Misinformation: going antiviral

CS 278 | Stanford University | Old Tom Holland

Announcements

Final projects are due a week from Friday

- Unfortunately, no late days on the final project…because, unfortunately, we don’t get late days on reporting final grades for graduating students to the Registrar

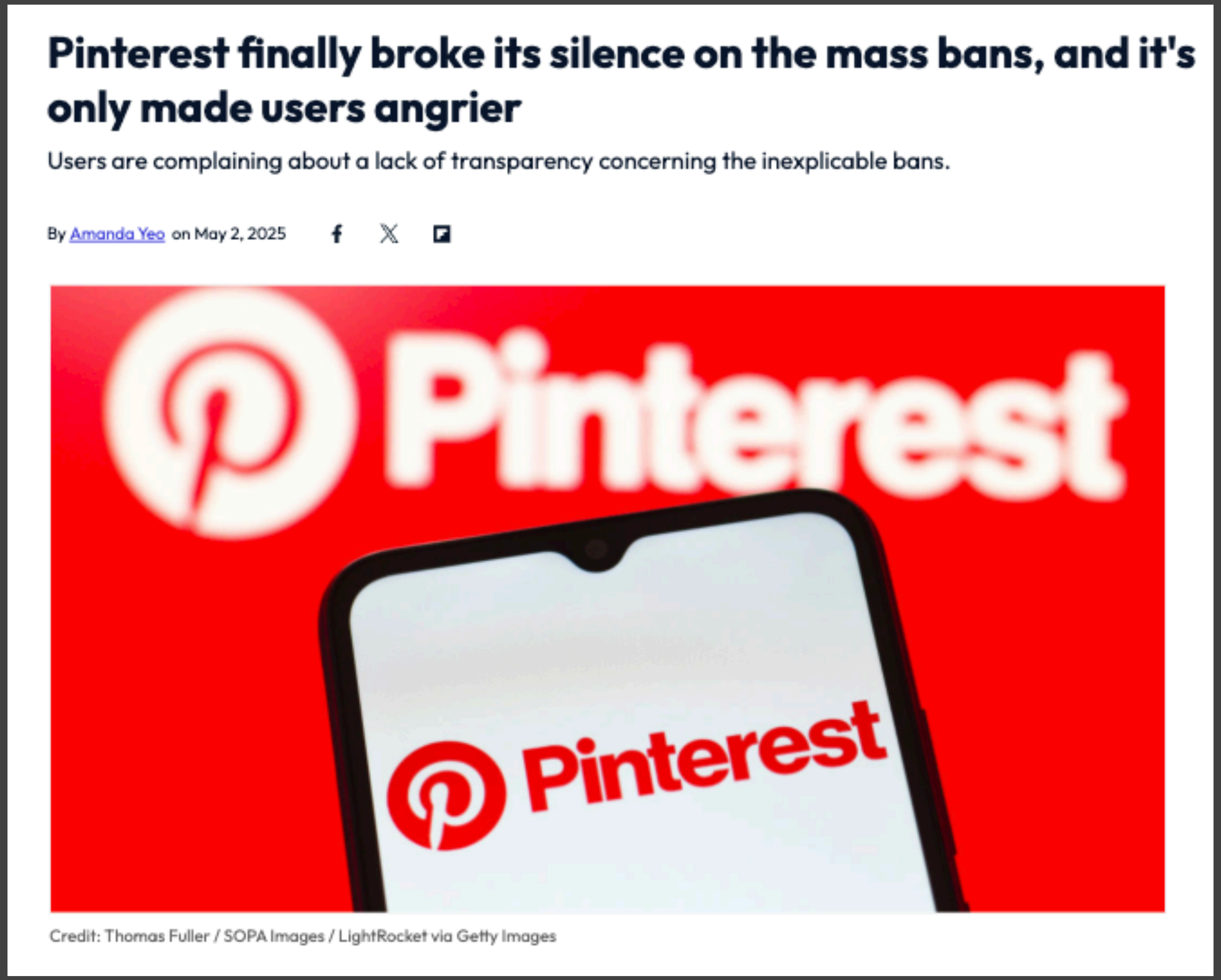
Last time

As Gillespie argues, moderation is the commodity of the platform: it sets apart what is allowed on the platform, and has downstream influences on descriptive norms.

Moderation works: it can change the community's behavior

Moderation classification rules are fraught and challenging — they reify what many of us carry around as unreflective understandings.

