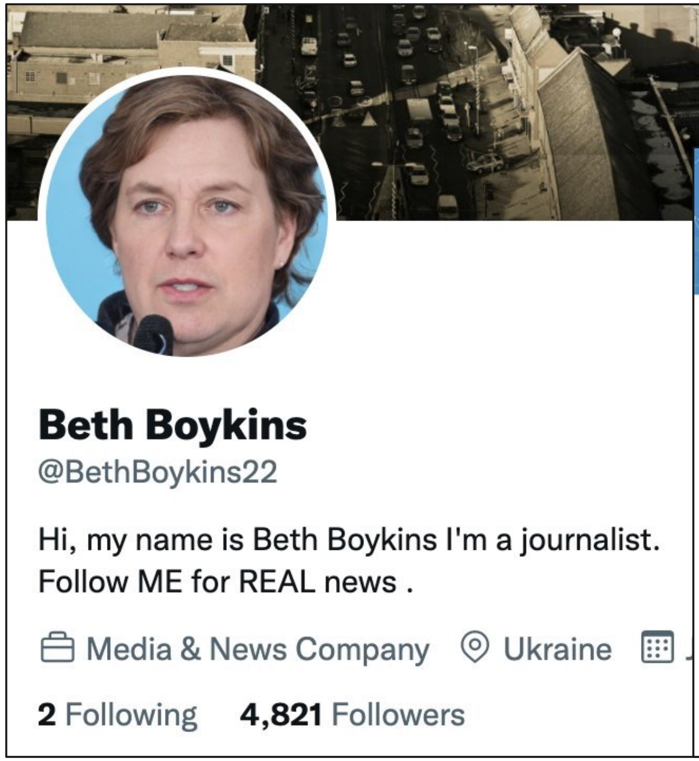
# DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks

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

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Santa Clara University

# Deceptive Profiles are Prevalent

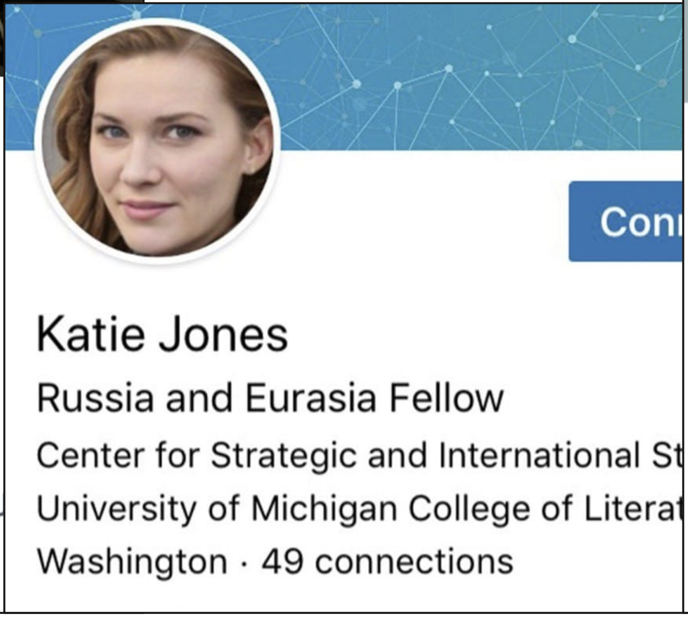


**Beth Boykins**
@BethBoykins22

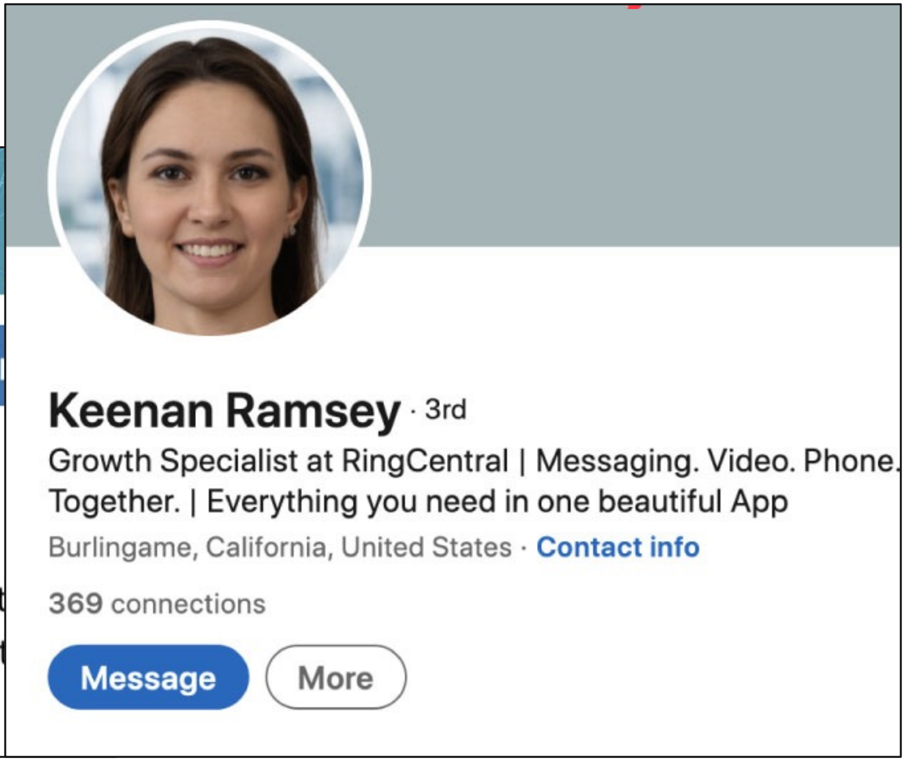
Hi, my name is Beth Boykins I'm a journalist. Follow ME for REAL news .

Media & News Company · Ukraine

**2** Following **4,821** Followers

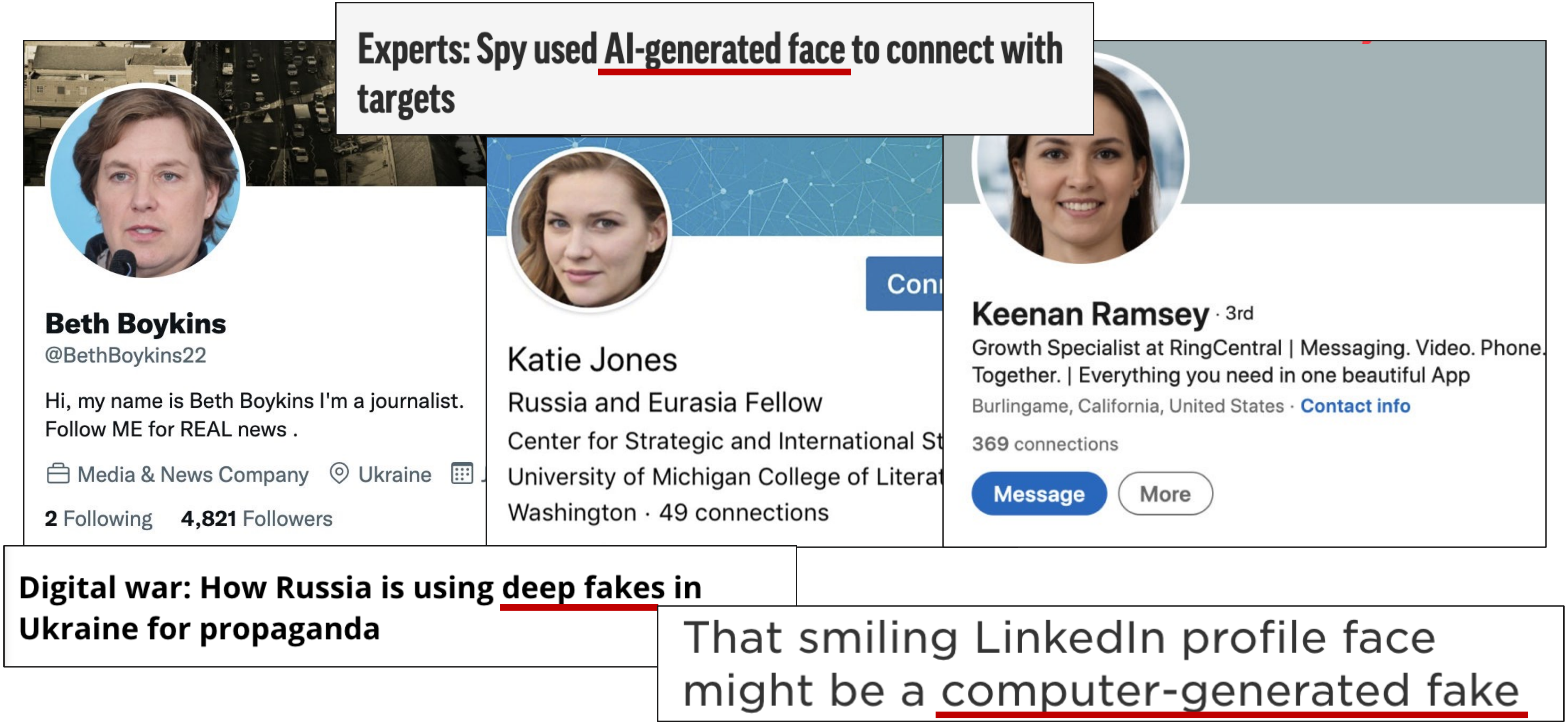


Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International St
University of Michigan College of Literat
Washington · 49 connections



**Keenan Ramsey** · 3rd
Growth Specialist at RingCentral | Messaging. Video. Phone. Together. | Everything you need in one beautiful App
Burlingame, California, United States · Contact info
369 connections

# Deceptive Profiles are Prevalent



Experts: Spy used AI-generated face to connect with targets

Beth Boykins
@BethBoykins22
Hi, my name is Beth Boykins I'm a journalist. Follow ME for REAL news .
Media & News Company · Ukraine
2 Following 4,821 Followers

Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International St
University of Michigan College of Literat
Washington · 49 connections

Keenan Ramsey · 3rd
Growth Specialist at RingCentral | Messaging. Video. Phone. Together. | Everything you need in one beautiful App
Burlingame, California, United States · Contact info
369 connections

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

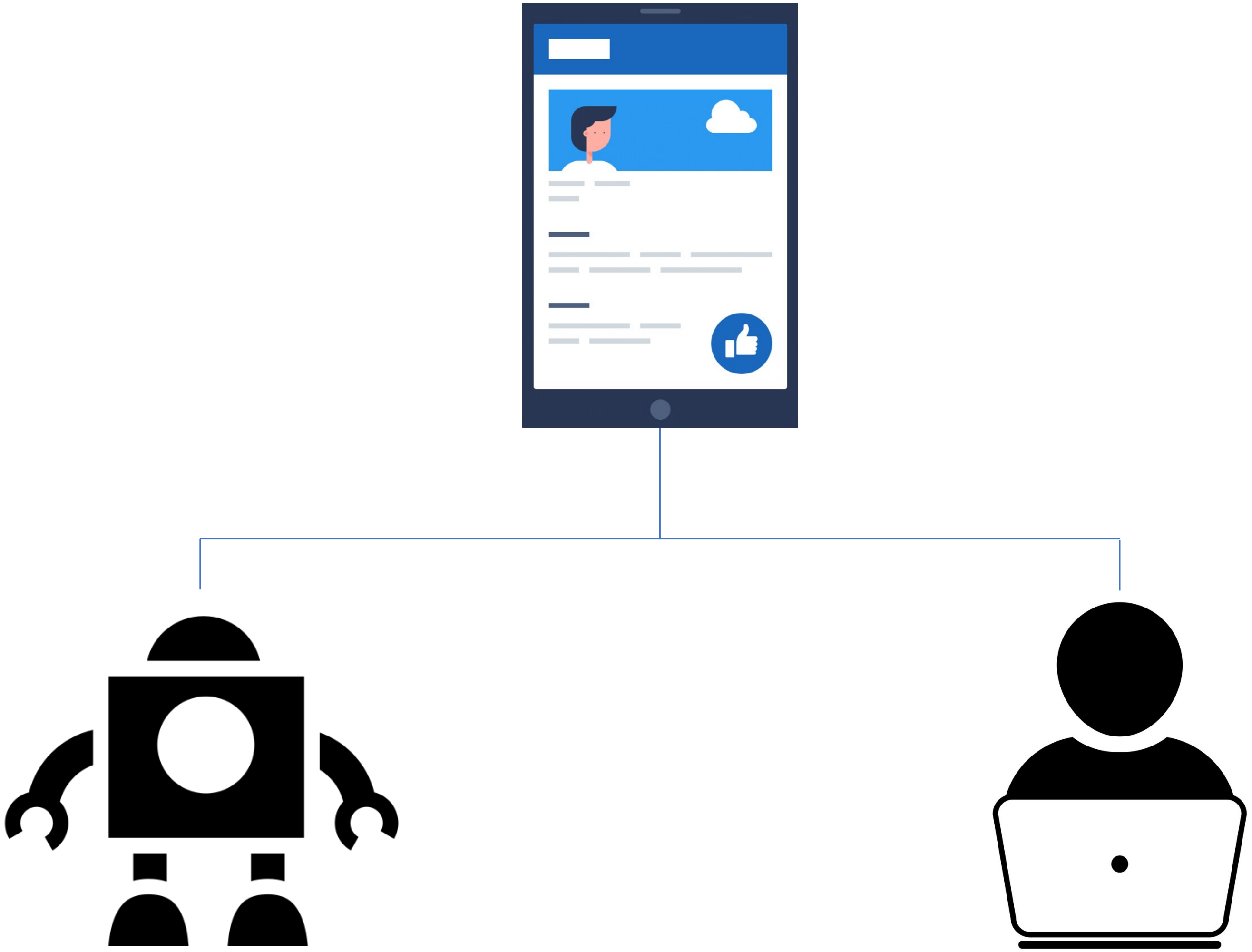
That smiling LinkedIn profile face might be a computer-generated fake

