

# Does It Matter Who Said It?

## Exploring the Impact of Deepfake Profiles On User Perception towards Disinformation

Margie Ruffin, Haeseung Seo, Aiping Xiong, Gang Wang



*<https://mruffin.github.io>*



UNIVERSITY OF  
**ILLINOIS**  
URBANA - CHAMPAIGN

# Deepfake profiles are on the rise in social media...

- Social Media has been plagued by disinformation campaigns
- Most of these “fake” profiles have been easy to spot
- Research shows, that’s not the case anymore

# Deepfake profiles are on the rise in social media...

- Social Media has been plagued by disinformation campaigns
- Most of these “fake” profiles have been easy to spot
- Research shows, that’s not the case anymore

What does this mean for the users of social media?



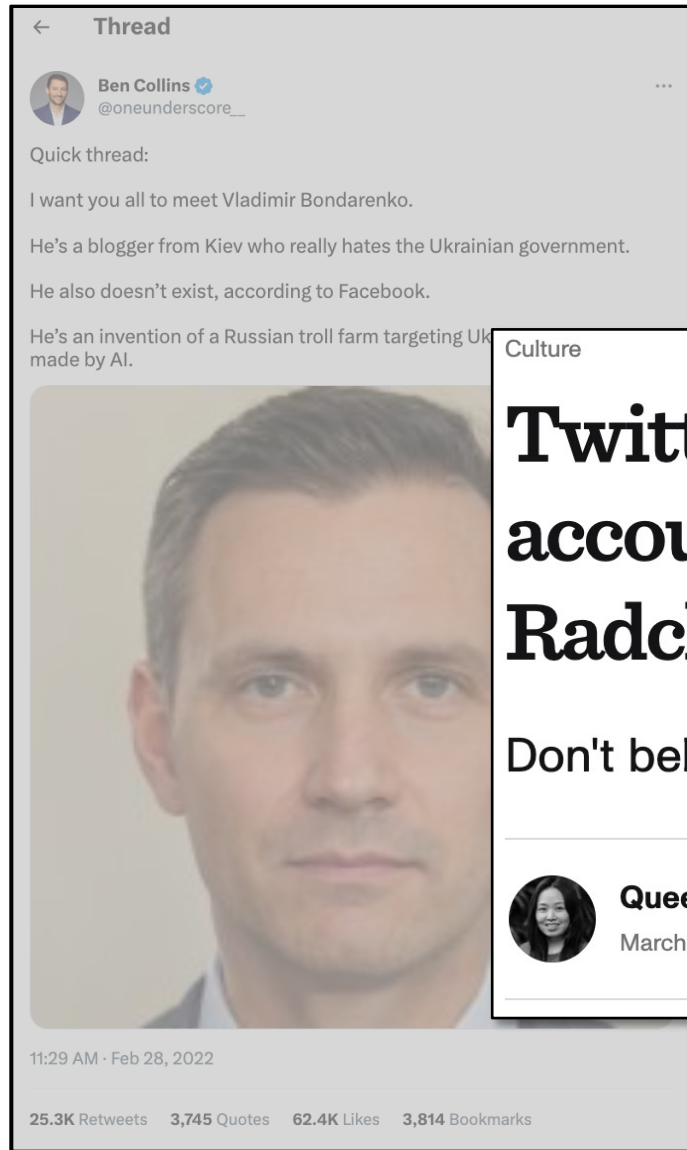
**Quick Thread:**

**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.**



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**

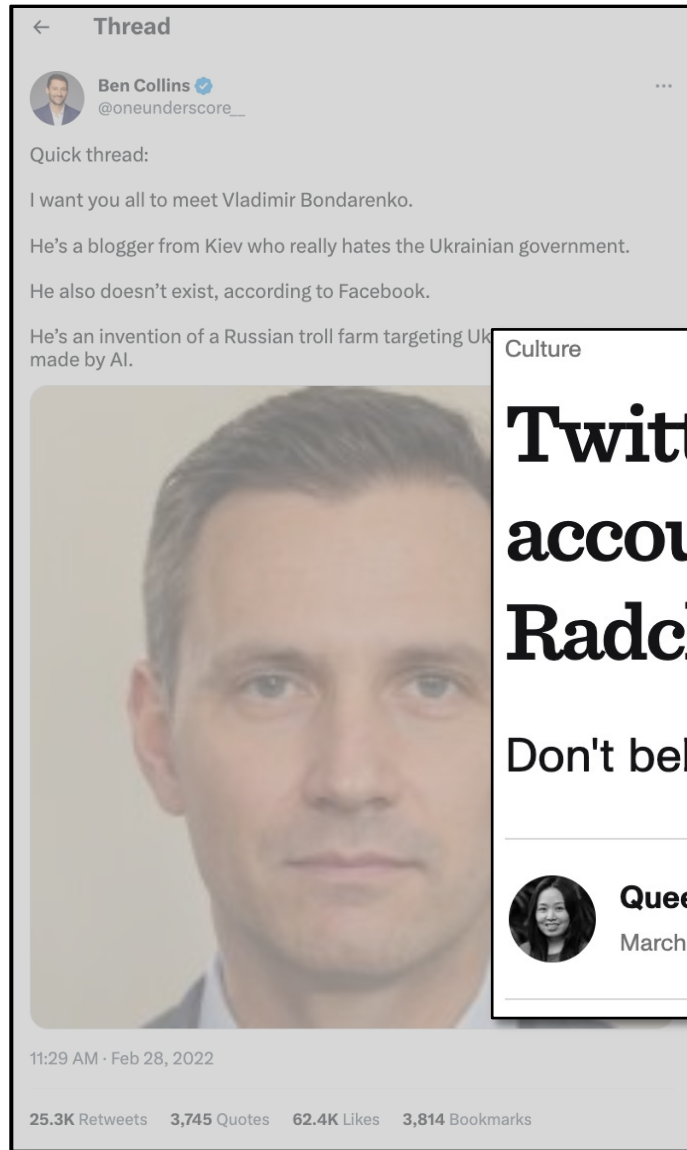
March 10, 2020 2:39 p.m. PT