“Moderation” example submitted by Julia Gendy

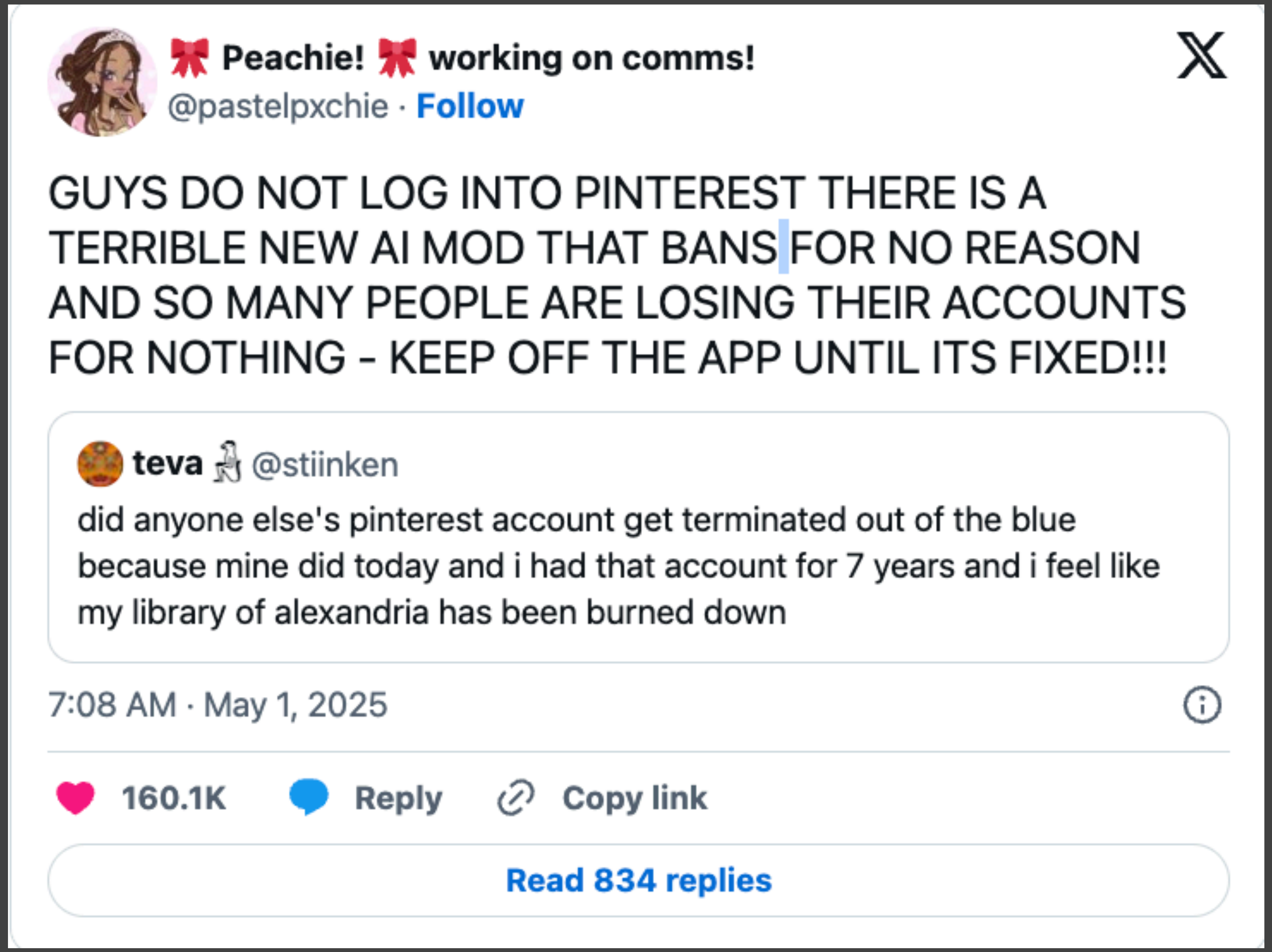


Pinterest finally broke its silence on the mass bans, and it's only made users angrier

Users are complaining about a lack of transparency concerning the inexplicable bans.

By Amanda Yeo on May 2, 2025

Credit: Thomas Fuller / SOPA Images / LightRocket via Getty Images



🎀 Peachie! 🎀 working on comms!
@pastelpxchie · Follow

GUYS DO NOT LOG INTO PINTEREST THERE IS A TERRIBLE NEW AI MOD THAT BANS FOR NO REASON AND SO MANY PEOPLE ARE LOSING THEIR ACCOUNTS FOR NOTHING - KEEP OFF THE APP UNTIL ITS FIXED!!!