[1] Aparna Banerjea. 2022. *Digital war: How Russia is using deep fakes in Ukraine for propaganda.* Business Today. https://tinyurl.com/2p5jftuh

[2] Raphael Satter. 2019. *Experts: Spy used AI-generated face to connect with targets*. AP News. https://tinyurl.com/npkfp8fh

[3] Shannon Bond. 2022. *That smiling LinkedIn profile face might be a computer-generated fake*. Georgia Public Broadcasting. https://tinyurl.com/mj9yes

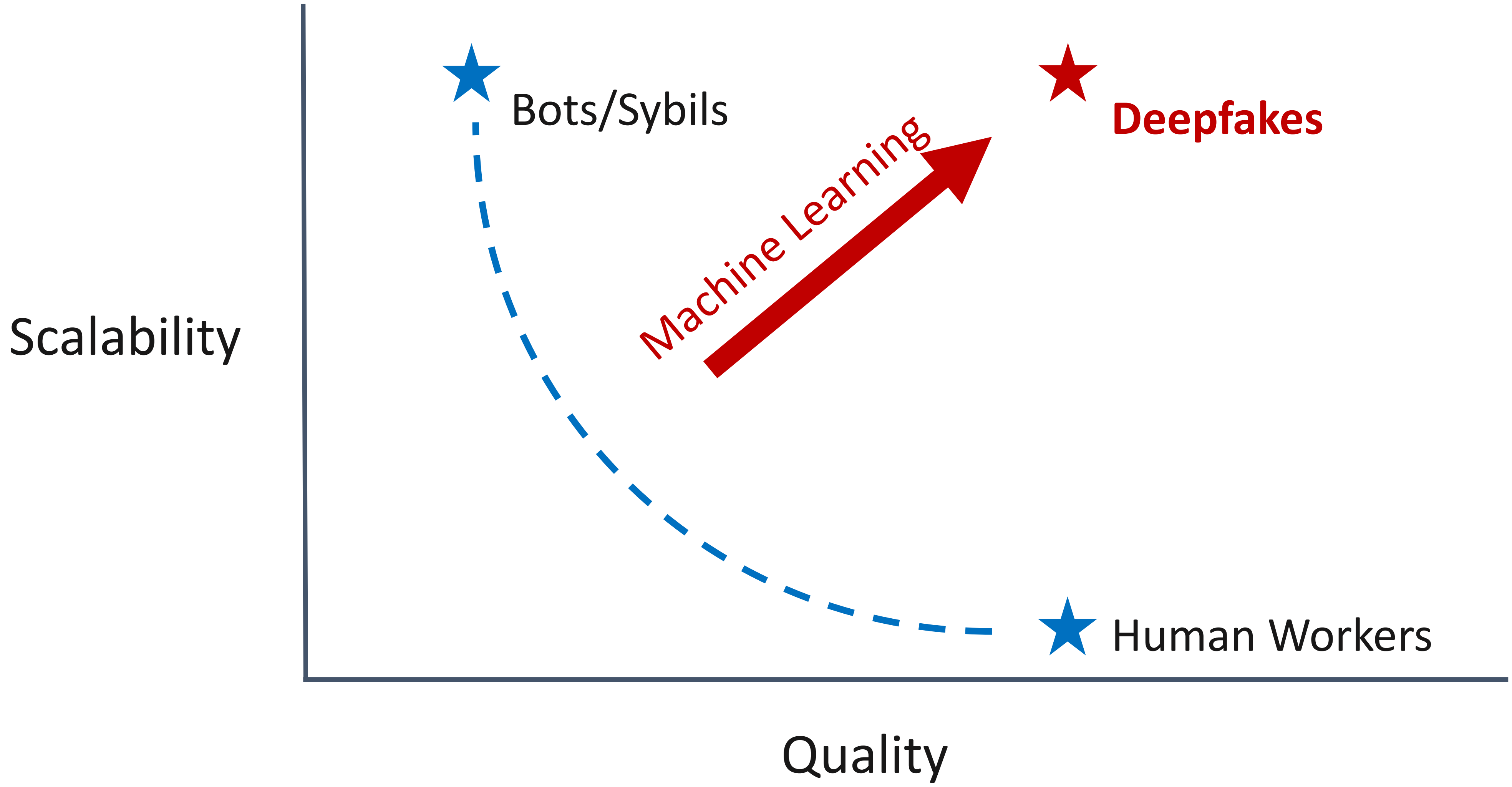
# Constructing Fake Profiles



**Automation**
Bots/Sybils

**Human Labor**
Human Workers

# Constructing Social Profiles



# Generative Deepfakes

**“Hi my name is Alice!**

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!. ”

StyleGAN2/3[1]

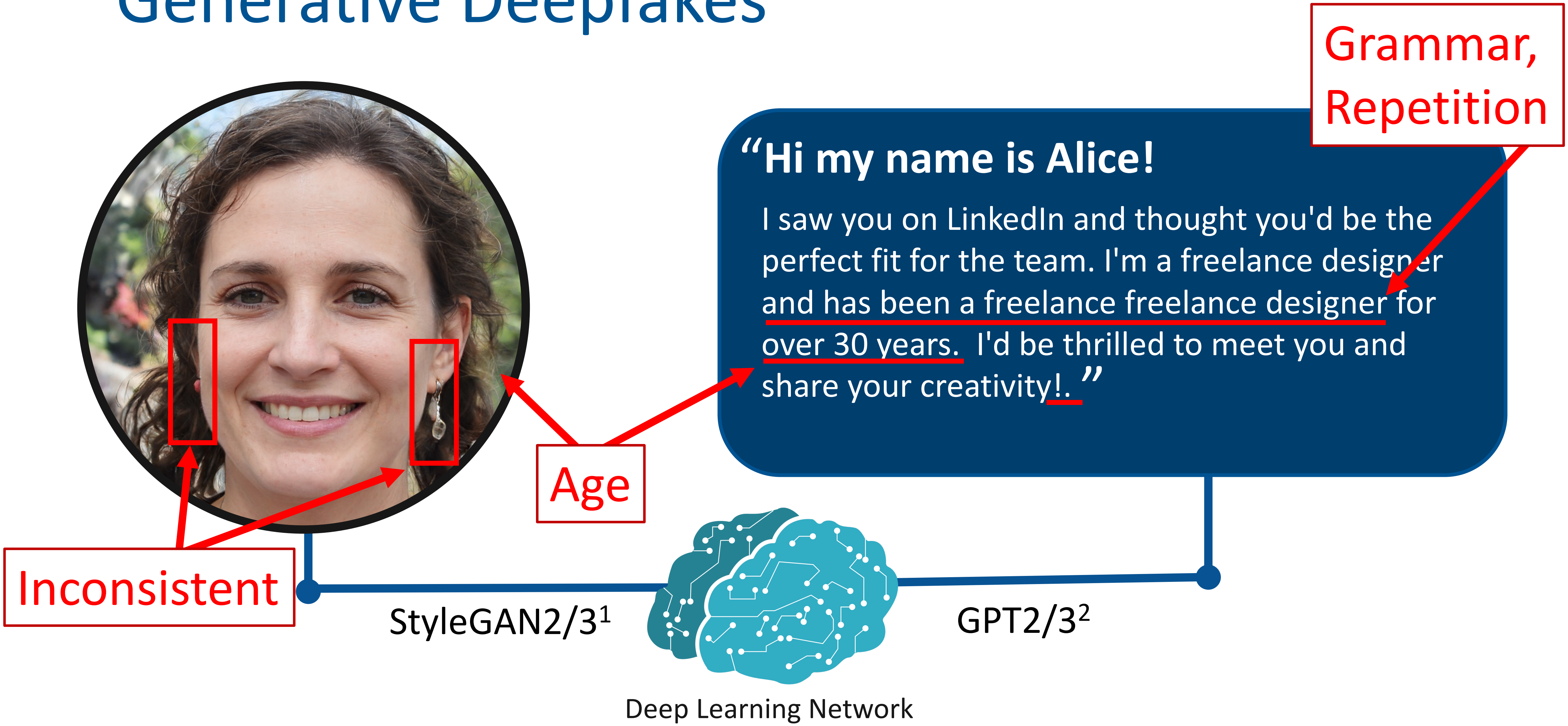
GPT2/3[2]

Deep Learning Network

[1] Karras, Tero et al. “Alias-free generative adversarial networks.” *Proceedings of NeurIPS*. 2021.

[2] Brown, Tom, et al. "Language models are few-shot learners."  *Proceedings of NeurIPS*. 2020.

# Generative Deepfakes



“Hi my name is Alice!

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!. ”

Grammar, Repetition

Age

Inconsistent

StyleGAN2/3[1]

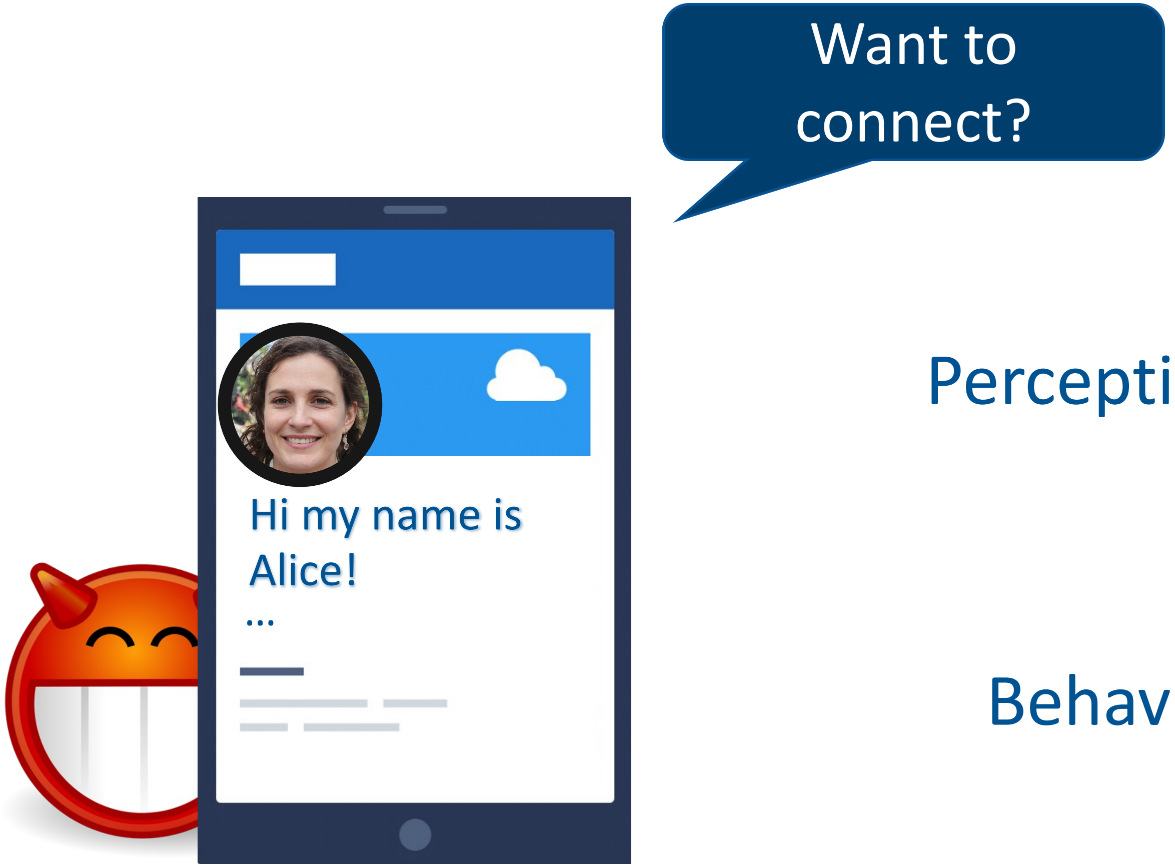
GPT2/3[2]

Deep Learning Network

[1] Karras, Tero et al. “Alias-free generative adversarial networks.” *Proceedings of NeurIPS*. 2021.