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/>.



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 alter perceptions?

- RQ1: Do participants increase their perceived accuracy of tweets if deepfake profiles were also presented compared to showing the tweets only?

# We want to know: Do profiles alter perceptions?

- 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?



# We want to know: Do profiles alter perceptions?

- 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?

# 3 Conditions

- Deepfake
- Organization
- Simplefake

# 3 Conditions

- Deepfake
- Organization
- Simplefake

# 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.”

## 3 Conditions

- Deepfake
- Organization
- Simplefake


## 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

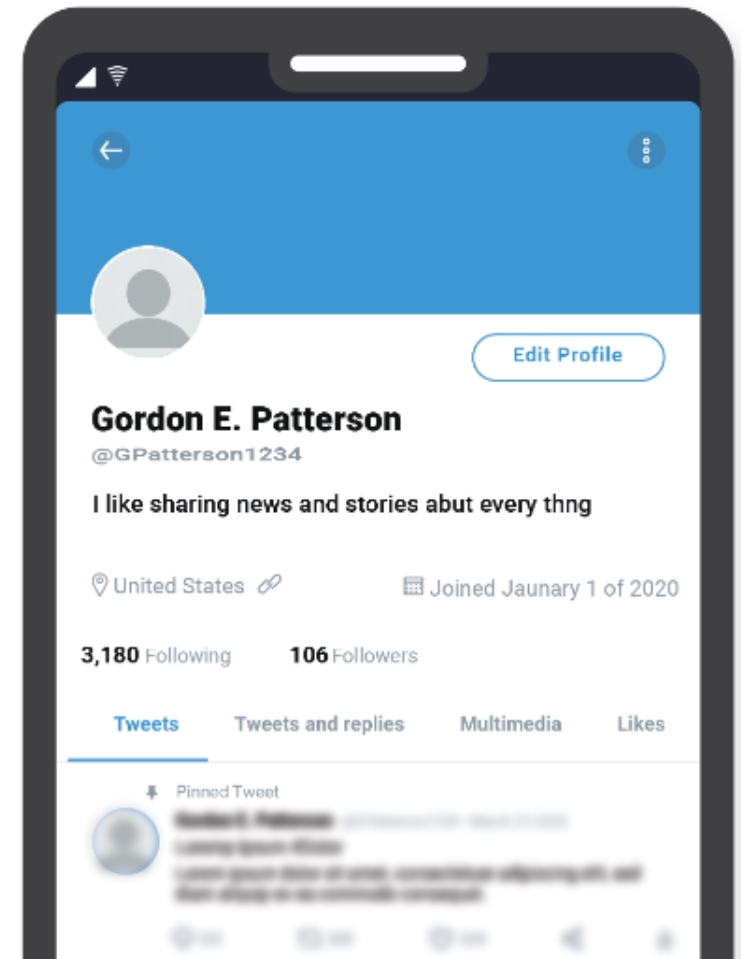
- Q1: Information accuracy
  - Q2: Engagement
  - Q3: Profile authenticity
  - Q4: Profile features
  - Q5: Recollection
  - Q6: Reason to engage
- 



