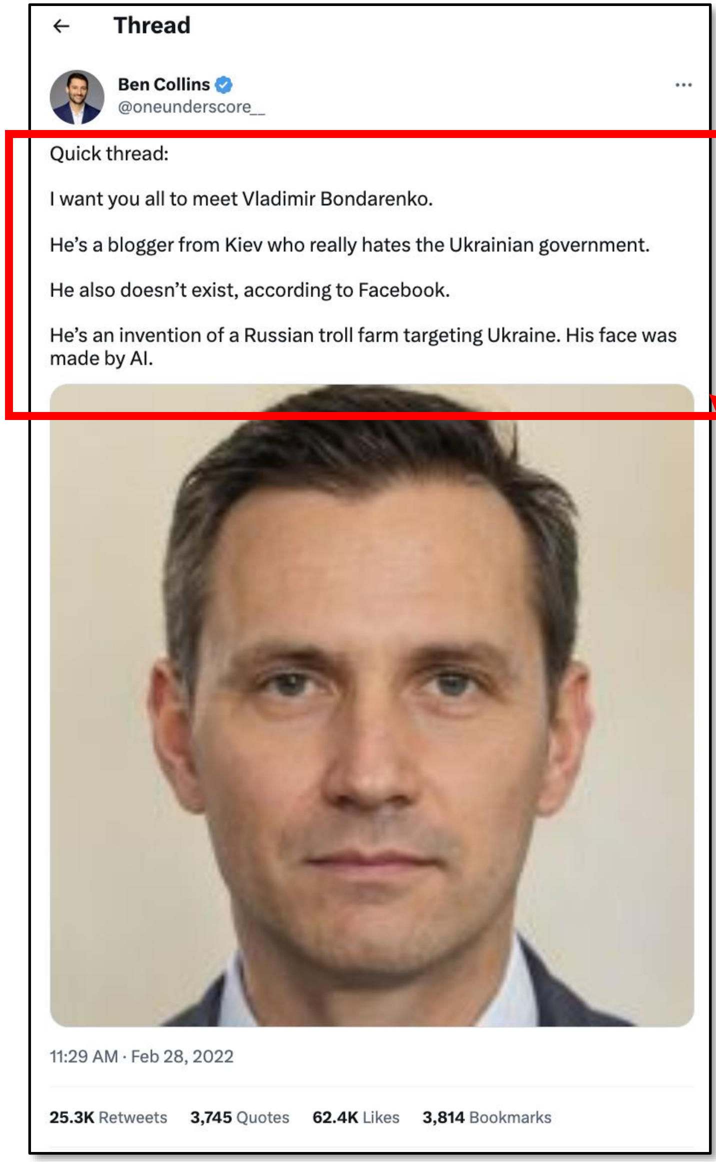
Does it Matter Who Said It?

Exploring the Impact of Deep-fake Enabled Profiles on User Perception Towards Disinformation

Margie Ruffin, Haesenug Seo, Aiping Xiong, Gang Wang

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

PennState



Thread

Ben Collins
@oneunderscore_

Quick thread:

I want you all to meet Vladimir Bondarenko.

He’s a blogger from Kiev who really hates the Ukrainian government.

He also doesn’t exist, according to Facebook.

He’s an invention of a Russian troll farm targeting Ukraine. His face was made by AI.

11:29 AM · Feb 28, 2022

25.3K Retweets 3,745 Quotes 62.4K Likes 3,814 Bookmarks

Quick Thread:

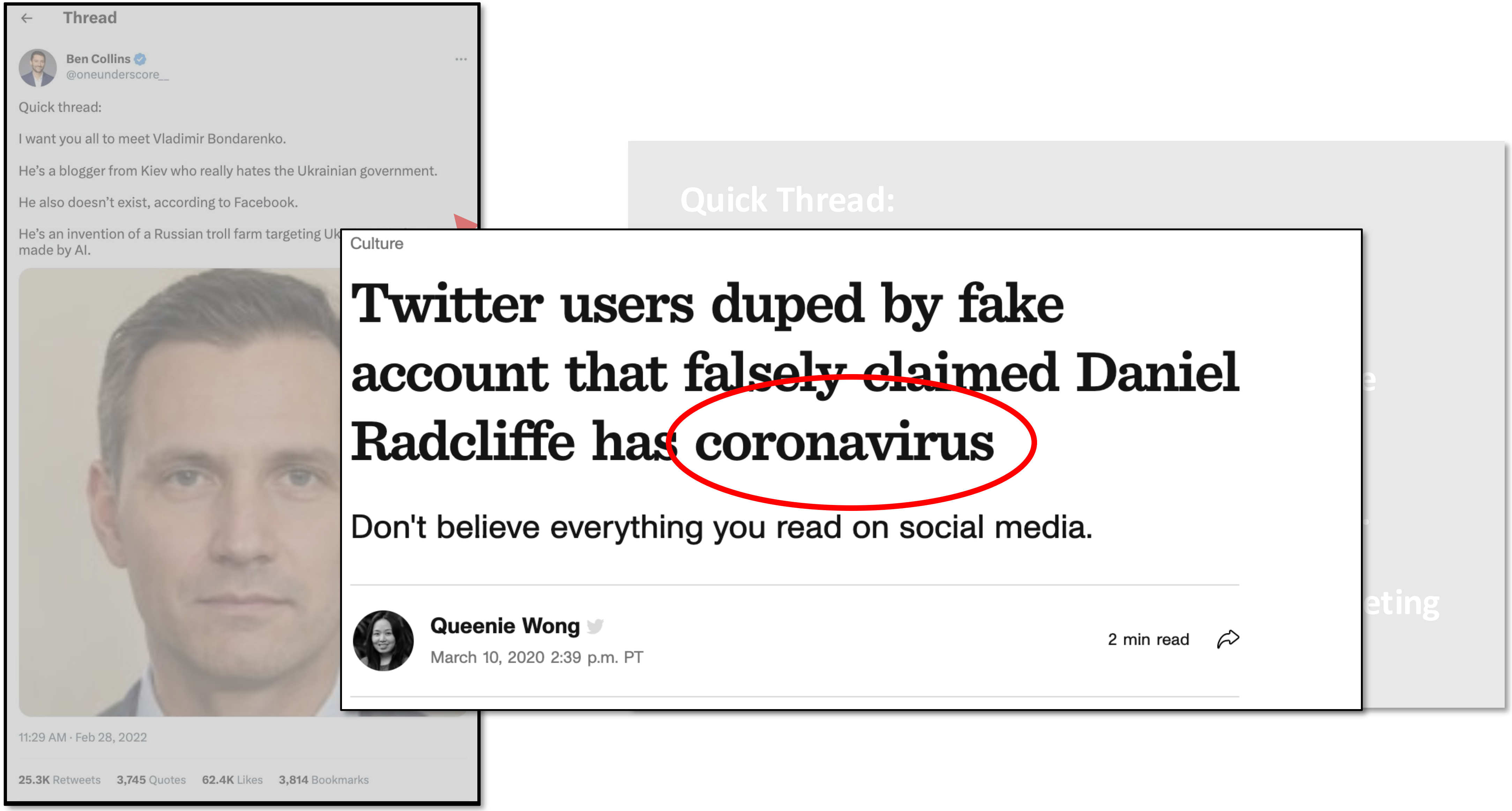
I want you all to meet Valdimir Bondarenko.

He’s a blogger from Kiev who really hates the Ukrainian Gov.

He also doesn’t exist, according to Facebook.

He’s an invention of a Russian troll farm targeting Ukraine. His face was made by AI.

[1] Aparna Banerjea. Digital war: How russia is using deep fakes in ukraine for propaganda. Business Today, 2022. https://www.businesstoday.in/latest/world/story/digital-war-how-russia-is-using-deep-fakes-in-ukraine-for-propaganda-324531-2022-03-02.



Thread

Ben Collins
@oneunderscore_

Quick thread:

I want you all to meet Vladimir Bondarenko.

He's a blogger from Kiev who really hates the Ukrainian government.

He also doesn't exist, according to Facebook.

He's an invention of a Russian troll farm targeting Uk
made by AI.

11:29 AM · Feb 28, 2022

25.3K Retweets 3,745 Quotes 62.4K Likes 3,814 Bookmarks

Culture

Twitter users duped by fake account that falsely claimed Daniel Radcliffe has coronavirus

Don't believe everything you read on social media.