[2] Brown, Tom, et al. "Language models are few-shot learners."  *Proceedings of NeurIPS*. 2020.

# Deepfakes in Social Engineering



Perception:

Behavior:

1. Deepfake Artifacts

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

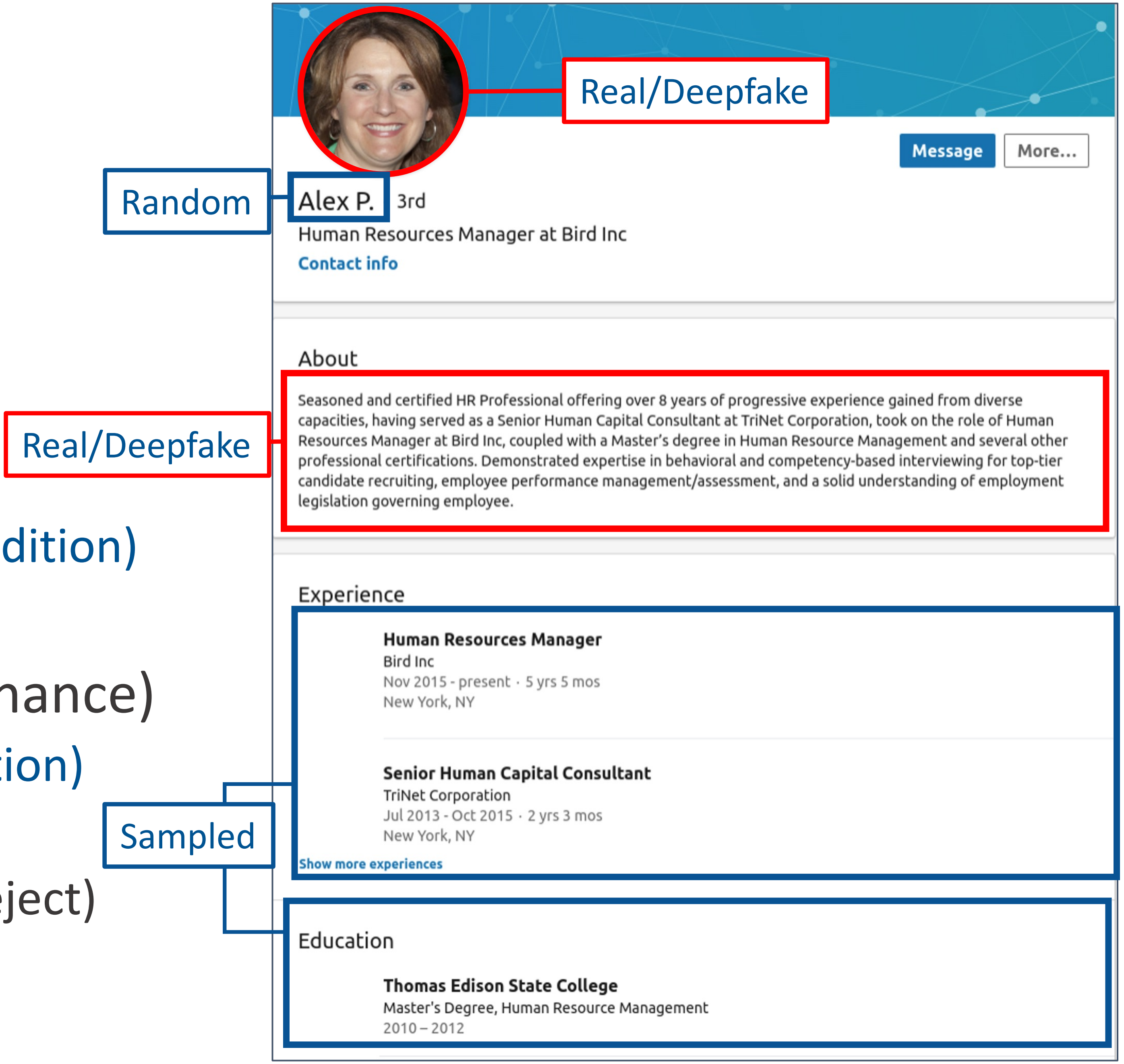
- LinkedIn connection requests

## 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)



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

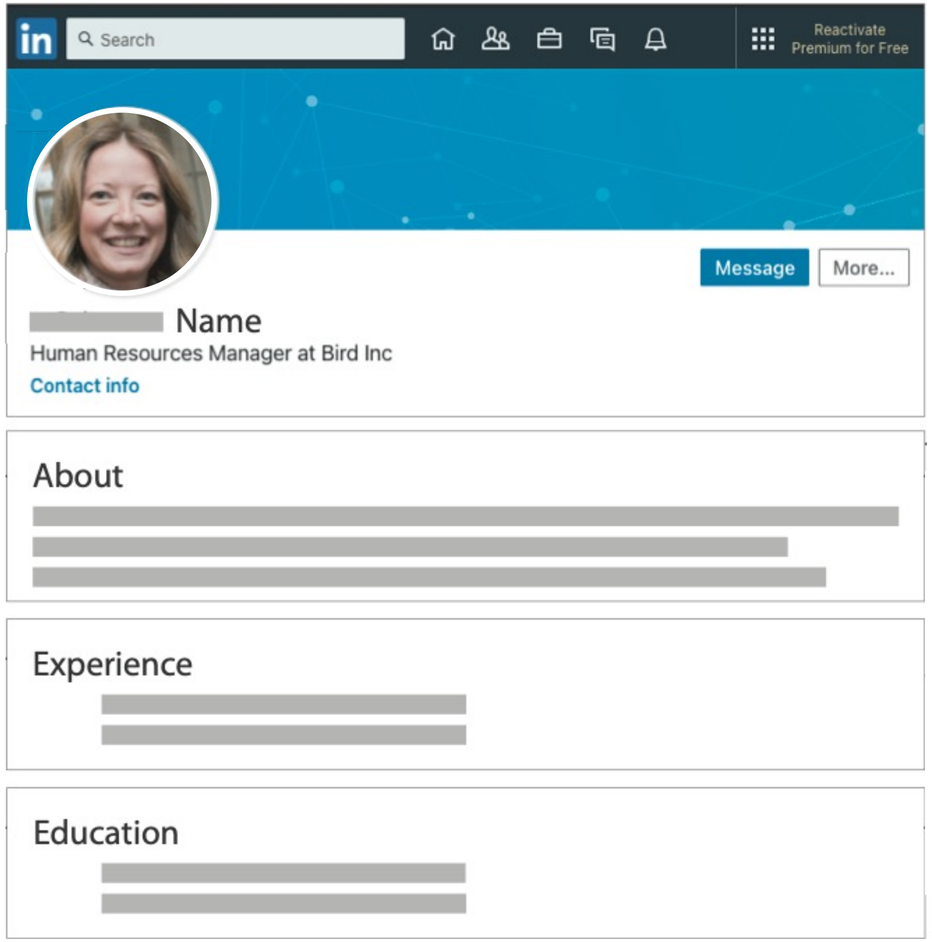
Show more experiences

Education

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

# Deepfake Artifacts: within-subjects

## ❶ Consistent Profile



[redacted] Name

Human Resources Manager at Bird Inc

Contact info

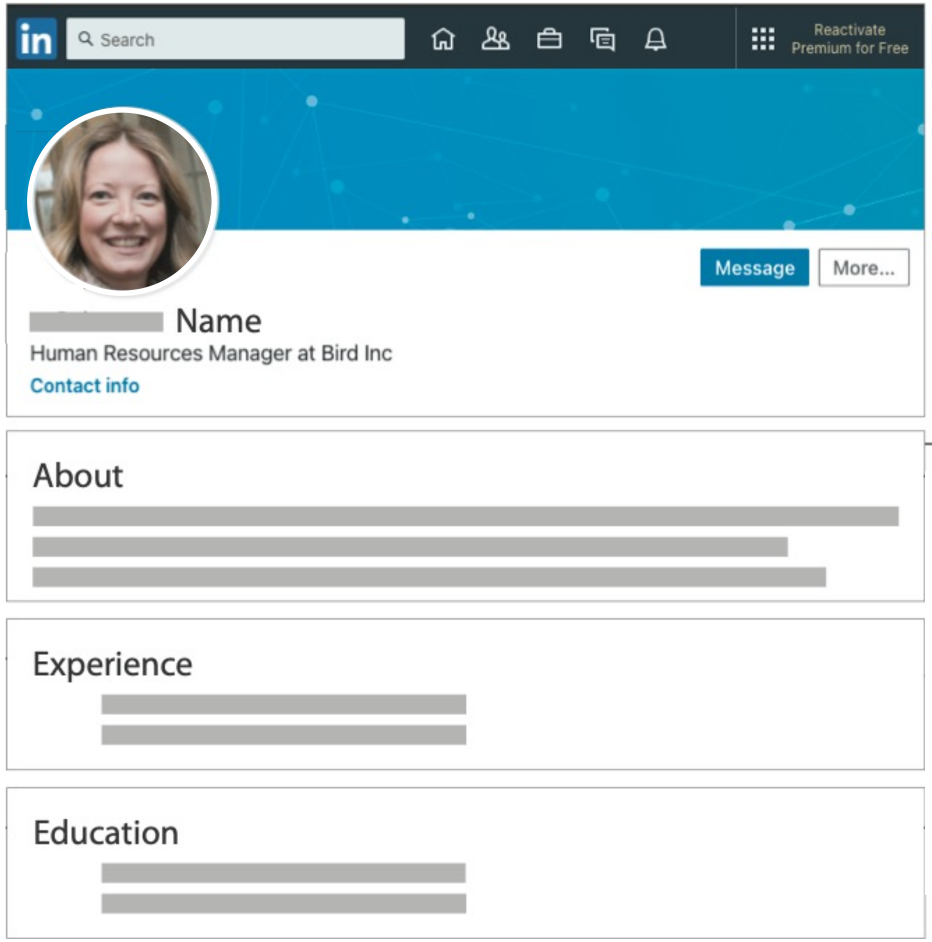
About

Experience

Education

# Deepfake Artifacts: within-subjects

## ❶ Consistent Profile



Name

Human Resources Manager at Bird Inc

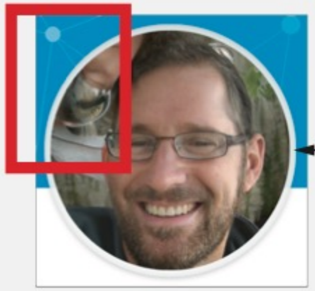
Contact info

About

Experience

Education

## ❷ Intra-field: Image



Using a profile image that contains artifacts

## ❹ Intra-field: Text

Introducing grammatical or semantic textual errors in the bio description (”about”)