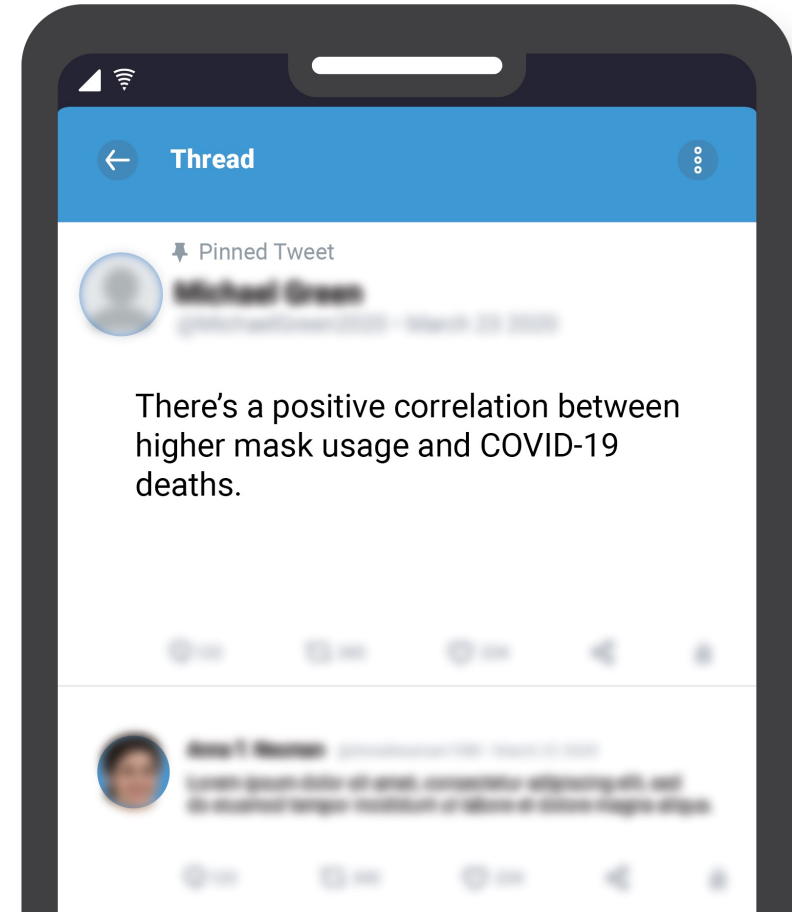
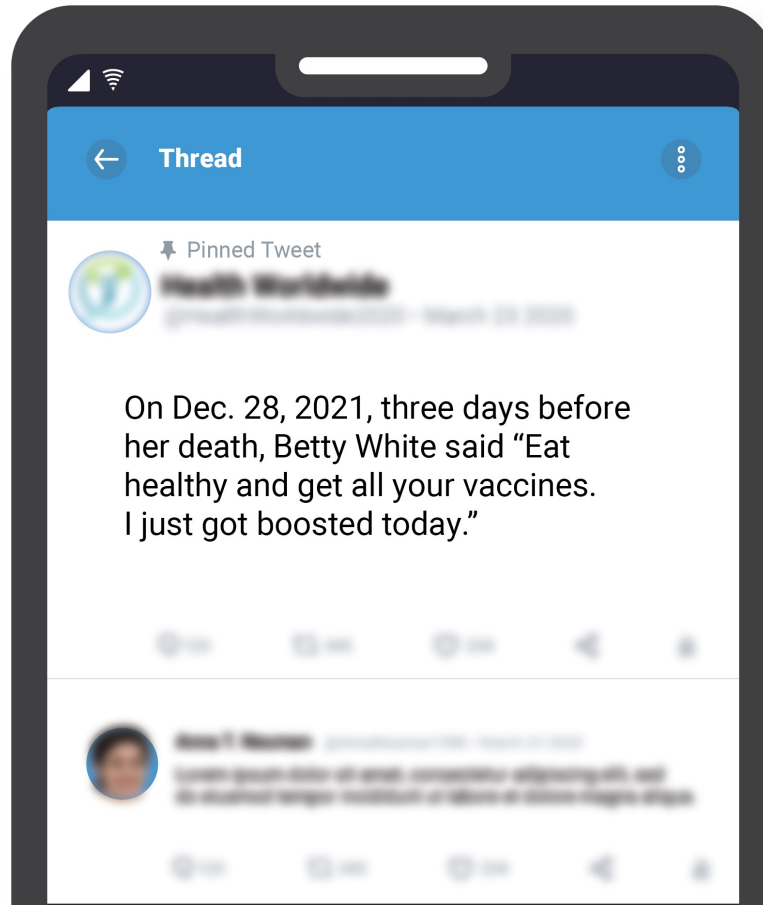