Queenie Wong
March 10, 2020 2:39 p.m. PT

2 min read

[1] Aparna Banerjea. Digital war: How russia is using deep fakes in ukraine for propaganda. Business Today, 2022. https://www.businesstoday.in/latest/world/story/digital-war-how-russia-is-using-deep-fakes-in-ukraine-for-propaganda-324531-2022-03-02.

[2] Queenie Wong. Twitter users duped by fake account that falsely claimed daniel radcliffe has coronavirus. CNET, Mar 2020. https://www.cnet.com/culture/twitter-users-duped-by-fake-account-that-falsely-claimed-daniel-radcliffe-has-coronavirus/.

We Want to Know:
Do Profiles Indicate Trustworthiness?

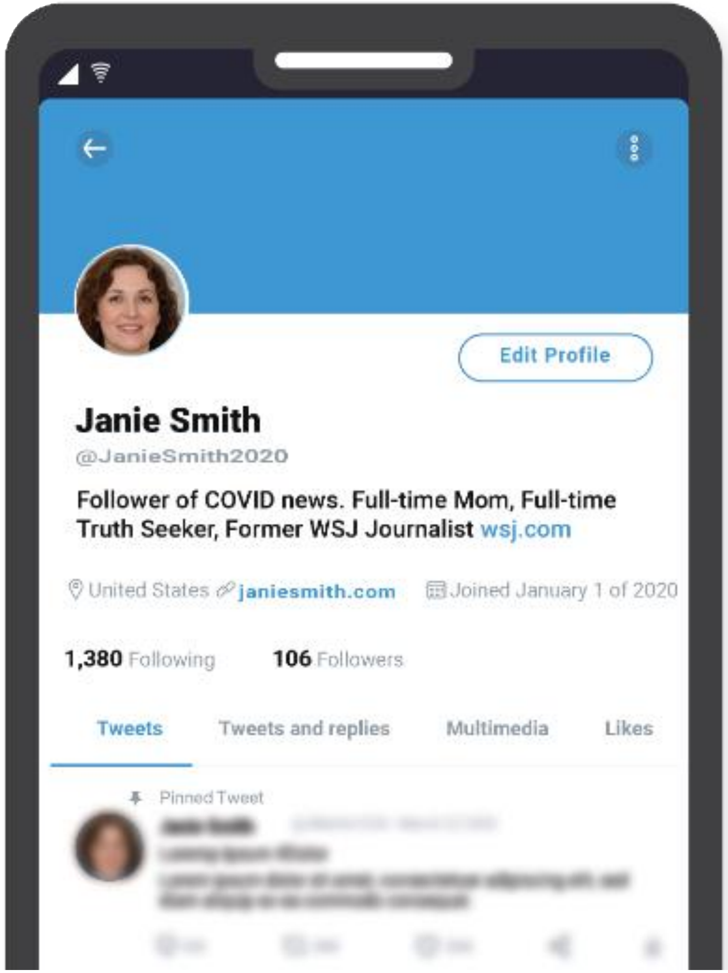
1. Do participants increase their perceived accuracy of tweets if deepfake profiles were also presented compared to showing the tweets only?

2. Do participants increase their engagement with the tweets if deepfake profiles were also presented compared to showing the tweets only?

3. 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?

3 Conditions

- Deepfake
  - Profile imitating journalists using deepfake photo
- Organization
  - Profile imitating a health organization
- Simplefake
  - Profile imitating a simplebot with no image



Edit Profile

Janie Smith

@JanieSmith2020

Follower of COVID news. Full-time Mom, Full-time Truth Seeker, Former WSJ Journalist wsj.com

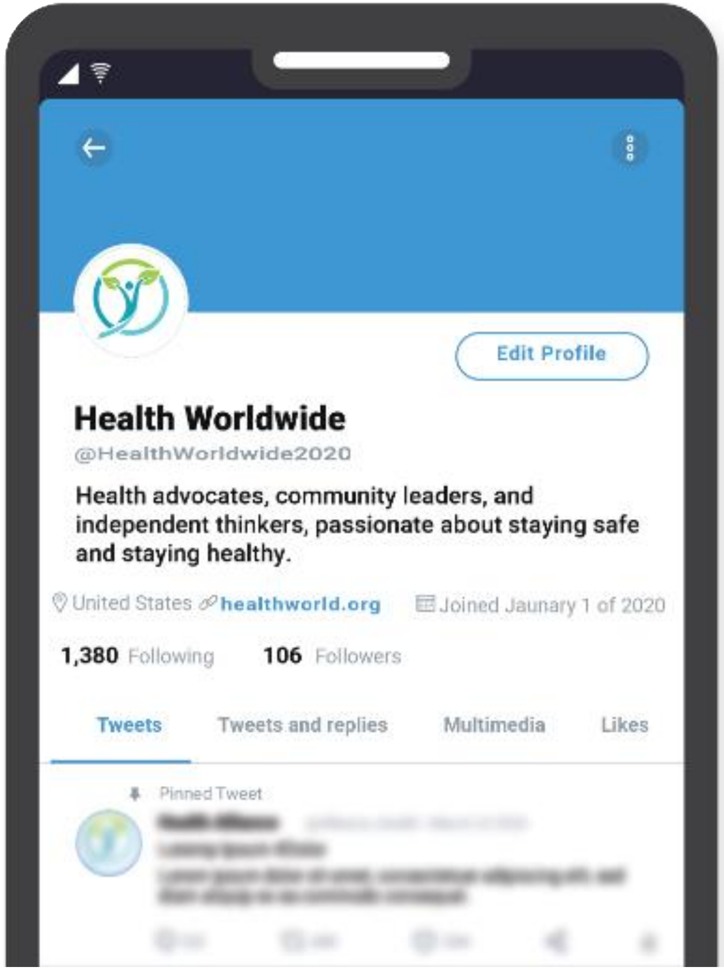
United States janiesmith.com Joined January 1 of 2020

1,380 Following 106 Followers

Tweets Tweets and replies Multimedia Likes

Pinned Tweet

(a) Deepfake



Edit Profile

Health Worldwide

@HealthWorldwide2020

Health advocates, community leaders, and independent thinkers, passionate about staying safe and staying healthy.

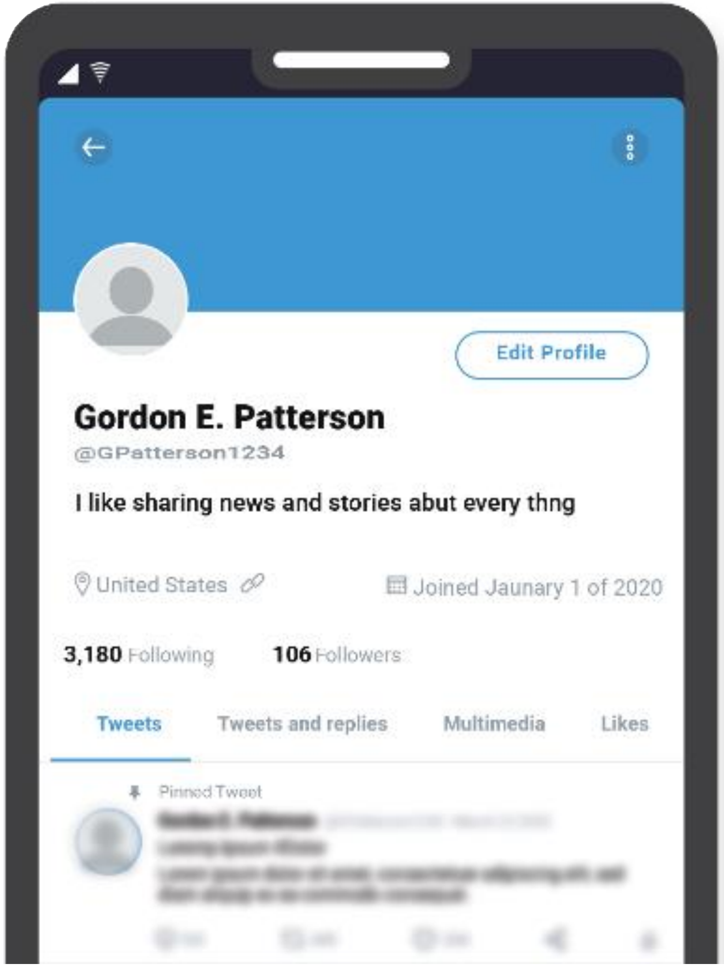
United States healthworld.org Joined Jaunary 1 of 2020