# Intra-Field Artifacts

Deepfake Artifacts: within-subjects

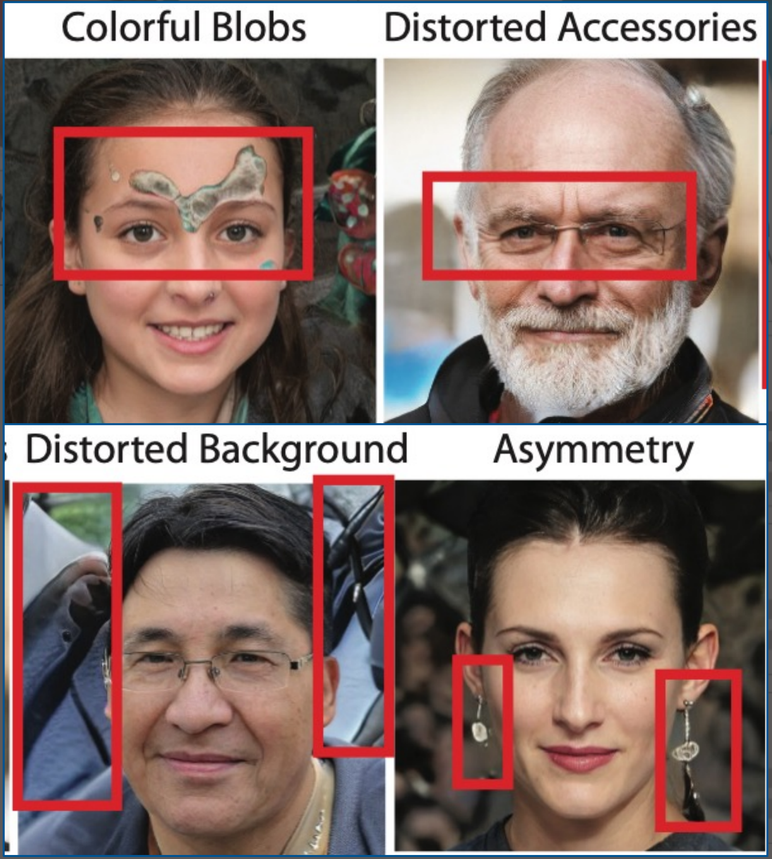
② Intra-field: Image

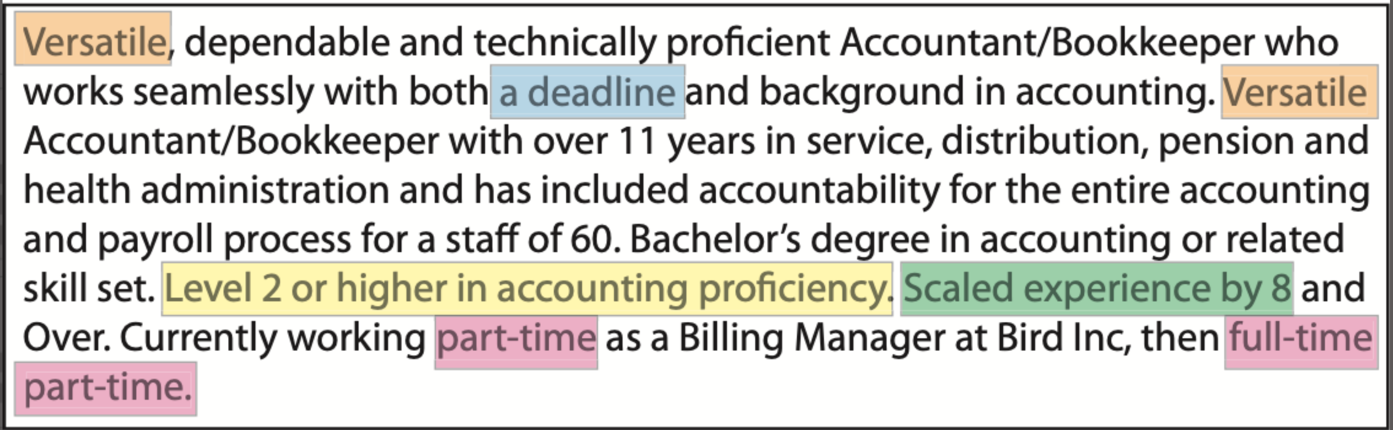
Using a p
that con

① Consistent Profile

④ Intra-field: Text

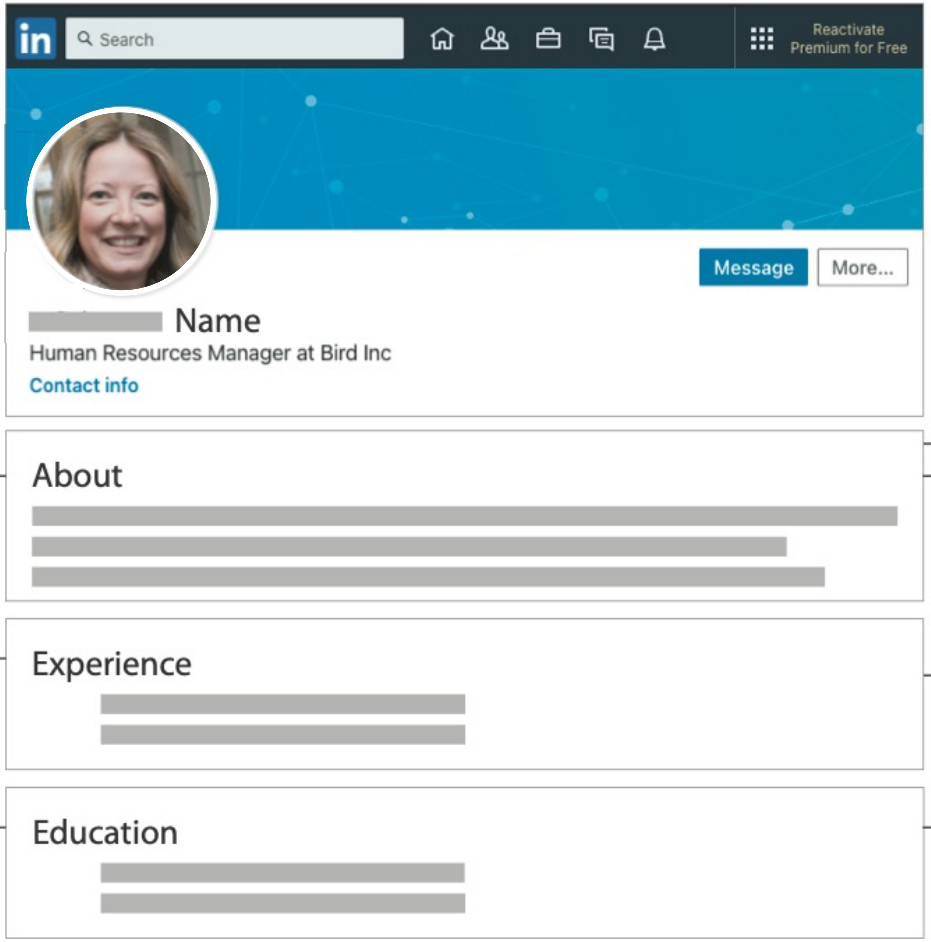
Introducing grammatical or semantic textual errors in the bio description ("about")




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

## 1 Consistent Profile



Name

Human Resources Manager at Bird Inc

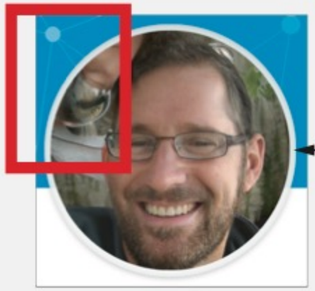
Contact info

About

Experience

Education

## 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

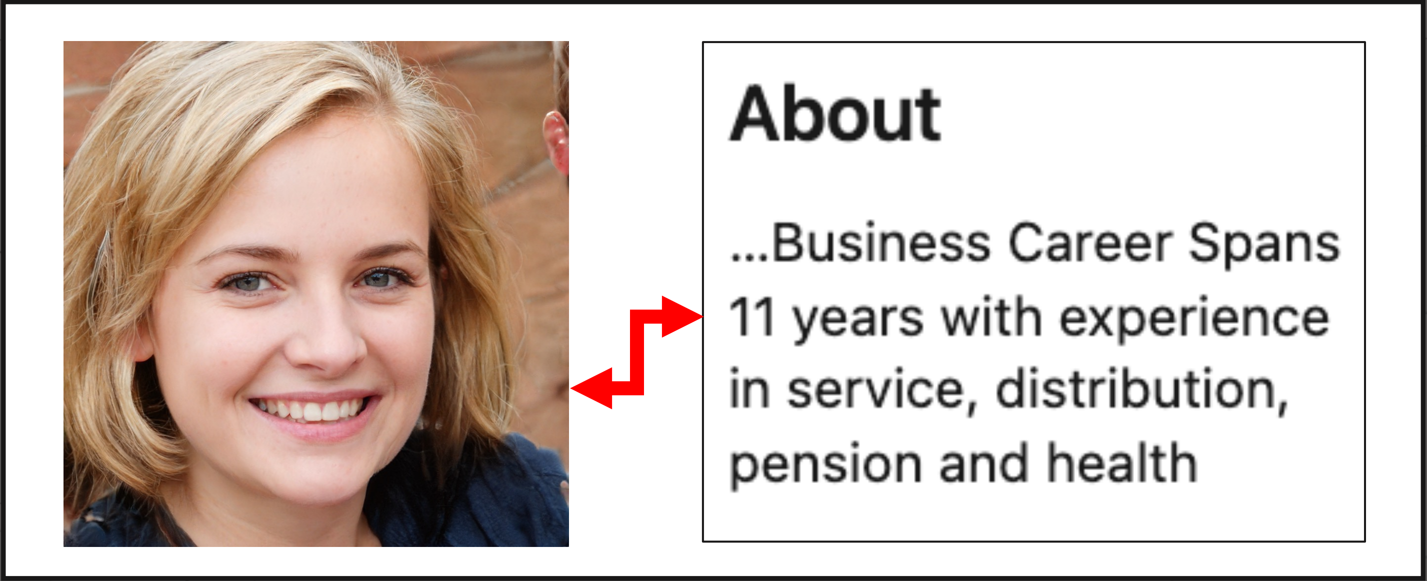
## 4 Intra-field: Text

Introducing grammatical or semantic textual errors in the bio description ("about")

## 5 Inter-field: Text

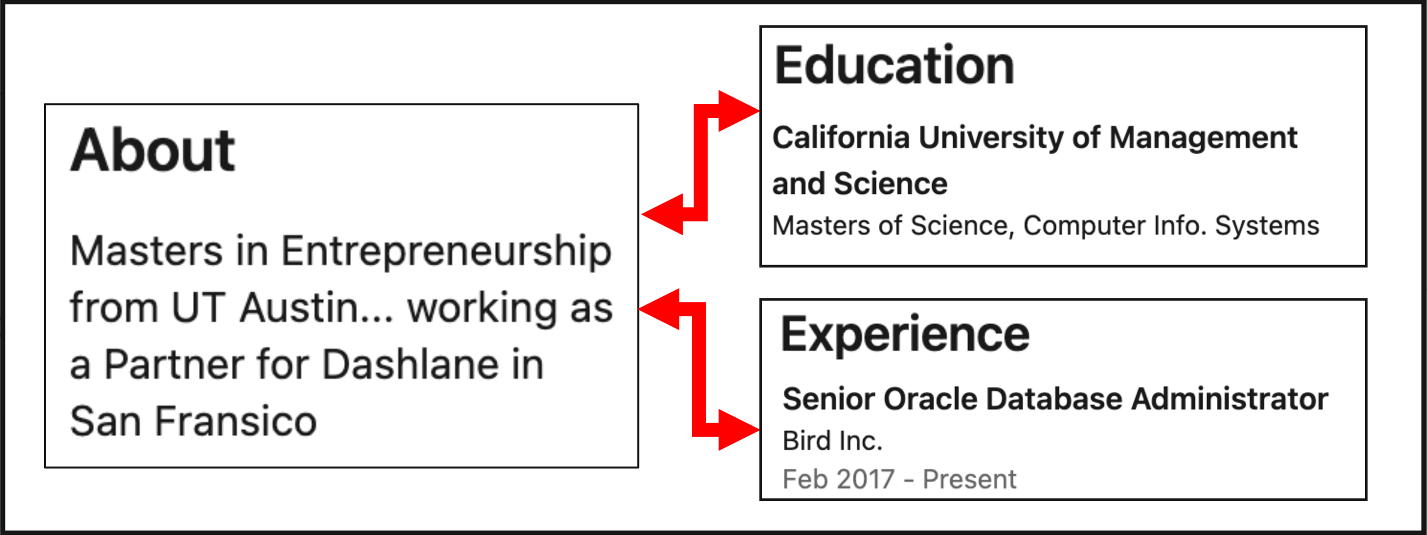
Introducing relational errors between profile bio and other fields, such as incorrect reference to school, degree, and work experience