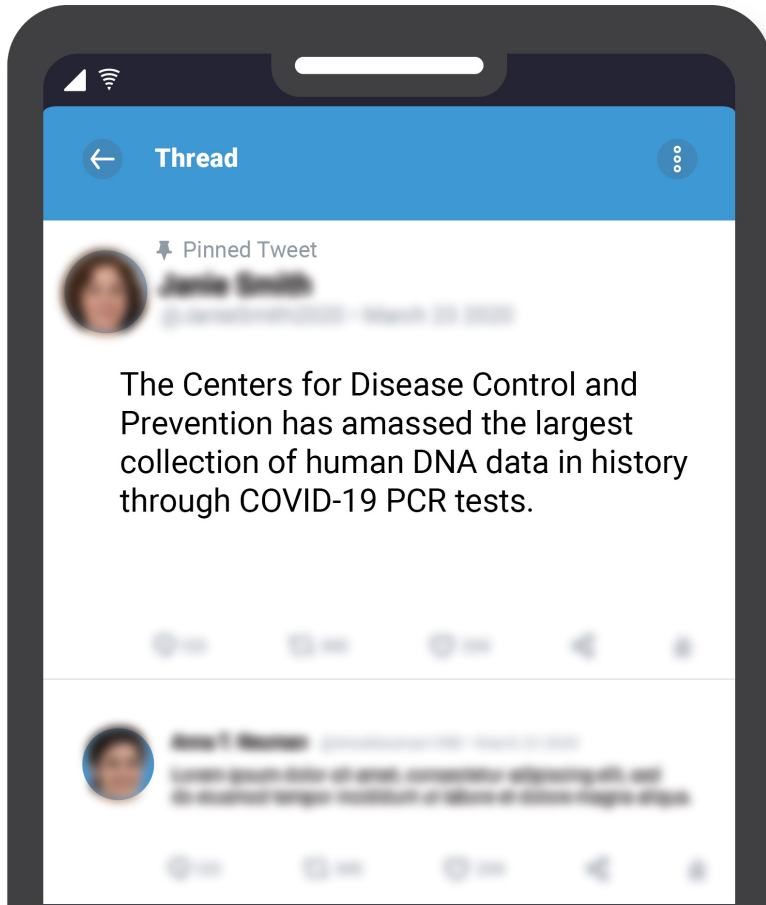
(a) Deepfake