1,380 Following 106 Followers

Tweets Tweets and replies Multimedia Likes

Pinned Tweet

(b) Organization



Edit Profile

Gordon E. Patterson

@GPatterson1234

I like sharing news and stories abut every thng

United States Joined Jaunary 1 of 2020

3,180 Following 106 Followers

Tweets Tweets and replies Multimedia Likes

Pinned Tweet

(c) Simplefake

3 Conditions

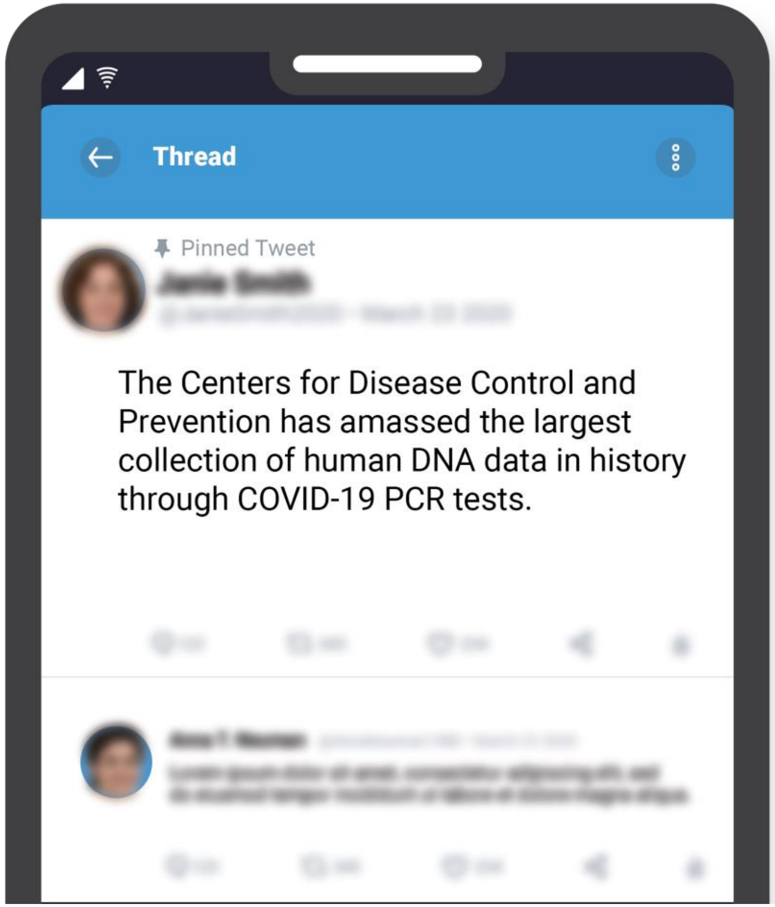
- Deepfake
  - Profile imitating journalists using deepfake photo
- Organization
  - Profile imitating a health organization
- Simplefake
  - Profile imitating a simplebot with no image

3 Tweets

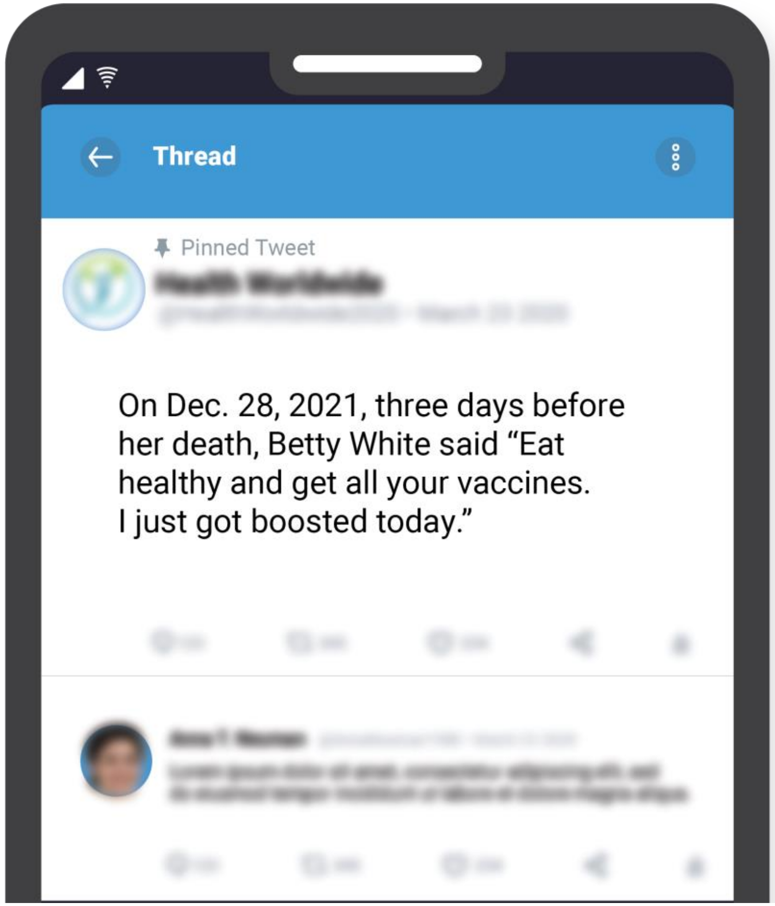
Tweet 1: “The Centers for Disease Control and Prevention has amassed the largest collection of human DNA data in history through COVID-19 PCR tests.”

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.”

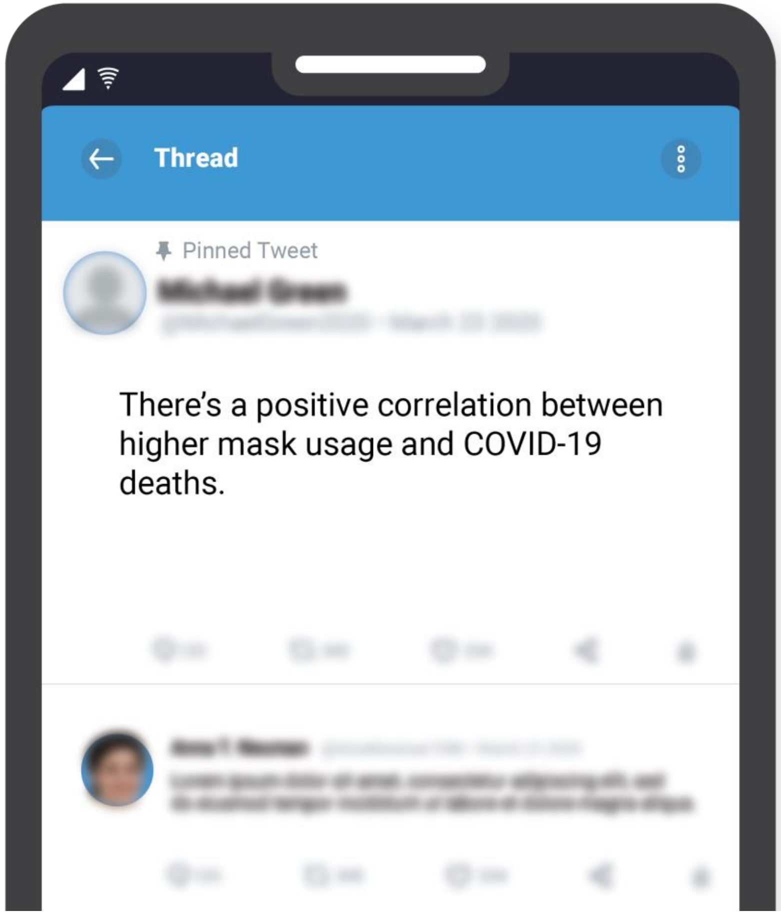
Tweet 3: “There’s a positive correlation between higher mask usage and COVID-19 deaths.”



The Centers for Disease Control and Prevention has amassed the largest collection of human DNA data in history through COVID-19 PCR tests.



On Dec. 28, 2021, three days before her death, Betty White said “Eat healthy and get all your vaccines. I just got boosted today.”



There’s a positive correlation between higher mask usage and COVID-19 deaths.

3 Conditions

- Deepfake
  - Profile imitating journalists using deepfake photo
- Organization
  - Profile imitating a health organization
- Simplefake
  - Profile imitating a simplebot with no image

3 Tweets

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."

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."