# 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

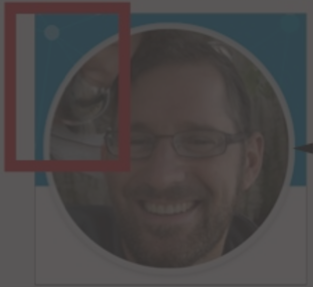
**Senior Oracle Database Administrator**

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 wit the person's experience

About

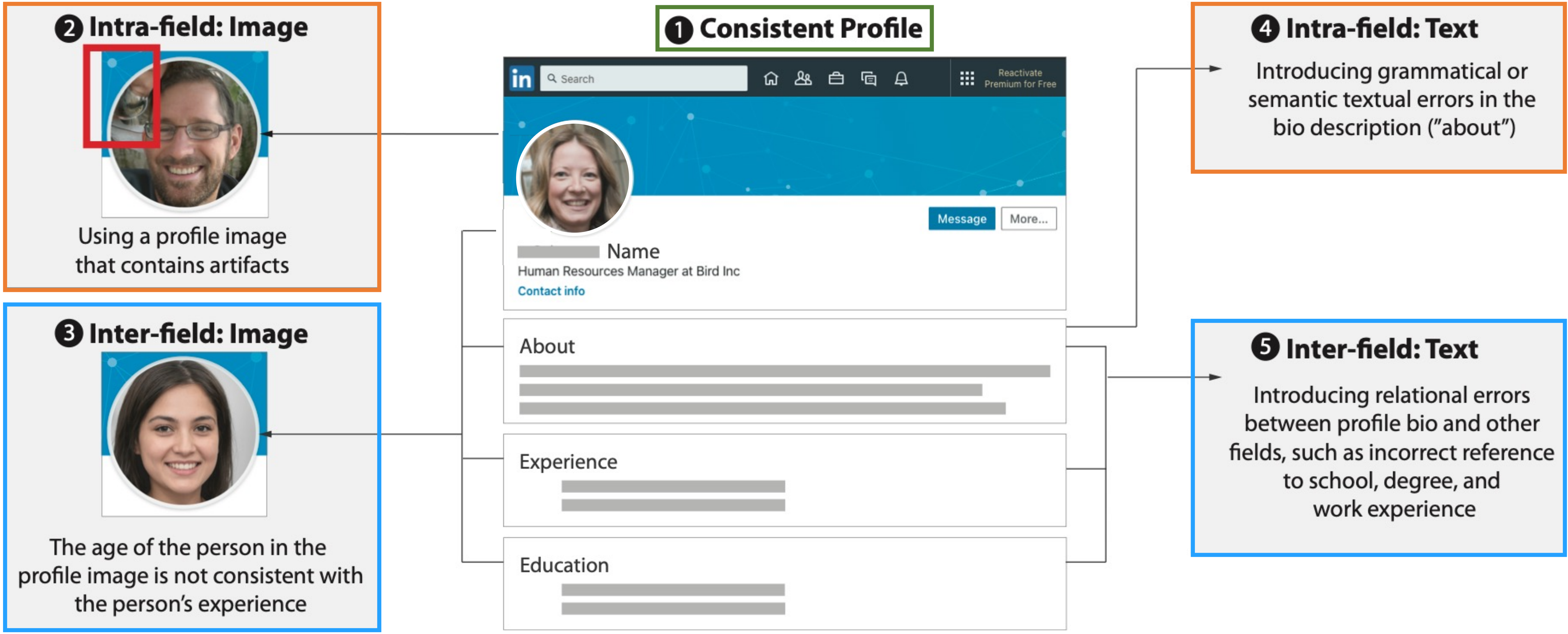
## 4 Intra-field: Text

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



❶ **Consistent Profile**

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

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

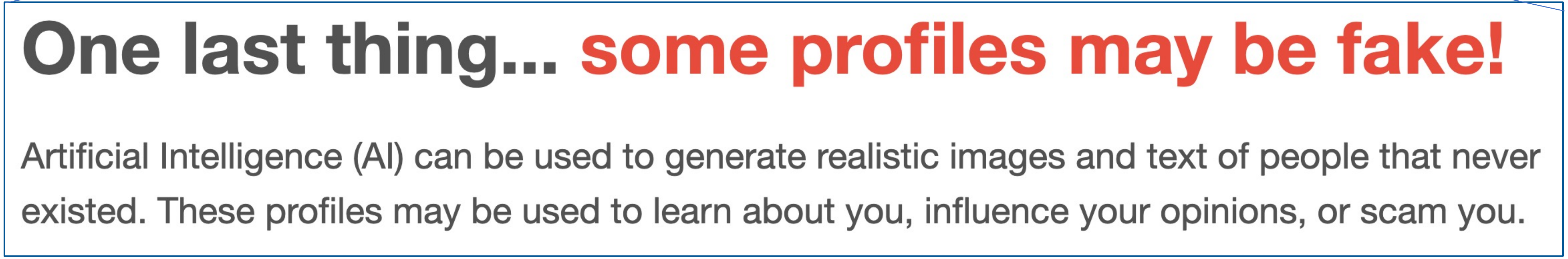
# Prior Knowledge: between-subjects

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

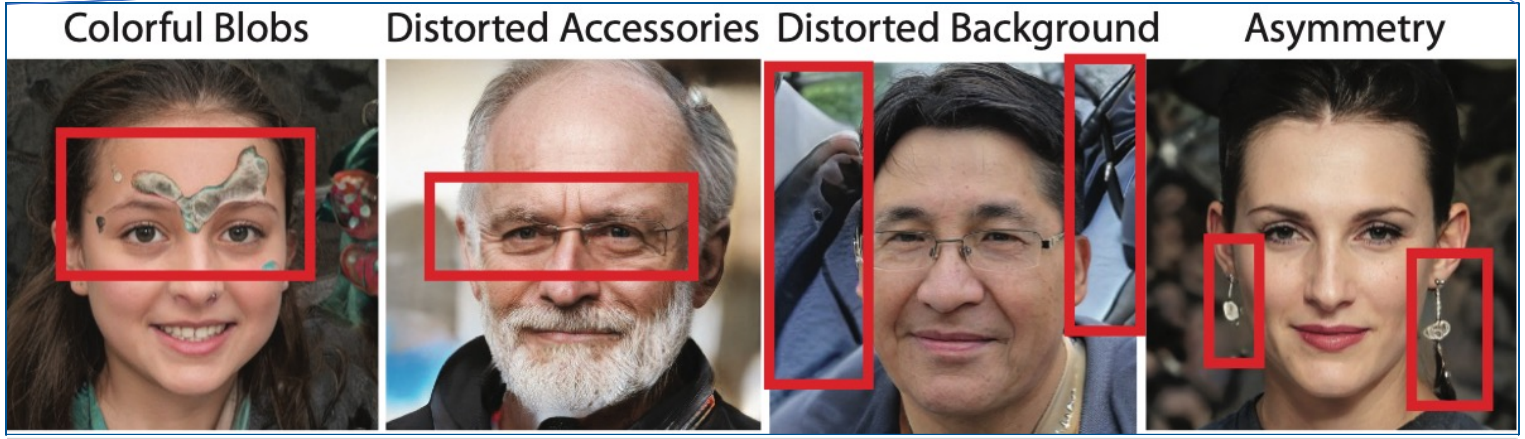


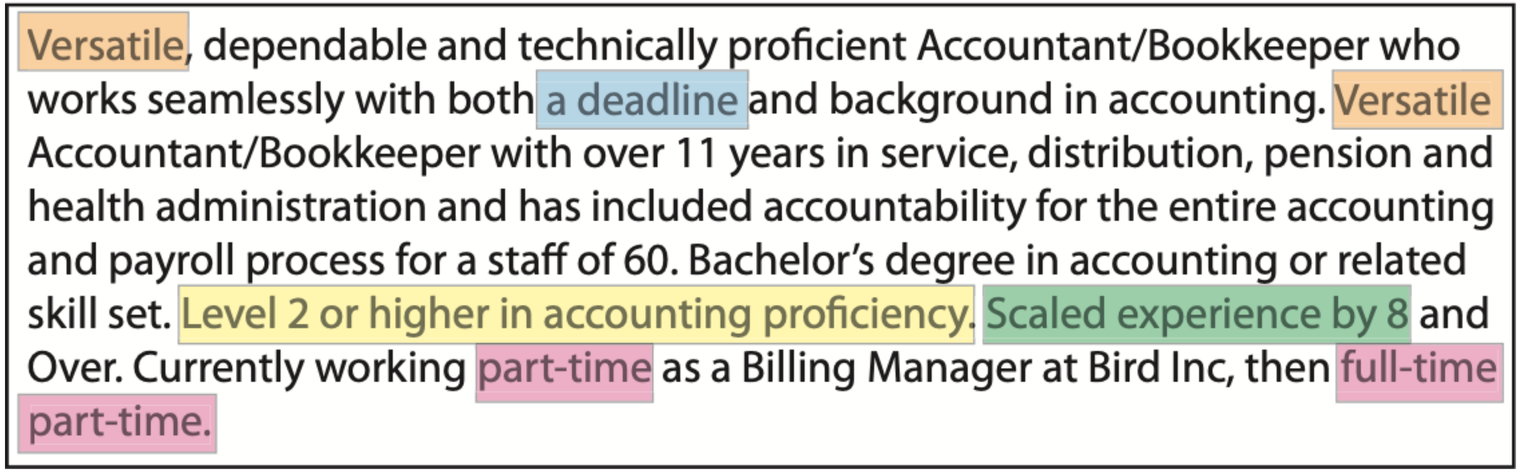
## One last thing... some profiles may be fake!