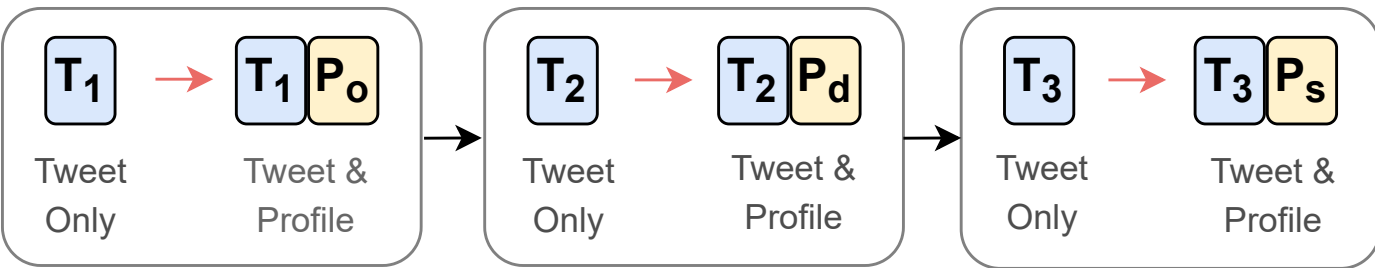
(b) Organization



(c) Simplefake



## 1 Rating Tweets



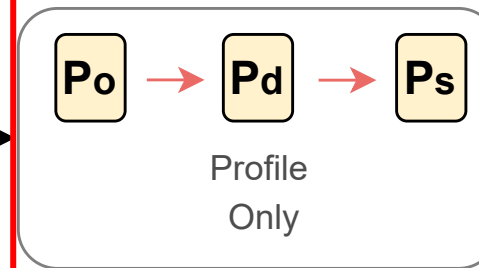
**P<sub>o</sub>**= Organization Profile

**P<sub>d</sub>**= Deepfake Profile

**P<sub>s</sub>**= Simplefake Profile

(Randomized profile order; Randomized tweet-profile pairing)

## 2 Rating Profiles



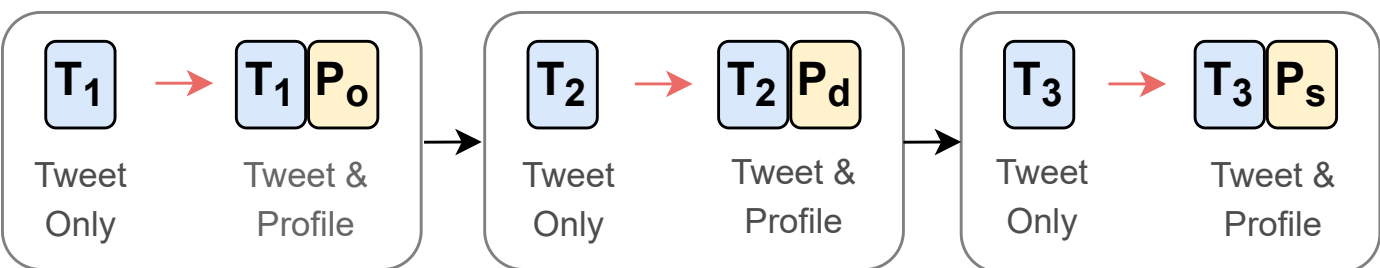
Same profiles used in 1

(Randomized profile order)

## 3 Exit Questions

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

## 1 Rating Tweets



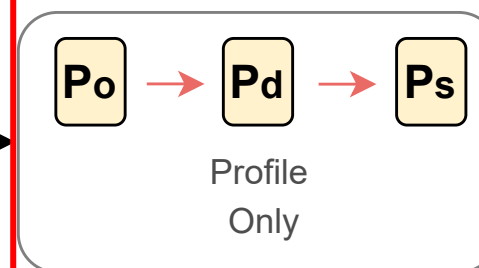
$P_o$  = Organization Profile

$P_d$  = Deepfake Profile

$P_s$  = Simplefake Profile

(Randomized profile order; Randomized tweet-profile pairing)