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

Experiment Design

Q1: Information accuracy

Q2: Engagement

Q3: Profile authenticity

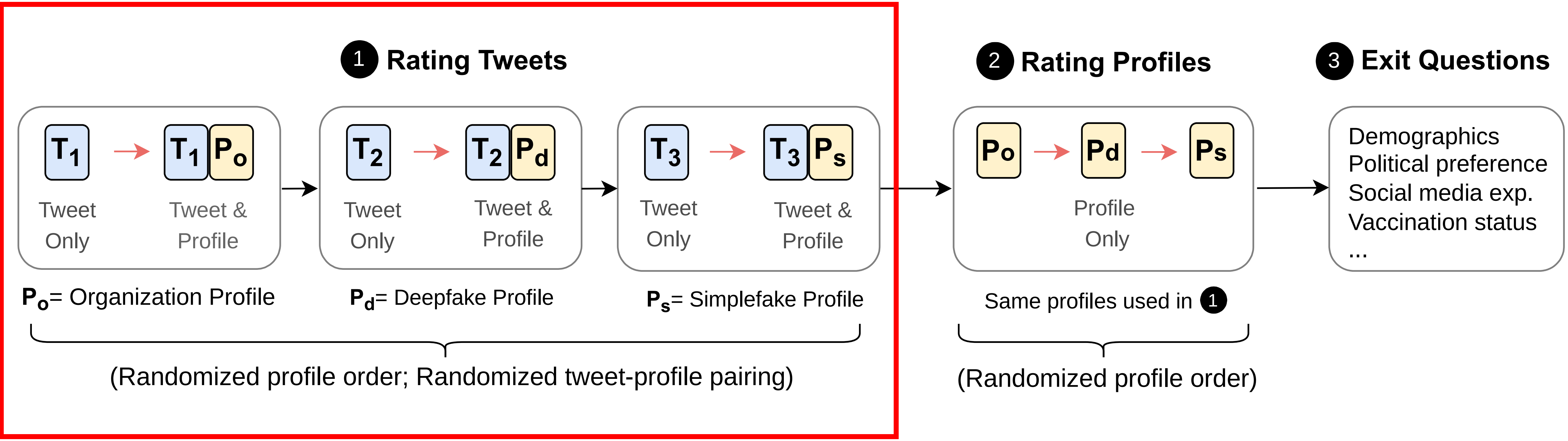
Q4: Profile features

Q5: Recollection

Q6: Reason to engage

Survey Methodology:

From a Participants Perspective



1 Rating Tweets

T_1 → T_1 P_o

Tweet Only

Tweet & Profile

T_2 → T_2 P_d

Tweet Only

Tweet & Profile

T_3 → T_3 P_s

Tweet Only

Tweet & Profile

P_o= Organization Profile

P_d= Deepfake Profile

P_s= Simplefake Profile

(Randomized profile order; Randomized tweet-profile pairing)

2 Rating Profiles

P_o → P_d → P_s

Profile Only

Same profiles used in 1

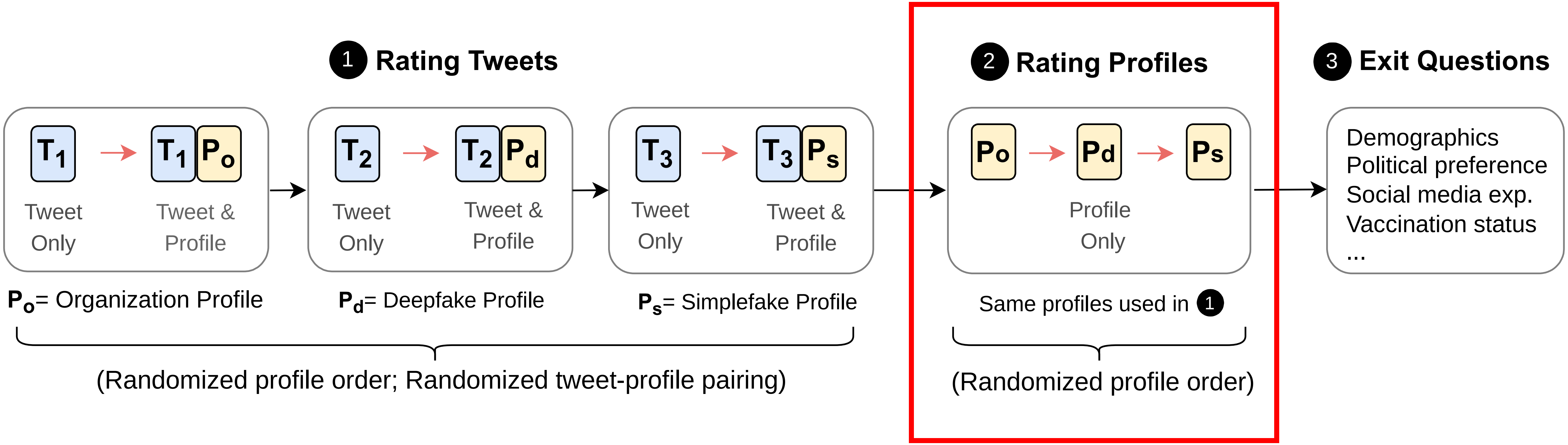
(Randomized profile order)

3 Exit Questions

Demographics
Political preference
Social media exp.
Vaccination status
...

Survey Methodology:

From a Participants Perspective



1 Rating Tweets

T_1 → T_1 P_o

Tweet Only — Tweet & Profile

T_2 → T_2 P_d

Tweet Only — Tweet & Profile

T_3 → T_3 P_s

Tweet Only — Tweet & Profile

P_o= Organization Profile

P_d= Deepfake Profile

P_s= Simplefake Profile

(Randomized profile order; Randomized tweet-profile pairing)

2 Rating Profiles

P_o → P_d → P_s

Profile Only

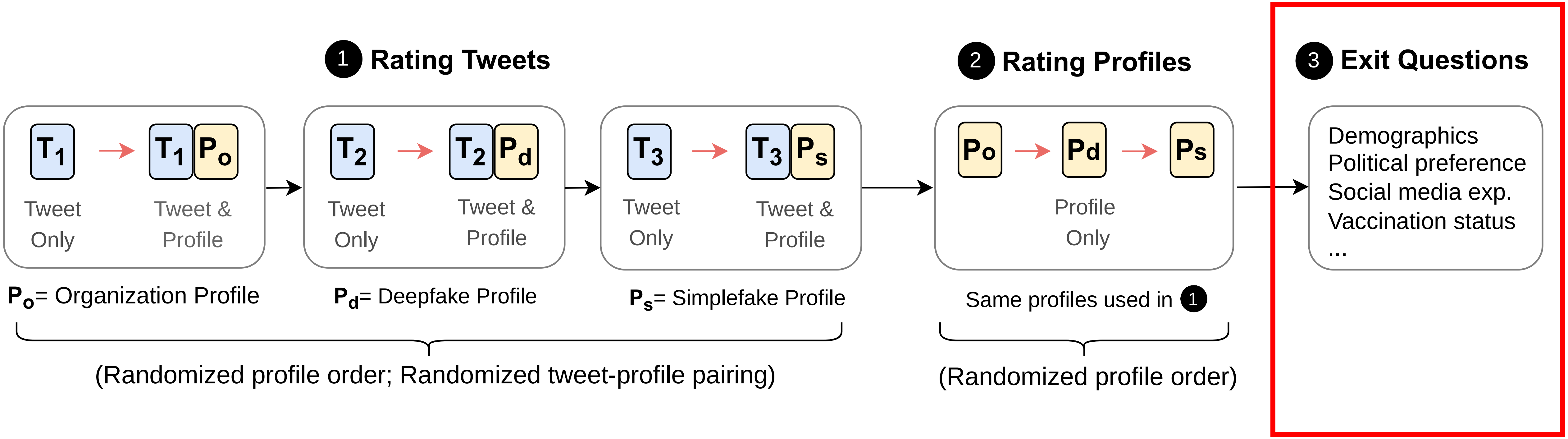
Same profiles used in 1

(Randomized profile order)

3 Exit Questions

Demographics
Political preference
Social media exp.
Vaccination status
...