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

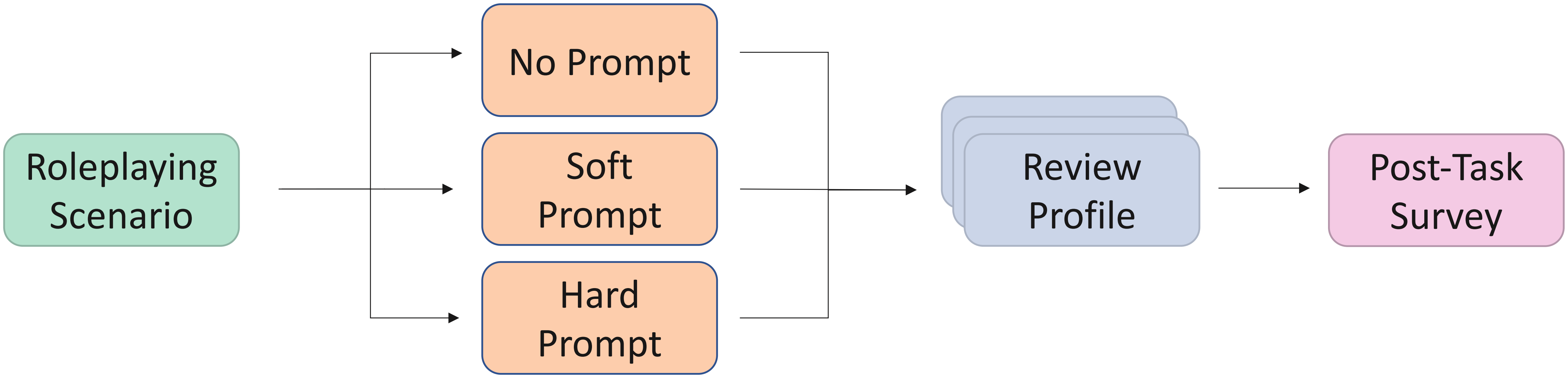
No Prompt | Soft Prompt | Hard Prompt





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



**Roleplay a LinkedIn user**

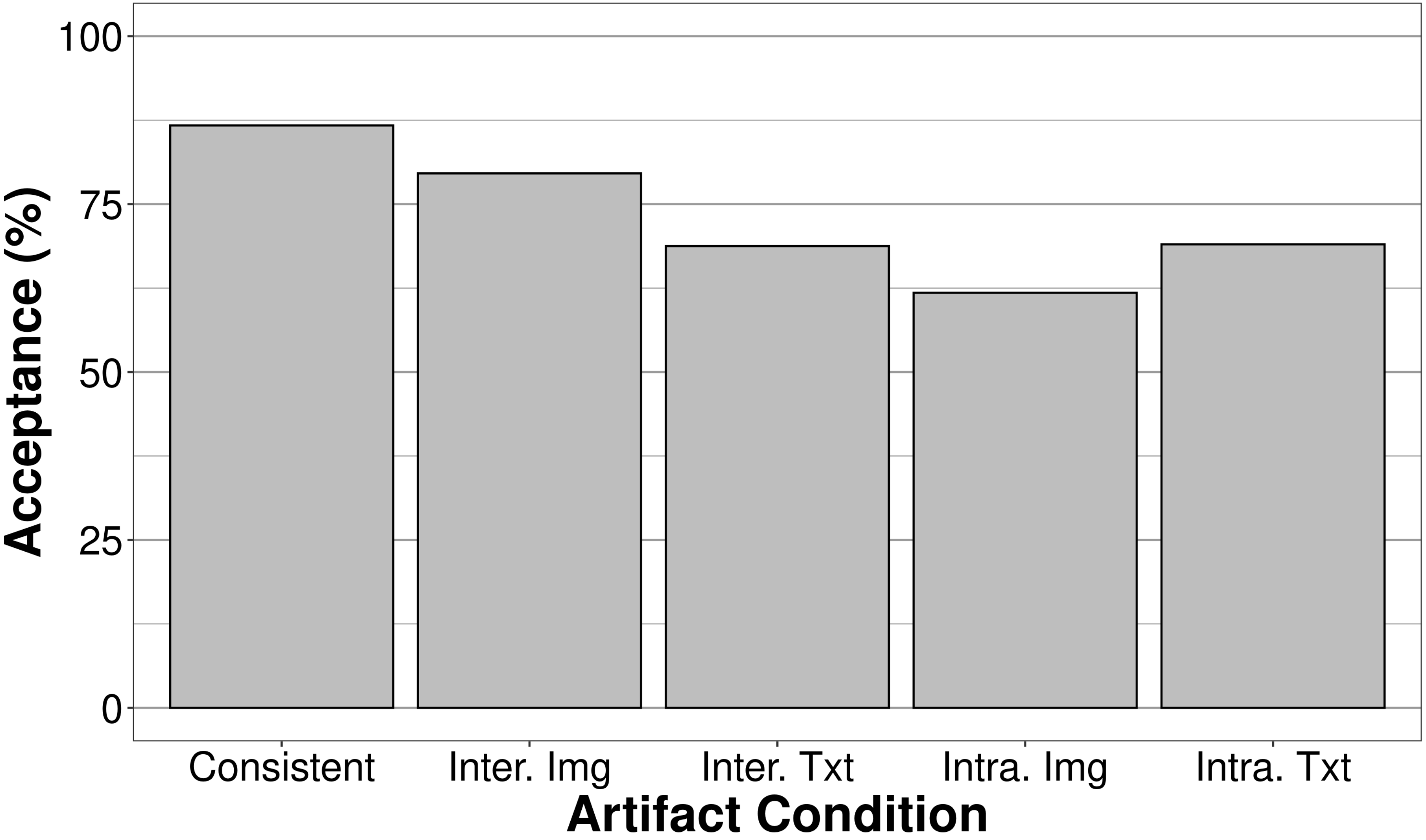
**View a prompt condition**

**Accept or Reject connection requests from three profiles**

**Complete a post-task survey**

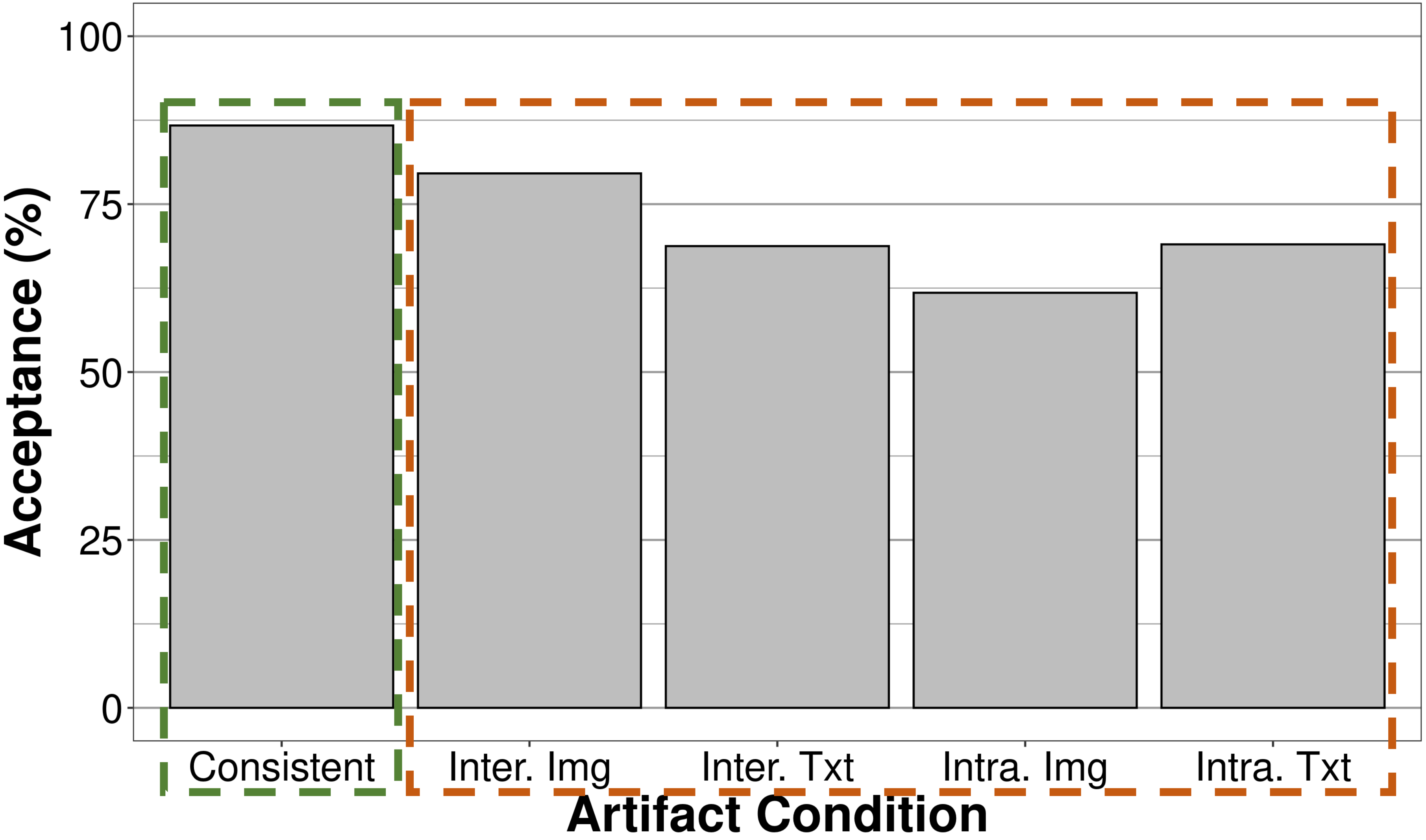
**Recruited N=286 Amazon Mturk participants**

# User Acceptance of Deepfakes (RQ1)



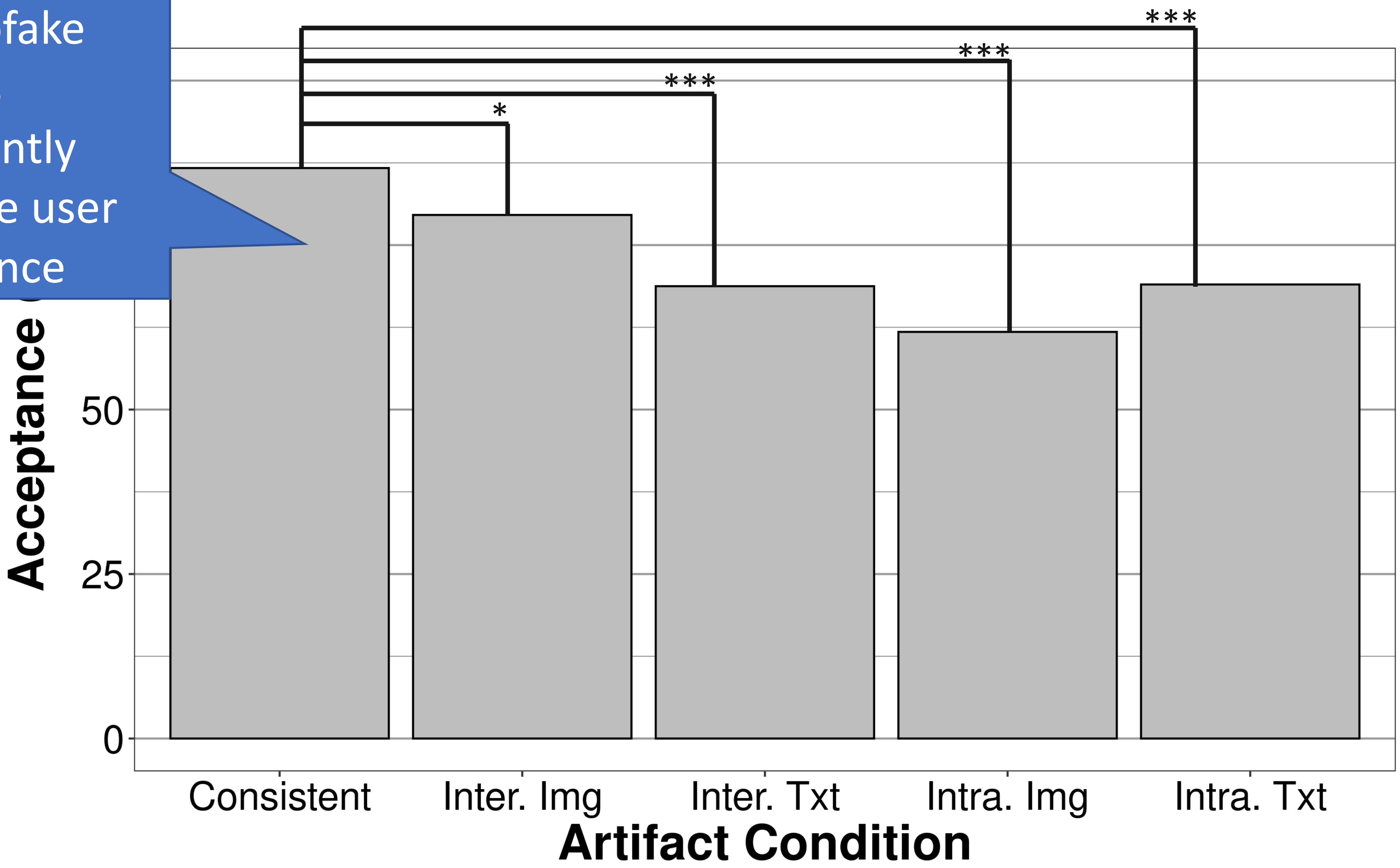
*,**,*** = statistically significant under binary mixed effect regression