## 2 Rating Profiles



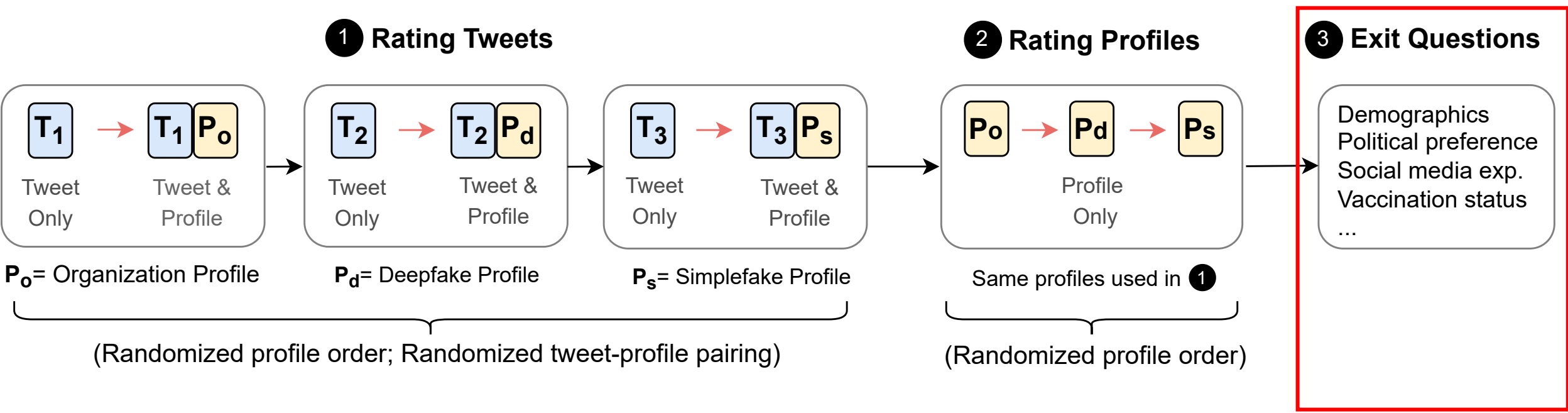
Same profiles used in 1

(Randomized profile order)

## 3 Exit Questions

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

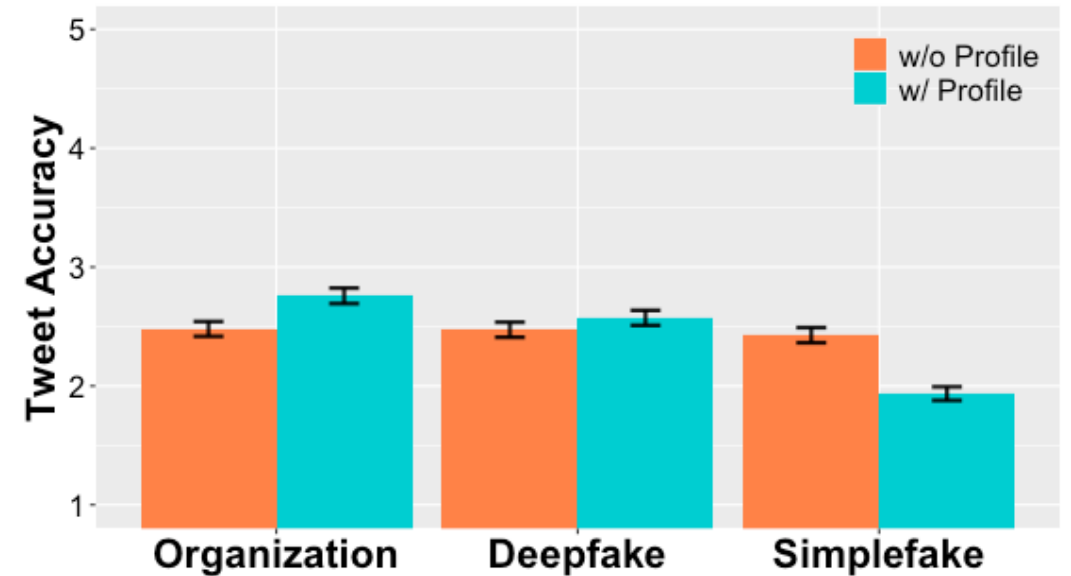




# What do users have to say about tweet accuracy?

The perceived tweet accuracy is increased when a profile of authority (organization and deepfake) is presented compared with showing the tweet alone.

The perceived accuracy decreases when the presented profile is a simple bot profile.



# Why do users engage inaccurate tweets?

The engagement of tweet increases when a profile of authority (organization and deepfake) is presented, vs showing the tweet alone.