Survey Methodology:
From a Participants Perspective



P_o= Organization Profile

P_d= Deepfake Profile

P_s= Simplefake Profile

Same profiles used in 1

(Randomized profile order; Randomized tweet-profile pairing)

(Randomized profile order)

Profile Effects on Accuracy and Engagement

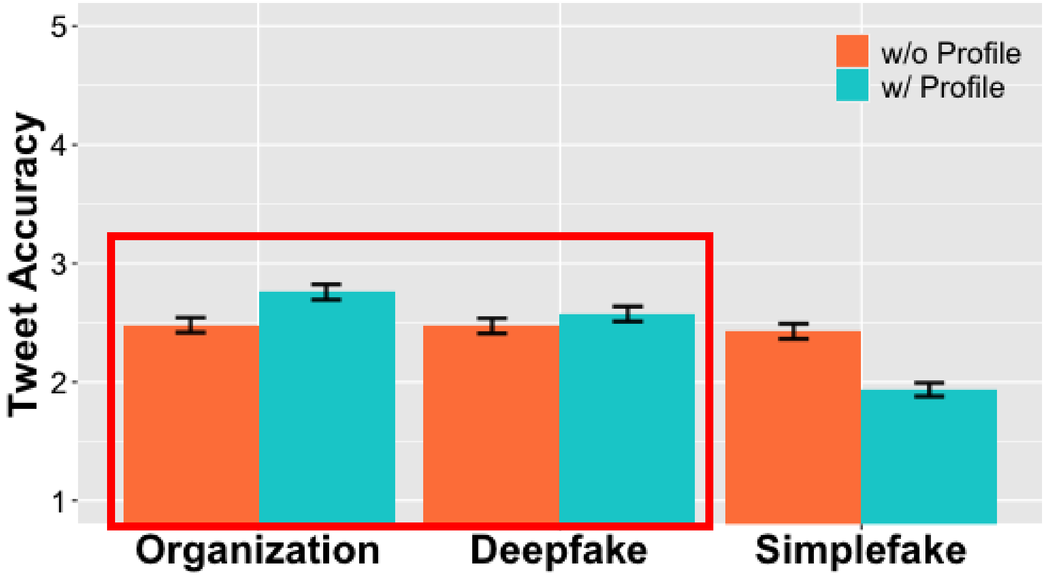


Figure1: Perceived tweet accuracy rating: mean and standard error.

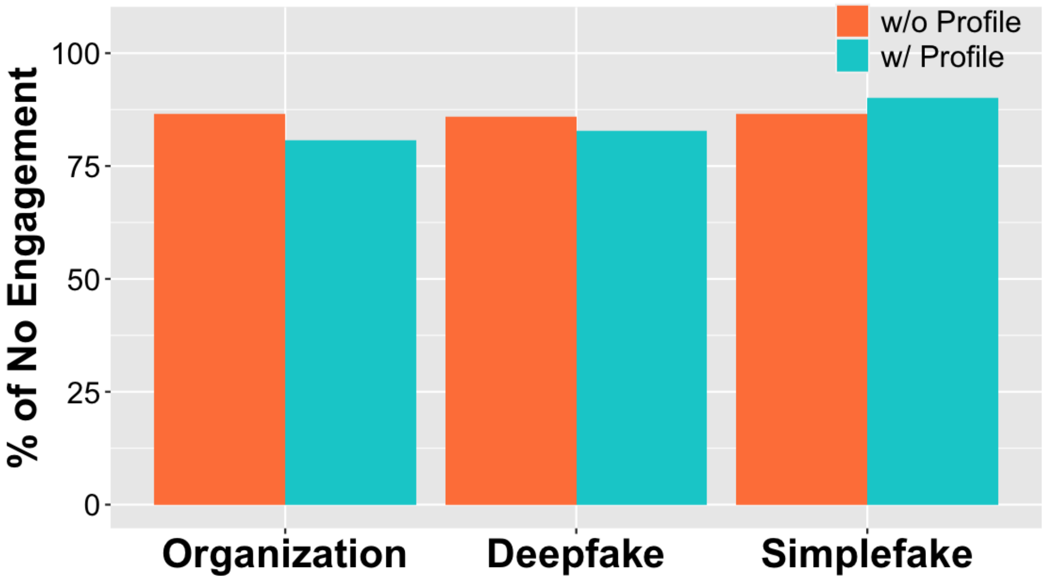


Figure 2: % of Participants who selected "no engagement" towards the tweet.

Profile Effects on Accuracy and Engagement

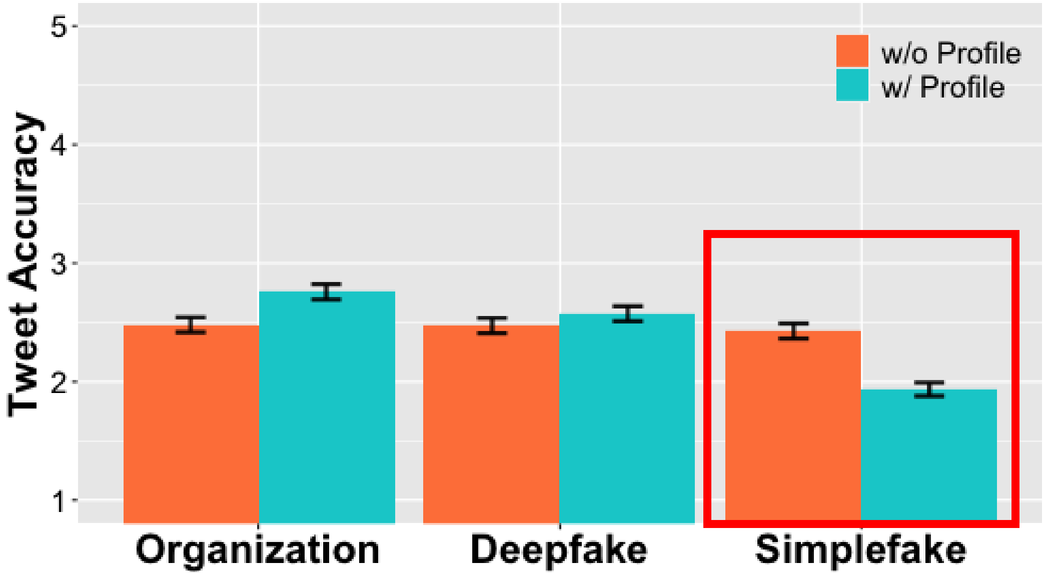


Figure1: Perceived tweet accuracy rating: mean and standard error.

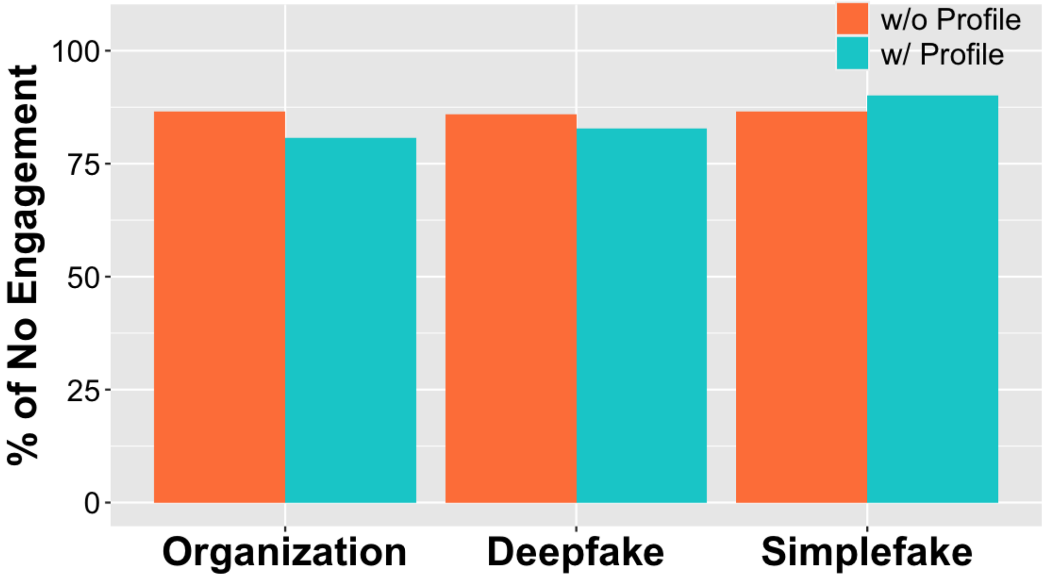


Figure 2: % of Participants who selected “no engagement” towards the tweet.