# User Acceptance of Deepfakes (RQ1)



*,**,*** = statistically significant under binary mixed effect regression

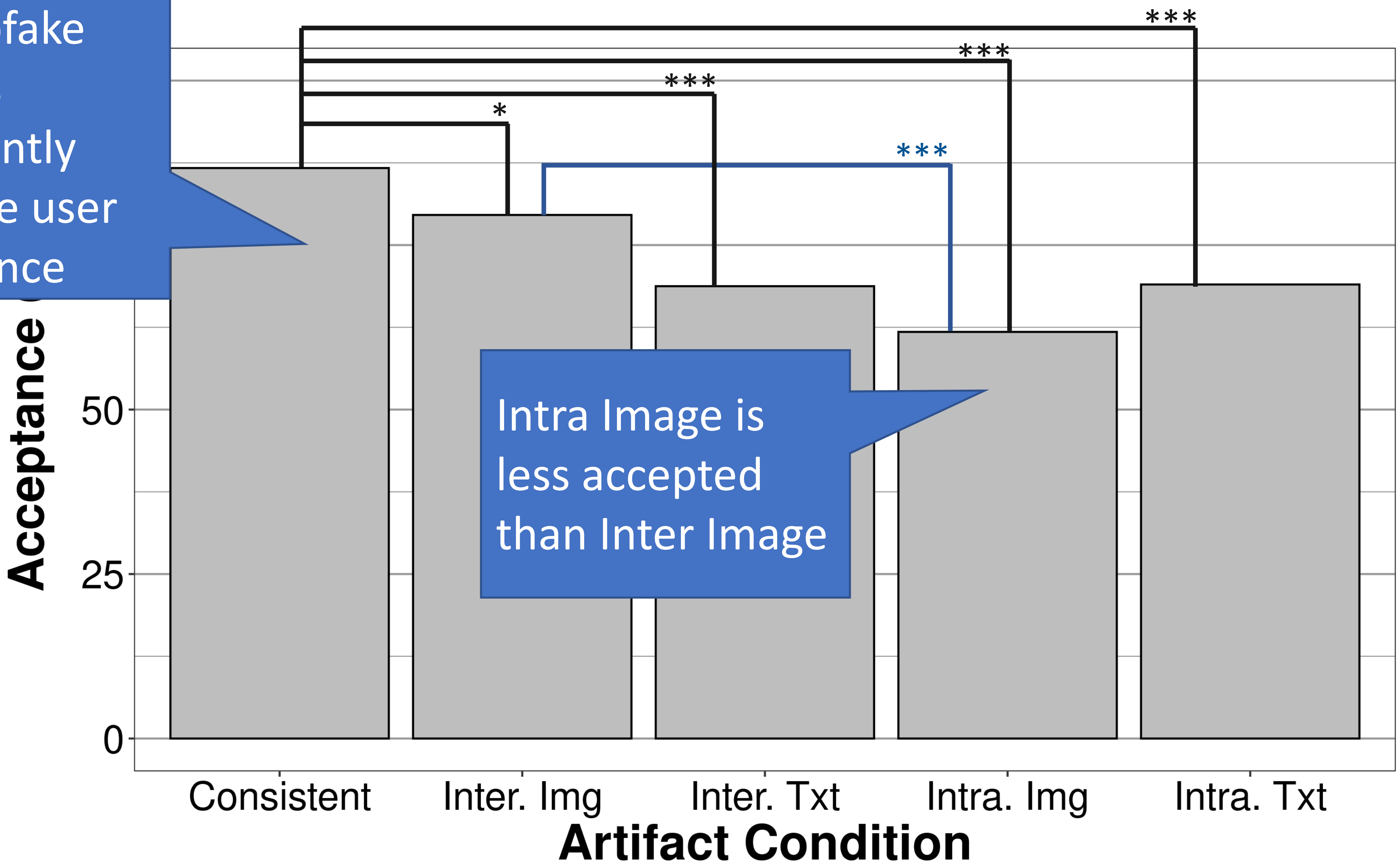
# User Acceptance of Deepfakes (RQ1)



All deepfake artifacts significantly decrease user acceptance

*,**,*** = statistically significant under binary mixed effect regression

# User Acceptance of Deepfakes (RQ1)

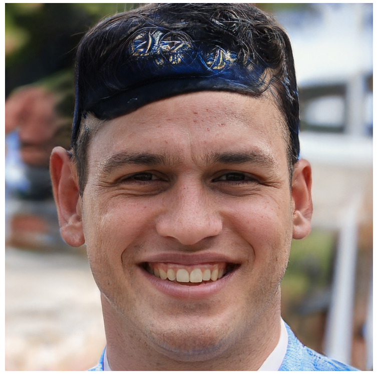


All deepfake artifacts significantly decrease user acceptance

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

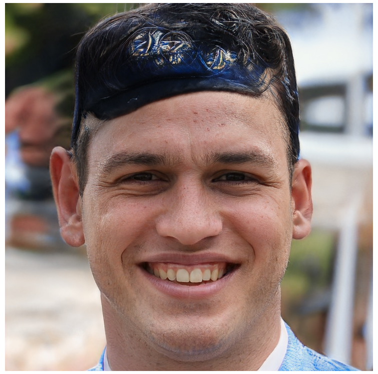
**No Prompt**

"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

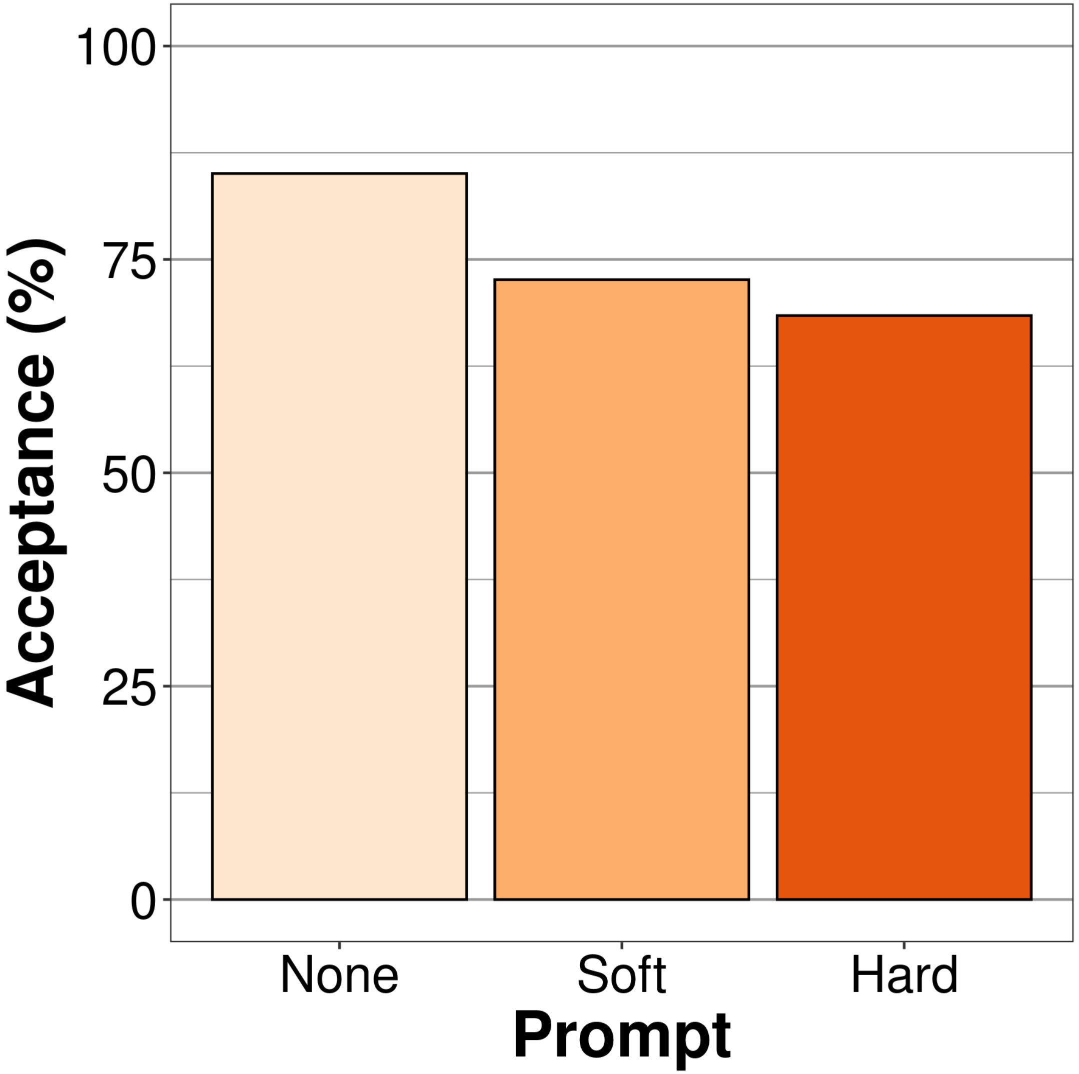
- Easy to detect
- [No Prompt] Hard to attribute
- [Hard Prompt] Easy to attribute

## Inter-Image Artifacts

- Hard to detect
  - Noted by one participant

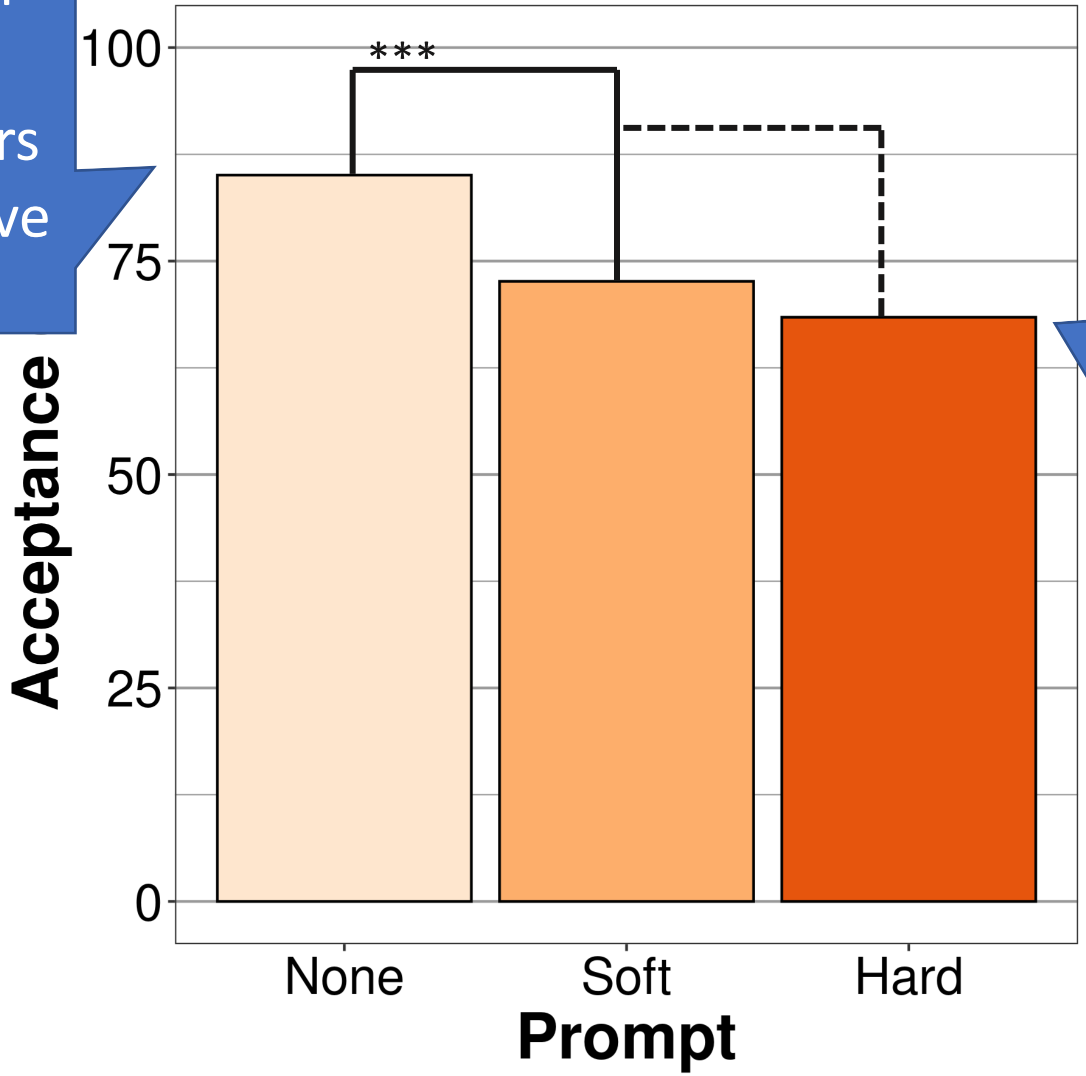
> “The individual looks very young in his profile image. Doesn't really align with what I was expecting after reading their bio… – P133

