; Tamora, L.; Johnson, K.; Smith, I.; Buziashvili, E.; Kann, A.; Karan, K.; de León Rosas, E. P.; and Rizzuto, M. 2019. OperationFFS: Fake Face Swarm.

Nimmo, B.; François, C.; Eib, C. S.; and Ronzaud, L. 2020. Facebook Takes Down Small, Recently Created Network Linked to Internet Research Agency.

Norman, G. 2010. Likert scales, levels of measurement and the "laws"of statistics. *Advances in Health Sciences Education*.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. OpenAI. https://openai.com/blog/chatgpt/.

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855): 590–595.

Petty, R. E.; and Cacioppo, J. T. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*, 1–24. Springer.

Photos, G. D. 2019. Unique, worry-free model photos. Generated Media, Inc. https://generated.photos/.

Pichai, S. 2023. Bard: An important next step on our AI journey. Google. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Prolific. 2022. Quickly find research participants you can trust. Prolific Inc. https://www.prolific.co/.

Redmiles, E. M.; Kross, S.; and Mazurek, M. L. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proc. of IEEE SP*.

Reyes, Y. 2022. Politifact - No, the CDC is not collecting human DNA from COVID-19 PCR tests.

Rodriguez, O. 2022. New LinkedIn profile features help verify identity, detect and remove fake accounts, boost authenticity. LinkedIn Official Blog.

Ruffin, M.; Seo, H.; Xiong, A.; and Wang, G. 2024. supplementary materials. https://tinyurl.com/3fvcnauu.

Satter, R. 2019. Experts: Spy used AI-generated face to connect with targets. APNews.

Seo, H.; Xiong, A.; Lee, S.; and Lee, D. 2022. If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation. In *Proc. of ICWSM*.

Settles, G. 2022. Politifact - No evidence that masks caused covid-19 deaths.

Singmann, H.; Bolker, B.; Westfall, J.; Aust, F.; and Ben-Shachar, M. S. 2015. afex: Analysis of factorial experiments. *R package version 0.13–145*.

SSCI. 2017. Assessing Russian activities and intentions in recent US elections. Director of National Intelligence.

Strauss, A.; and Corbin, J. M. 1997. *Grounded theory in practice*. Sage.

Tahir, R.; Batool, B.; Jamshed, H.; Jameel, M.; Anwar, M.; Ahmed, F.; Zaffar, M. A.; and Zaffar, M. F. 2021. Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proc. of CHI*.

Talwar, S.; Dhir, A.; Kaur, P.; Zafar, N.; and Alrasheedy, M. 2019. Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services*, 51: 72–82.

Tang, J.; Birrell, E.; and Lerner, A. 2022. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proc. of SOUPS*.

Ternovski, J.; Kalla, J.; and Aronow, P. M. 2021. Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. *OSF Preprints*.

Thomas, K.; McCoy, D.; Grier, C.; Kolcz, A.; and Paxson, V. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *Proc. of USENIX Security*.

Twitter. 2023. Twitter's recommendation algorithm. Twitter Engineering. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

Vecteezy. 2022. Download free vector art, Stock Photos and Stock Video Footage. Eezy Inc. https://www.vecteezy.com/.

Vraga, E. K.; and Bode, L. 2017. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5): 621–645.

Vraga, E. K.; Bode, L.; and Tully, M. 2022. Creating news literacy messages to enhance expert corrections of misinformation on Twitter. *Communication Research*, 49(2): 245–267.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. 2020. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In *Proc. of IJCAI*.

Wong, Q. 2020. Twitter users duped by fake account that falsely claimed Daniel Radcliffe has coronavirus. CNET.

Works, C. 2022. Fake name generator. Corban Works, LLC. https://www.fakenamegenerator.com/.

Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *SIGKDD Explorations Newsletter*, 21(2): 80–90.

Xiao, C.; Freeman, D. M.; and Hwa, T. 2015. Detecting Clusters of Fake Accounts in Online Social Networks. In *Proc. of AISec*.

Yaqub, W.; Kakhidze, O.; Brockman, M. L.; Memon, N.; and Patil, S. 2020. Effects of credibility indicators on social media news sharing intent. In *Proc. of CHI*.

## Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Methodology, and Discussion and Conclusion.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Methodology.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Choice of Fake News Topic and Other Limitations.
   (e) Did you describe the limitations of your work? Yes, see the Choice of Fake News Topic and Other Limitations
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Broader Perspective, Ethics and Competing Interests.
   (g) Did you discuss any potential misuse of your work? No, because the potential risk of misuse is minimal.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Methodology and Broader Perspective, Ethics and Competing Interests.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, see the Methodology.
   (b) Have you provided justifications for all theoretical results? Yes, see the Methodology.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see the Discussion and Conclusion.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see the Results, and Discussion and Conclusion.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see the Discussion and Conclusion
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see the Background and Related Work, and Discussion and Conclusion
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see the Discussion and Conclusion
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see (Ruffin et al. 2024) for the new assets (user survey questionnaire and images used in the survey).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see the Broader Perspective, Ethics, and Competing Interests statement.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Broader Perspective, Ethics, and Competing Interests statement.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA, the new assets are not datasets.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA, the new assets are not datasets.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, see the supplementary materials (Ruffin et al. 2024).
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see the Broader Perspective, Ethics, and Competing Interests.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, we provided the estimated hourly wage paid to participants (and how much each participant was compensated). The total amount can be calculated by multiplying it by the number of participants in the study. See the Methodology (Recruitment) for details.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see the Methodology, and Broader Perspective, Ethics, and Competing Interests.