Profile Effects on Accuracy and Engagement

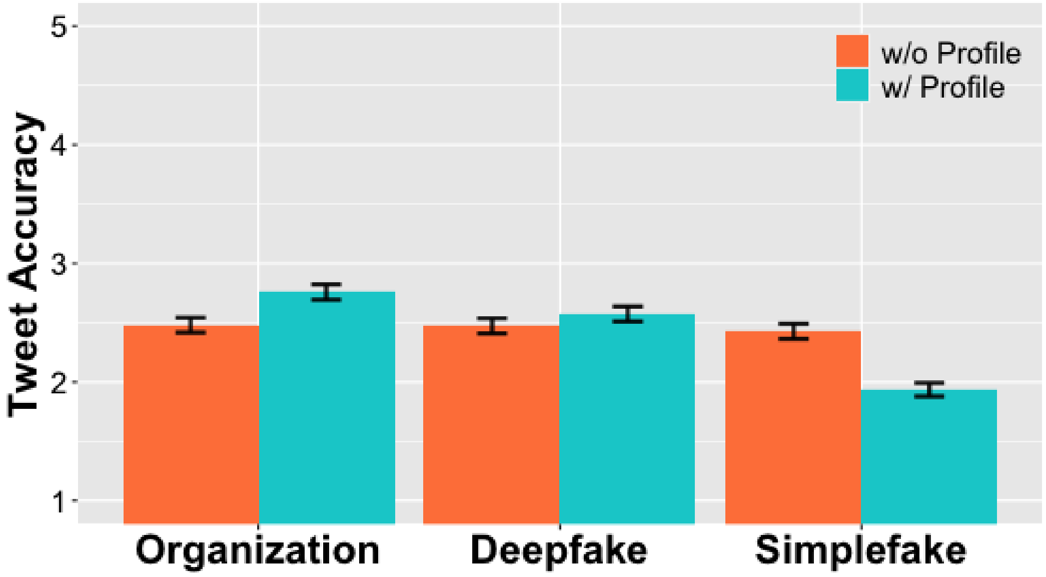


Figure1: Perceived tweet accuracy rating: mean and standard error.

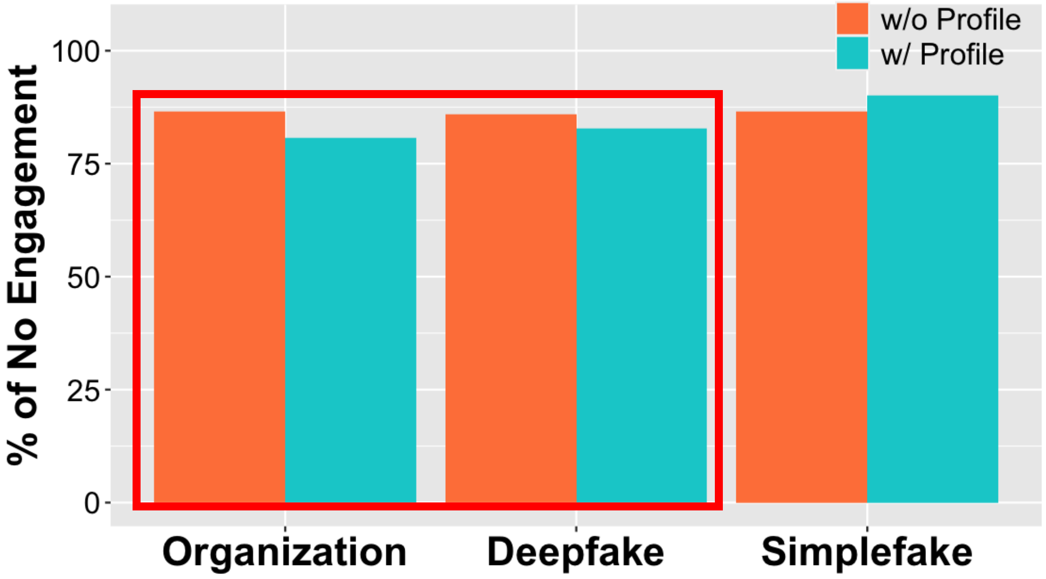
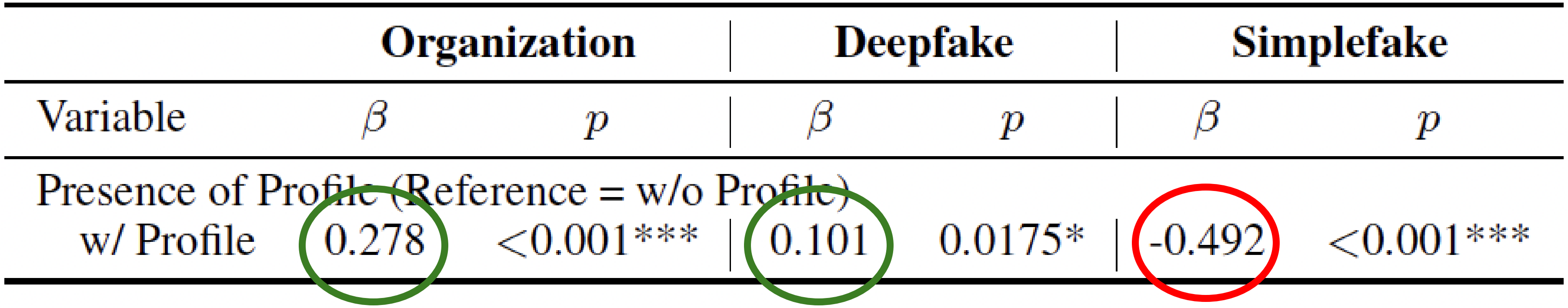


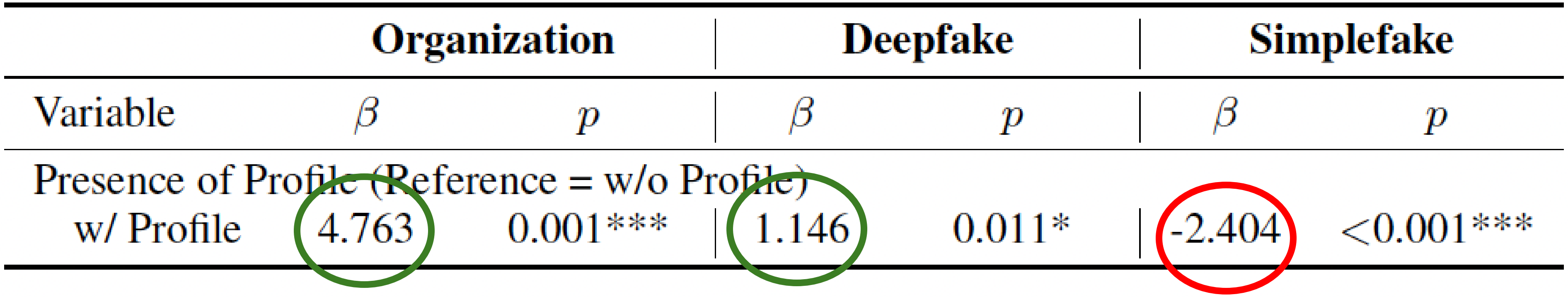
Figure 2: % of Participants who selected "no engagement" towards the tweet.

Profile Effects on Accuracy and Engagement



	Organization		Deepfake		Simplefake	
Variable	β	p	β	p	β	p
Presence of Profile (Reference = w/o Profile)						
w/ Profile	0.278	<0.001***	0.101	0.0175*	-0.492	<0.001***

Table 1: Effect of the presence of profile on tweet accuracy



	Organization		Deepfake		Simplefake	
Variable	β	p	β	p	β	p
Presence of Profile (Reference = w/o Profile)						
w/ Profile	4.763	0.001***	1.146	0.011*	-2.404	<0.001***

Table 2: Effect of the presence of profile on engagement.

Why Do Users Engage With Tweets Perceived as Inaccurate?

