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

Jaron Mink, Licheng Luo, Natã M. Barbosa, Olivia Figueria, Yang Wang, Gang Wang
Deceptive Profiles are Prevalent
Deceptive Profiles are Prevalent

Digital war: How Russia is using deep fakes in Ukraine for propaganda

That smiling LinkedIn profile face might be a computer-generated fake

Constructing Fake Profiles

Automation
Bots/Sybils

Human Labor
Human Workers
Constructing Social Profiles

Scalability

Quality

Bots/Sybils

Deepfakes

Machine Learning

Human Workers
I saw you on LinkedIn and thought you'd be the perfect fit for the team. I'm a freelance designer and have been a freelance designer for over 30 years. I'd be thrilled to meet you and share your creativity! 

Hi my name is Alice!

Deep Learning Network

I saw you on LinkedIn and thought you'd be the perfect fit for the team. I'm a freelance designer and has been a freelance freelance designer for over 30 years. I'd be thrilled to meet you and share your creativity!

"Hi my name is Alice!"

Deep Learning Network

Deepfakes in Social Engineering

1. Deepfake Artifacts

Want to connect?

Perception: Hrm...

Behavior: Yes!

2. User Knowledge
Research Questions

1. Do deepfake artifacts influence users’ perceived trustworthiness of a profile or decision to accept a connection request?

2. Does priming/training influence users’ perceived trustworthiness of a profile or decision to accept a connection request?

3. What strategies do users employ to assess a profile?
Participant Task

Scenario
• LinkedIn connection requests

Tutorial
• Prompt (between-subjects condition)

Review three profiles (HR/IT/Finance)
• Artifact (within-subjects condition)
• Rate Trust (0-100)
• Respond to Request (accept/reject)
• Provide reason (open-text)
Deepfake Artifacts: within-subjects
Deepfake Artifacts: within-subjects

2 Intra-field: Image
Using a profile image that contains artifacts

1 Consistent Profile

4 Intra-field: Text
Introducing grammatical or semantic textual errors in the bio description ("about")
Deepfake Artifacts: within-subjects

Intra-Field Artifacts

Using a profile that contains

**Consistent Profile**

- **Intra-field: Image**
  - Colorful Blobs
  - Distorted Accessories
  - Distorted Background
  - Asymmetry

- **Intra-field: Text**
  - Introducing grammatical or semantic textual errors in the bio description ("about")

**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.
Deepfake Artifacts: within-subjects

2 Intra-field: Image
Using a profile image that contains artifacts

3 Inter-field: Image
The age of the person in the profile image is not consistent with the person’s experience

1 Consistent Profile

4 Intra-field: Text
Introducing grammatical or semantic textual errors in the bio description ("about")

5 Inter-field: Text
Introducing relational errors between profile bio and other fields, such as incorrect reference to school, degree, and work experience
Deepfake Artifacts: within-subjects

Inter-Field Artifacts

About

Business Career Spans 11 years with experience in service, distribution, pension and health

About

Masters in Entrepreneurship from UT Austin... working as a Partner for Dashlane in San Fransico

Education

California University of Management and Science
Masters of Science, Computer Info. Systems

Experience

Senior Oracle Database Administrator
Bird Inc.
Feb 2017 - Present
Deepfake Artifacts: within-subjects

2 Intra-field: Image
Using a profile image that contains artifacts

3 Inter-field: Image
The age of the person in the profile image is not consistent with the person’s experience

1 Consistent Profile

4 Intra-field: Text
Introducing grammatical or semantic textual errors in the bio description ("about")

5 Inter-field: Text
Introducing relational errors between profile bio and other fields, such as incorrect reference to school, degree, and work experience
Prior Knowledge: between-subjects

| No Prompt | Soft Prompt | Hard Prompt |

...You have received a number of connection requests on LinkedIn... This is your first time meeting these users.
Prior Knowledge: between-subjects

No Prompt | **Soft Prompt** | Hard Prompt

One last thing... **some profiles may be fake!**

Artificial Intelligence (AI) can be used to generate realistic images and text of people that never existed. These profiles may be used to learn about you, influence your opinions, or scam you.
Prior Knowledge: between-subjects

<table>
<thead>
<tr>
<th>Colorful Blobs</th>
<th>Distorted Accessories</th>
<th>Distorted Background</th>
<th>Asymmetry</th>
</tr>
</thead>
</table>

**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.
Experimental Procedure

Roleplaying Scenario

- No Prompt
- Soft Prompt
- Hard Prompt

View a prompt condition

Accept or Reject connection requests from three profiles

Post-Task Survey

Recruited N=286 Amazon Mturk participants
User Acceptance of Deepfakes (RQ1)

* ** *** = statistically significant under binary mixed effect regression
User Acceptance of Deepfakes (RQ1)

*,**,*** = statistically significant under binary mixed effect regression
All deepfake artifacts significantly decrease user acceptance.
### User Acceptance of Deepfakes (RQ1)

All deepfake artifacts significantly decrease user acceptance.

Intra Image is less accepted than Inter Image.

* *, **, *** = statistically significant under binary mixed effect regression.
Artifact Attributability (RQ1)

Intra-Image Artifacts
- Easy to detect
- [No Prompt] Hard to attribute
- [Hard Prompt] Easy to attribute

No Prompt
“The profile looks excellent but I am wondering about her photo what’s going on with that?” – P43

Hard Prompt
“In this image there blop in her hand. It was an AI generated image.” – P117
Artifact Attributability (RQ1)

Intra-Image Artifacts
- Easy to detect
- [No Prompt] Hard to attribute
- [Hard Prompt] Easy to attribute

Inter-Image Artifacts
- Hard to detect
  - Noted by one participant

“The individual looks very young in his profile image. Doesn't really align with what I was expecting after reading their bio… – P133
User Acceptance of Deepfakes (RQ2)

*, **, *** = statistically significant under binary mixed effect regression
User Acceptance of Deepfakes (RQ2)

Just warning of deepfakes is enough for users to take protective behavior

No evidence that additional training further protects users

*,**,*** = statistically significant under binary mixed effect regression
User Acceptance of Deepfakes (RQ1/RQ2)
User Acceptance of Deepfakes (RQ1/RQ2)

> 43% of participants accept requests
User Trust of Deepfakes (RQ1/RQ2)

See paper for details!
Users’ Profile Searching Strategies (RQ3)

Personal Qualities (28%)
- Aptitude
- Personality

Inconsistencies (31%)
- “Bot-like” behavior
  - Intra/Inter-field artifacts
  - Generic descriptions
- “Dishonest” behavior

Decreased w/ prompting
- None: 44%, Soft: 23%, Hard: 16%

Increased w/prompting
- None: 16%, Soft: 35%, Hard: 43%
Anti-social Behaviors (RQ3)

Prompted users occasionally perceive deepfake artifacts within real images/text

“There is something wrong with the applicant’s photo...the detail of one of her shoulders is impossible” – P224

Prompted users occasionally perceive artifacts that stem from *racial and gender stereotypes*

“the picture shows a Black woman but the name seems to belong to a White man” – P68

“Chris P.”
Platform Recommendations

- **Do Not Adopt Profile Warnings (Yet)**
- **Take Action Before Profiles Reach Users**
- **Conduct Deepfake Measurements**
Summary

Deepfake Artifacts
• Decrease profile acceptance
• Vary in perceptibility

User Prompting/Training
• May protect against some threats
• Still leaves a large user-base vulnerable

User Behavior
• Is affected by prompting/training
• May utilize harmful strategies

Jaron Mink, Licheng Luo, Natã M. Barbosa, Olivia Figueria, Yang Wang, Gang Wang