# User Acceptance of Deepfakes (RQ2)



*,**,*** = statistically significant under binary mixed effect regression

# User Acceptance of Deepfakes (RQ2)

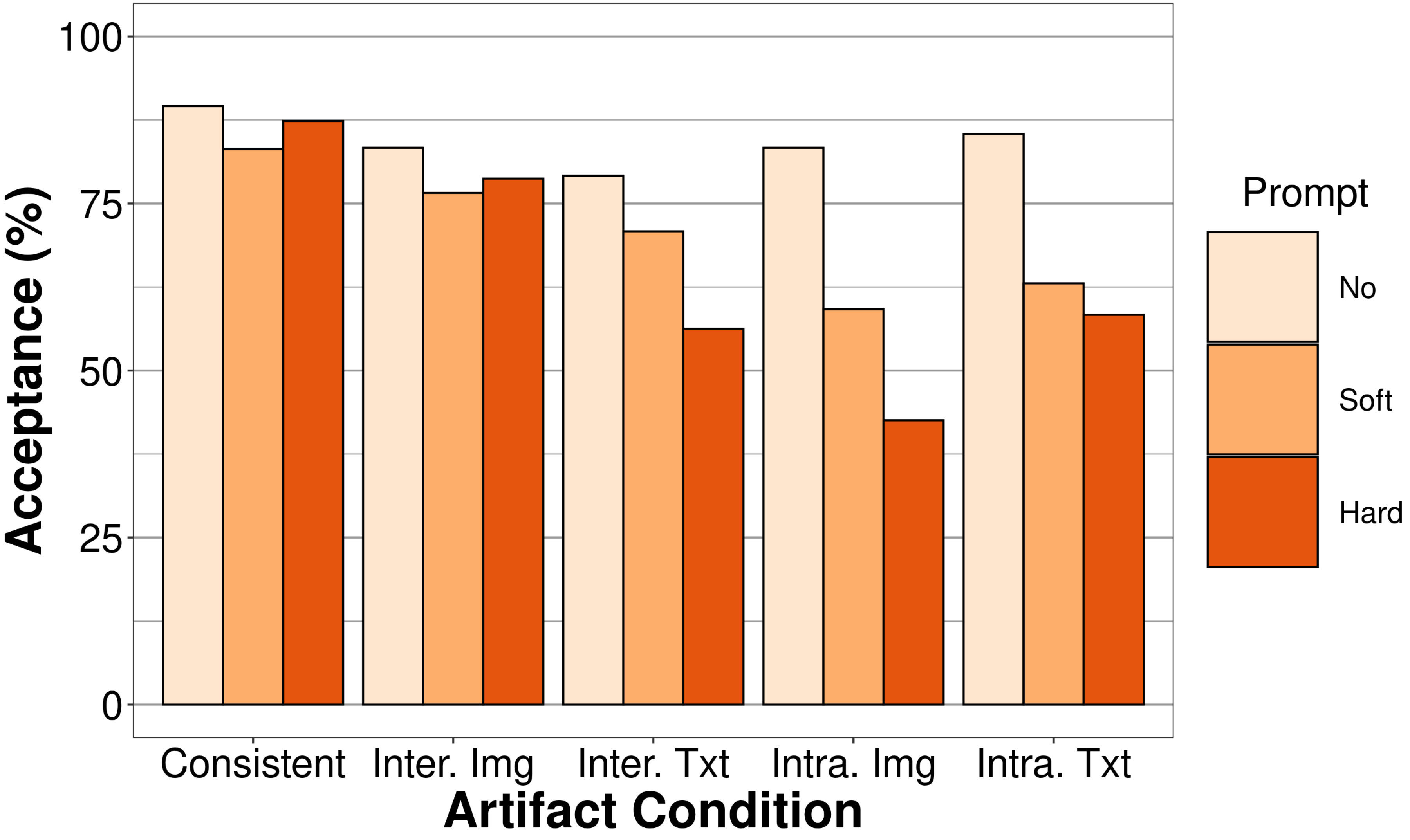


Just warning of deepfakes is enough for users to take protective behavior

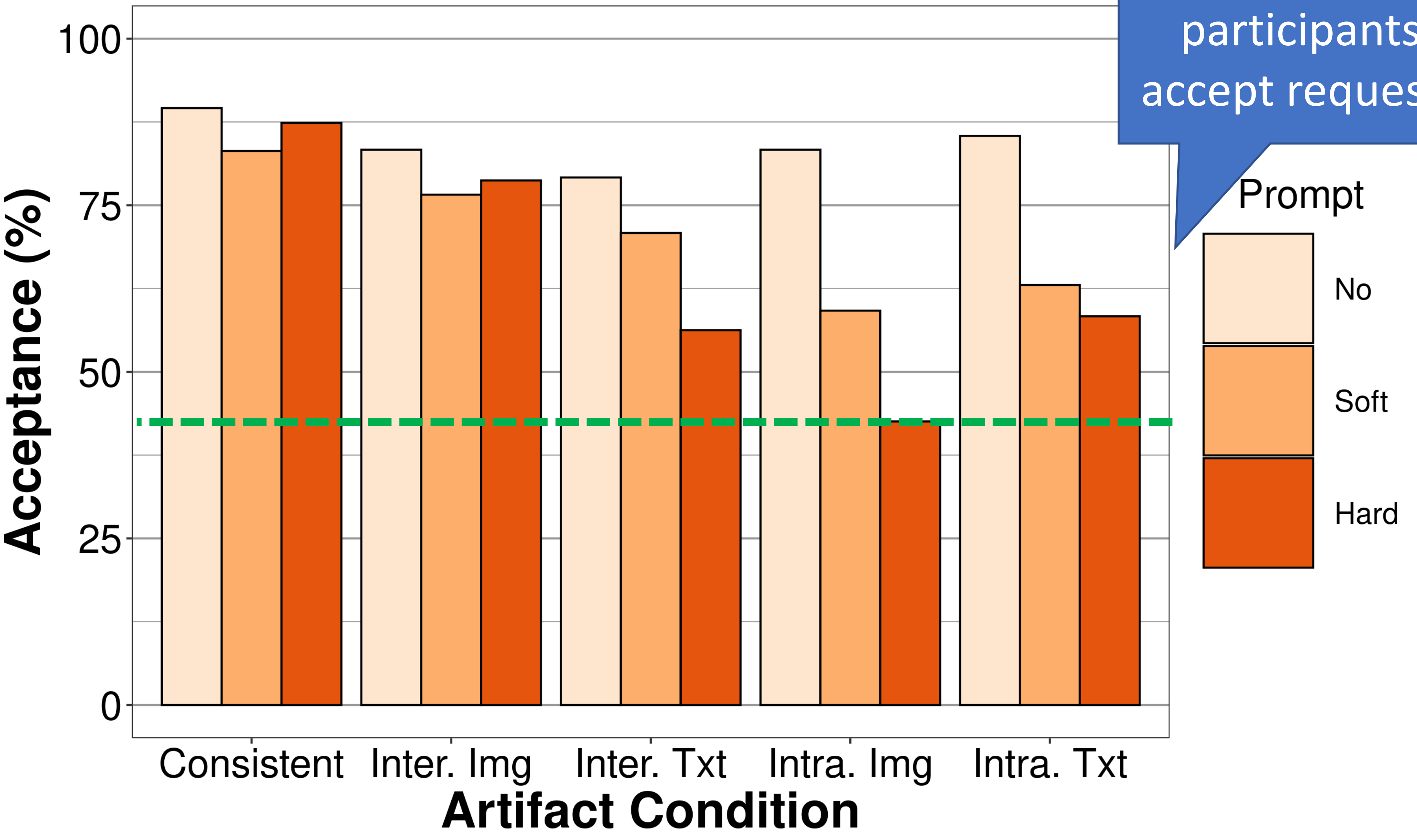
No evidence that additional training further protects users

*,**,*** = statistically significant under binary mixed effect regression

# User Acceptance of Deepfakes (RQ1/RQ2)

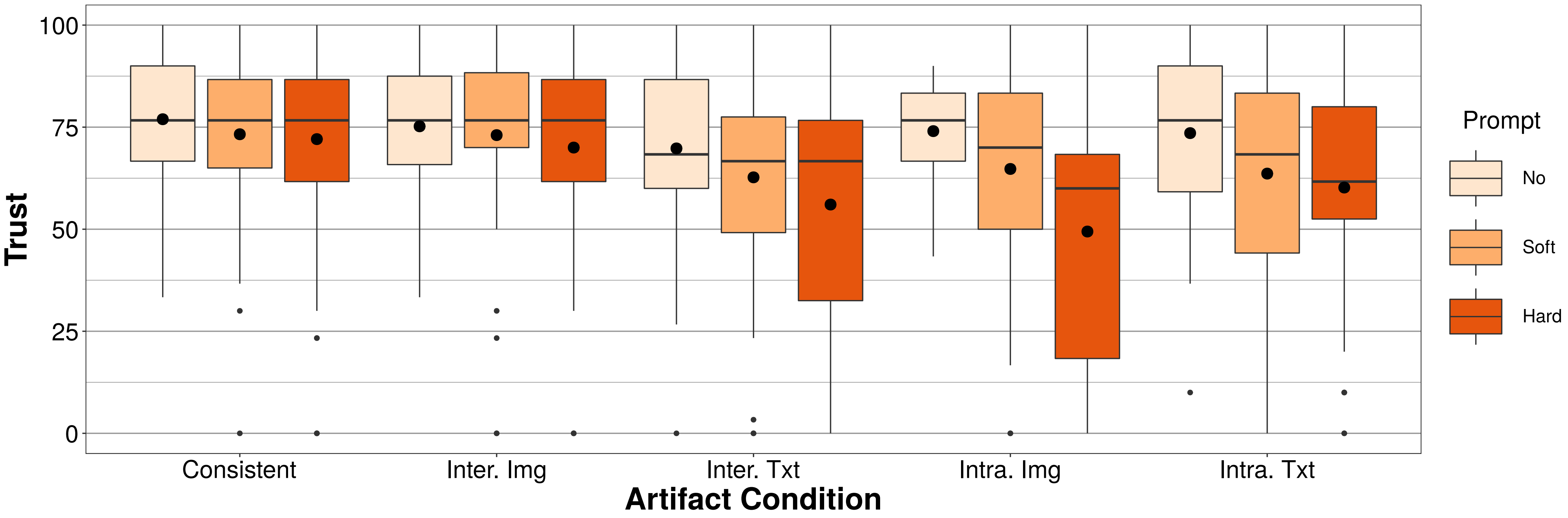


# User Acceptance of Deepfakes (RQ1/RQ2)



> 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

Decreased w/ prompting

- None: 44%, Soft: 23%, Hard: 16%

Inconsistencies (31%)

- “Bot-like” behavior
  - Intra/Inter-field artifacts
  - Generic descriptions
- “Dishonest” behavior

Increased w/prompting

- None: 16%, Soft: 35%, Hard: 43%

# Anti-social Behaviors (RQ3)

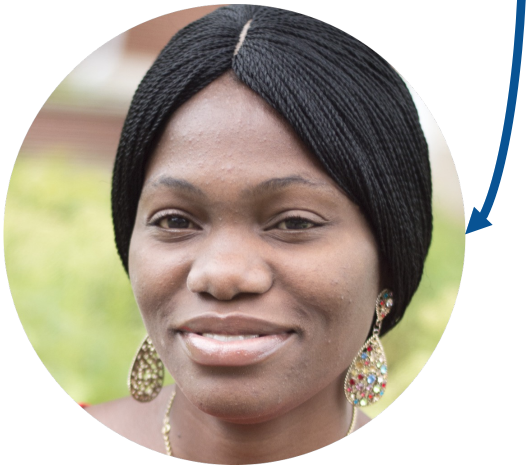
Prompted users occasionally perceive deepfake artifacts within real images/text

"There is something wrong with the applicant's photo...the detail of one of her shoulders is impossible" – P224

Real

"Chris P."

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

# 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

***https://jaronm.ink/deepphish***

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