12.9% of participants engaged with tweets they rated *“very inaccurate” or “somewhat inaccurate”*

“I replied to one of them to ask why”

“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”

Seek more information.

n = 20 participants

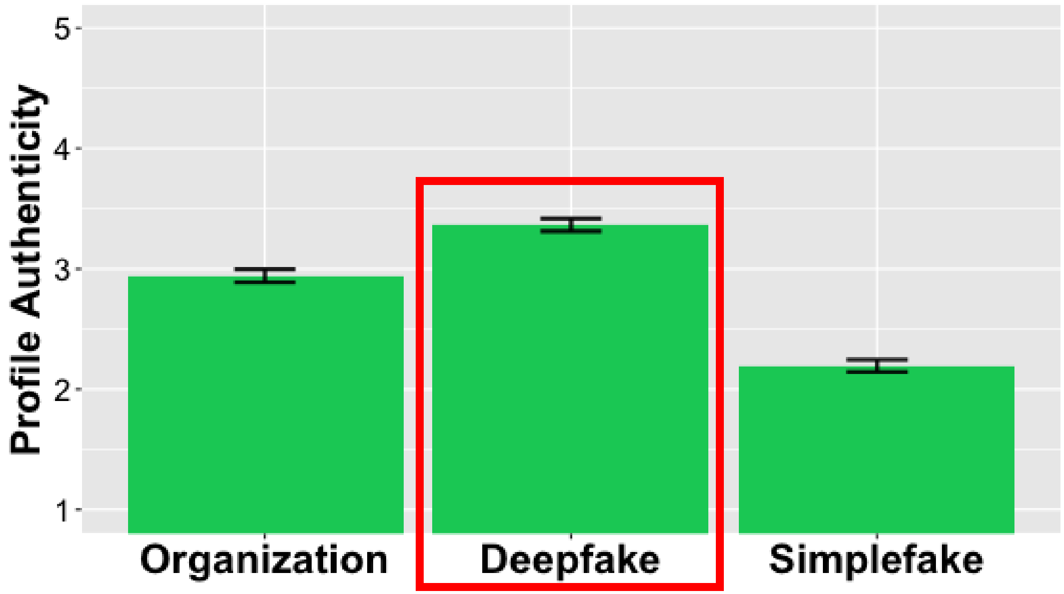
“Sometimes you need to speak some sense into people when they are incredibly wrong”

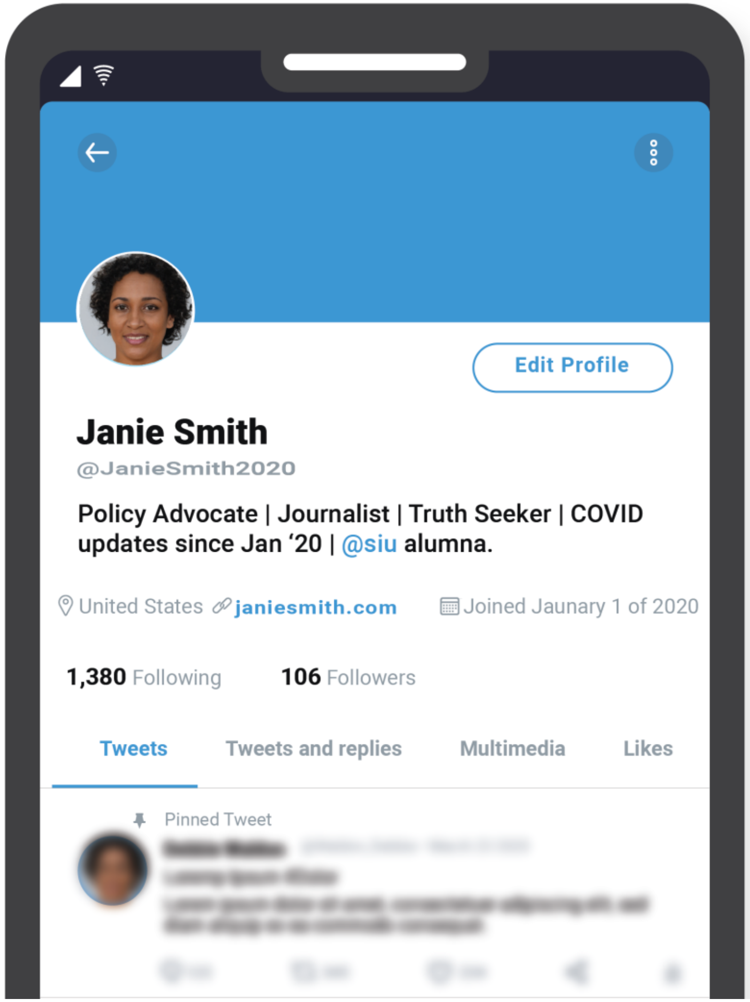
“Because if I see something so blatantly false I feel like I have to reply a response that sows seed of doubt and hope that people would think twice about false information”

Refute disinformation.

n = 13 participants

Which Profile Is The Most Convincing To Users?





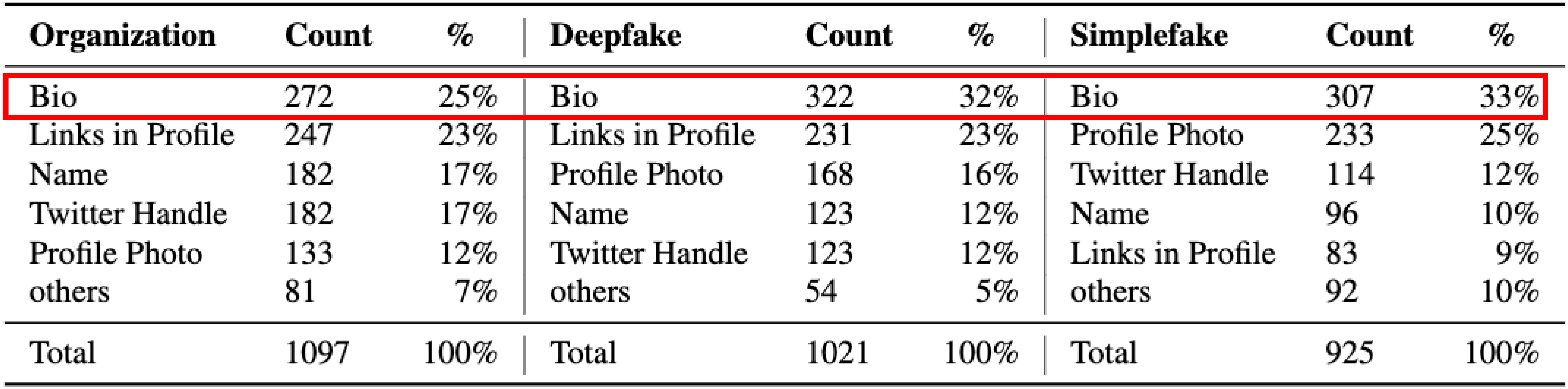
Janie Smith
@JanieSmith2020

Policy Advocate | Journalist | Truth Seeker | COVID updates since Jan '20 | @siu alumna.

United States janiesmith.com Joined Jaunary 1 of 2020

1,380 Following 106 Followers

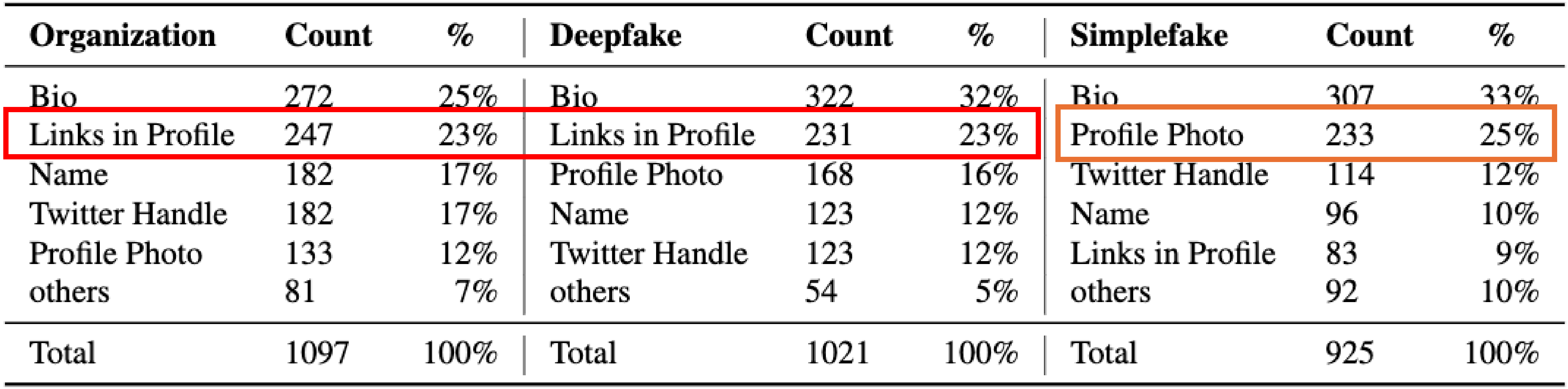
What Makes a Profile Authentic?