Engagement decreases when the presented profile is a simple bot profile

## **Seek more information.**

*“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.**

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



8,866



33.4K



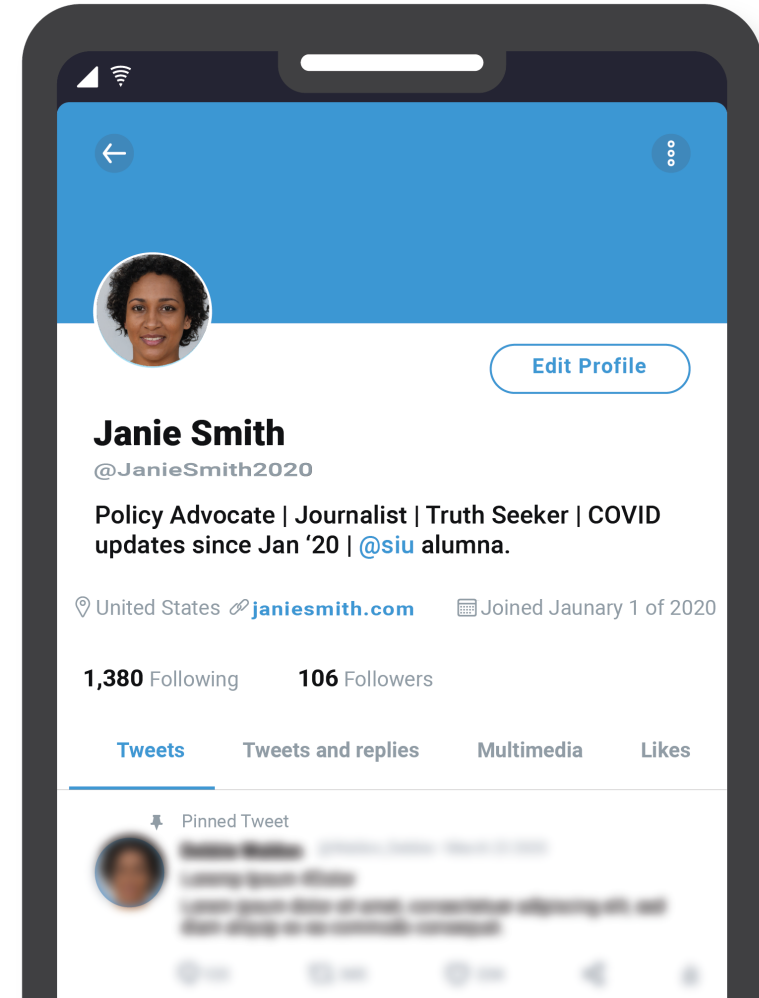
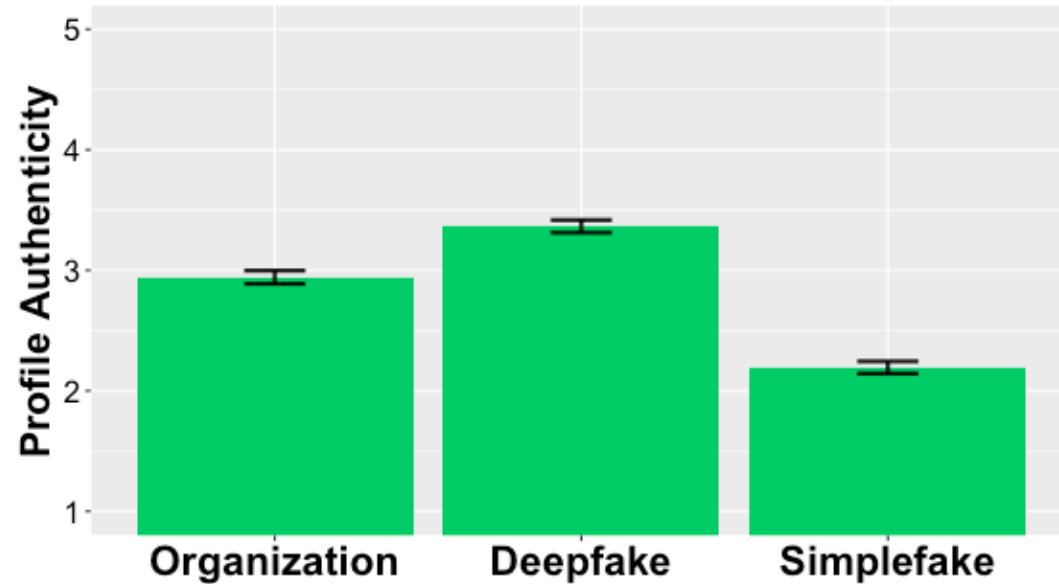
234.2K



30.1M



# Which profile is the most convincing to viewers?



# 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: 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: 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?

- 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
- 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

## + Future Work

- Examine the effects that various combinations of deepfake techniques have on downstream attacks
  - How sophisticated of an attack can be orchestrated using these techniques
- Consider the integration of technical- and human-aspect solutions for deepfake detection and mitigation
  - Can we teach people how to identify deepfakes (photos, videos, text, etc.)



# Thank You!



## Summary

- The plague of disinformation in social media platforms is worsened by the use of sophisticated deepfake profiles
- Users need assistance in identifying such profiles to help curb the spread of false information
- Alternative methods of correction by other users are needed for these platforms

Margie Ruffin, Haeseung Seo, Aiping Xiong, Gang Wang