Organization	Count	%	Deepfake	Count	%	Simplefake	Count	%
Bio	272	25%	Bio	322	32%	Bio	307	33%
Links in Profile	247	23%	Links in Profile	231	23%	Profile Photo	233	25%
Name	182	17%	Profile Photo	168	16%	Twitter Handle	114	12%
Twitter Handle	182	17%	Name	123	12%	Name	96	10%
Profile Photo	133	12%	Twitter Handle	123	12%	Links in Profile	83	9%
others	81	7%	others	54	5%	others	92	10%
Total	1097	100%	Total	1021	100%	Total	925	100%

Table 3: The Most Influential Profile Feature—We ask participants to select profile features that influence the information accuracy rating. The total numbers across the three conditions are different because participants can select multiple features per profile.

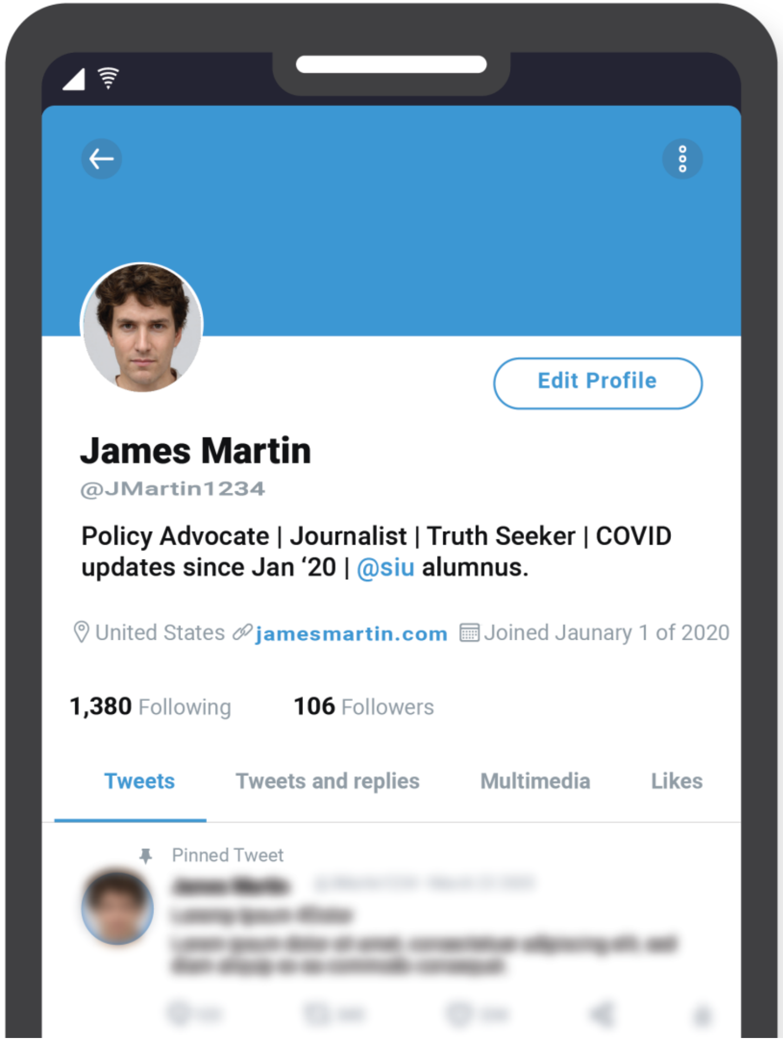
What Makes a Profile Authentic?



Organization	Count	%	Deepfake	Count	%	Simplefake	Count	%
Bio	272	25%	Bio	322	32%	Bio	307	33%
Links in Profile	247	23%	Links in Profile	231	23%	Profile Photo	233	25%
Name	182	17%	Profile Photo	168	16%	Twitter Handle	114	12%
Twitter Handle	182	17%	Name	123	12%	Name	96	10%
Profile Photo	133	12%	Twitter Handle	123	12%	Links in Profile	83	9%
others	81	7%	others	54	5%	others	92	10%
Total	1097	100%	Total	1021	100%	Total	925	100%

Table 3: The Most Influential Profile Feature—We ask participants to select profile features that influence the information accuracy rating. The total numbers across the three conditions are different because participants can select multiple features per profile.

What Does This All Mean?



James Martin
@JMartin1234

Policy Advocate | Journalist | Truth Seeker | COVID updates since Jan '20 | @siu alumnus.

United States jamesmartin.com Joined Jaunary 1 of 2020

1,380 Following 106 Followers

Pinned Tweet

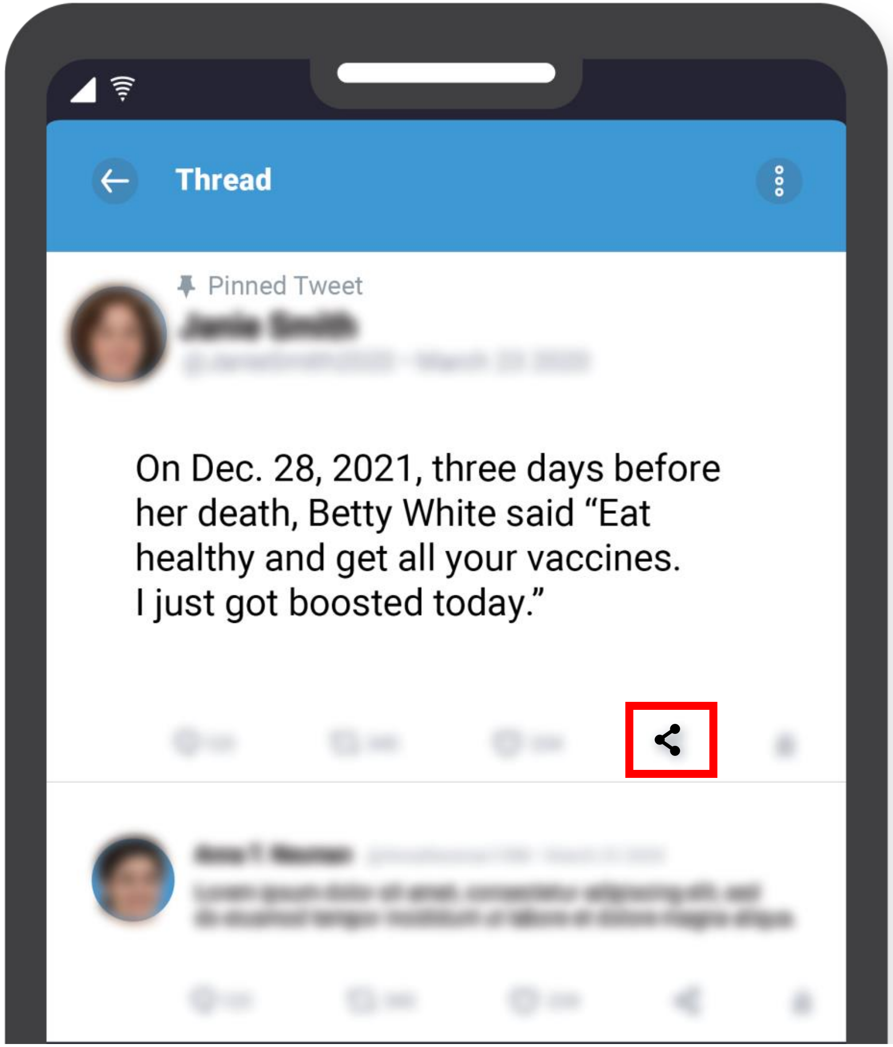
Our study shows the significant impact of deepfake profiles on participants' accuracy rating of and engagement with disinformation.

- Validates prior work suggesting deepfakes help in social engineering
- Users need help in identifying deepfake profiles on social media

What Does This All Mean?

Users unintentionally disseminate disinformation in an effort to correct the original poster by retweeting or replying.

- Alternative methods could be offered to help mitigate this



On Dec. 28, 2021, three days before her death, Betty White said “Eat healthy and get all your vaccines. I just got boosted today.”

Thank You!

Deepfake-enabled and organization profiles can affect how users view fake news.

Social Media users should be careful when engaging with fake news, because engagements helps to disseminate it.

https://margieruffin.com

mruffin2@Illinois.edu