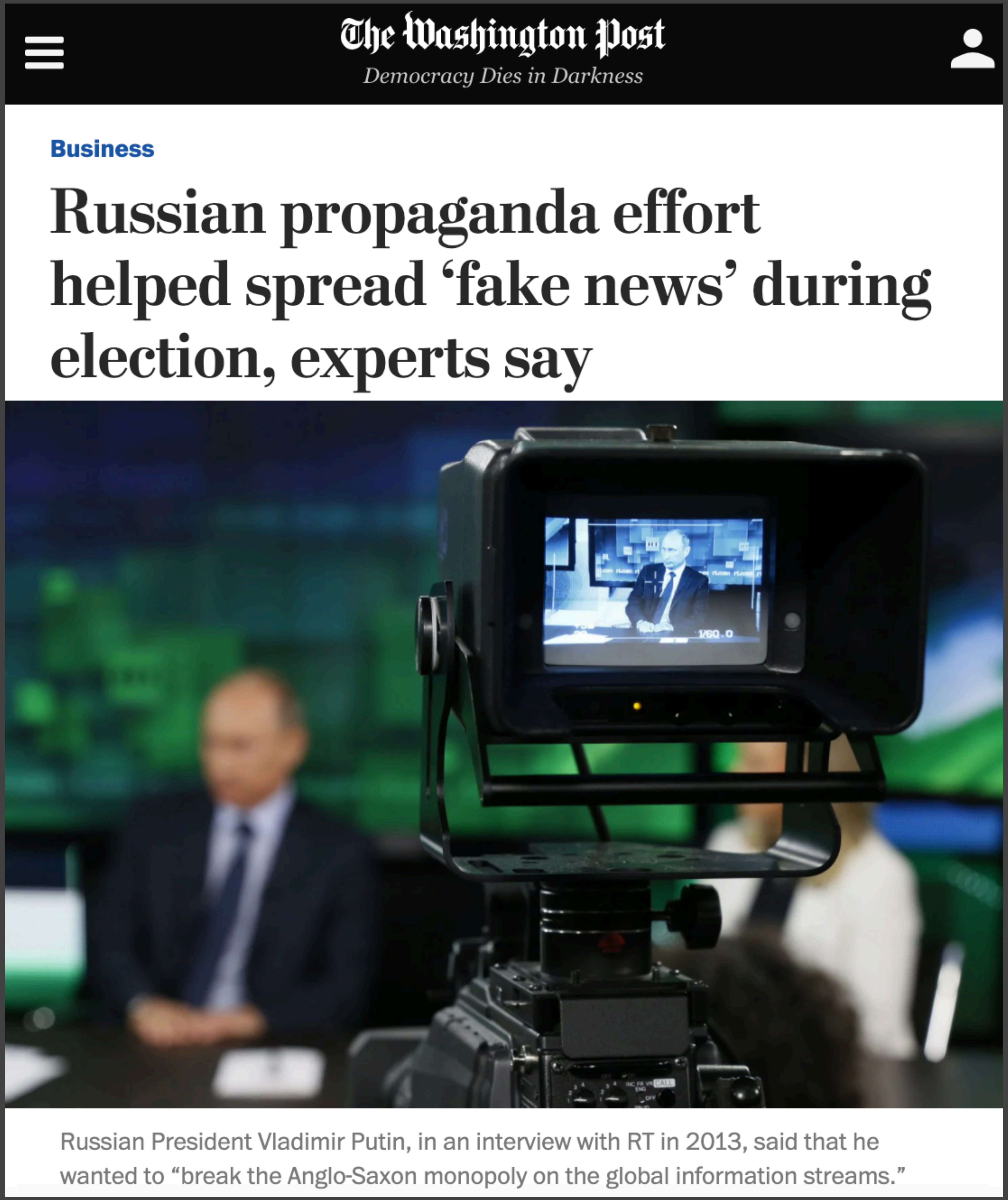
> teva @stiinken
>
> did anyone else's pinterest account get terminated out of the blue because mine did today and i had that account for 7 years and i feel like my library of alexandria has been burned down

7:08 AM · May 1, 2025

Pinterest mistakenly flagged artists’ posts as spam or misinformation. Despite widespread outrage, Pinterest stayed silent for days, leaving users to speculate. The Help Centre confirms AI is used to “improve content moderation,” but this opacity only fueled backlash and even a class action lawsuit.

0.5% extra credit for examples relevant to recent or upcoming lectures. Submit on Ed under the “Extra Credit” category

Attendance



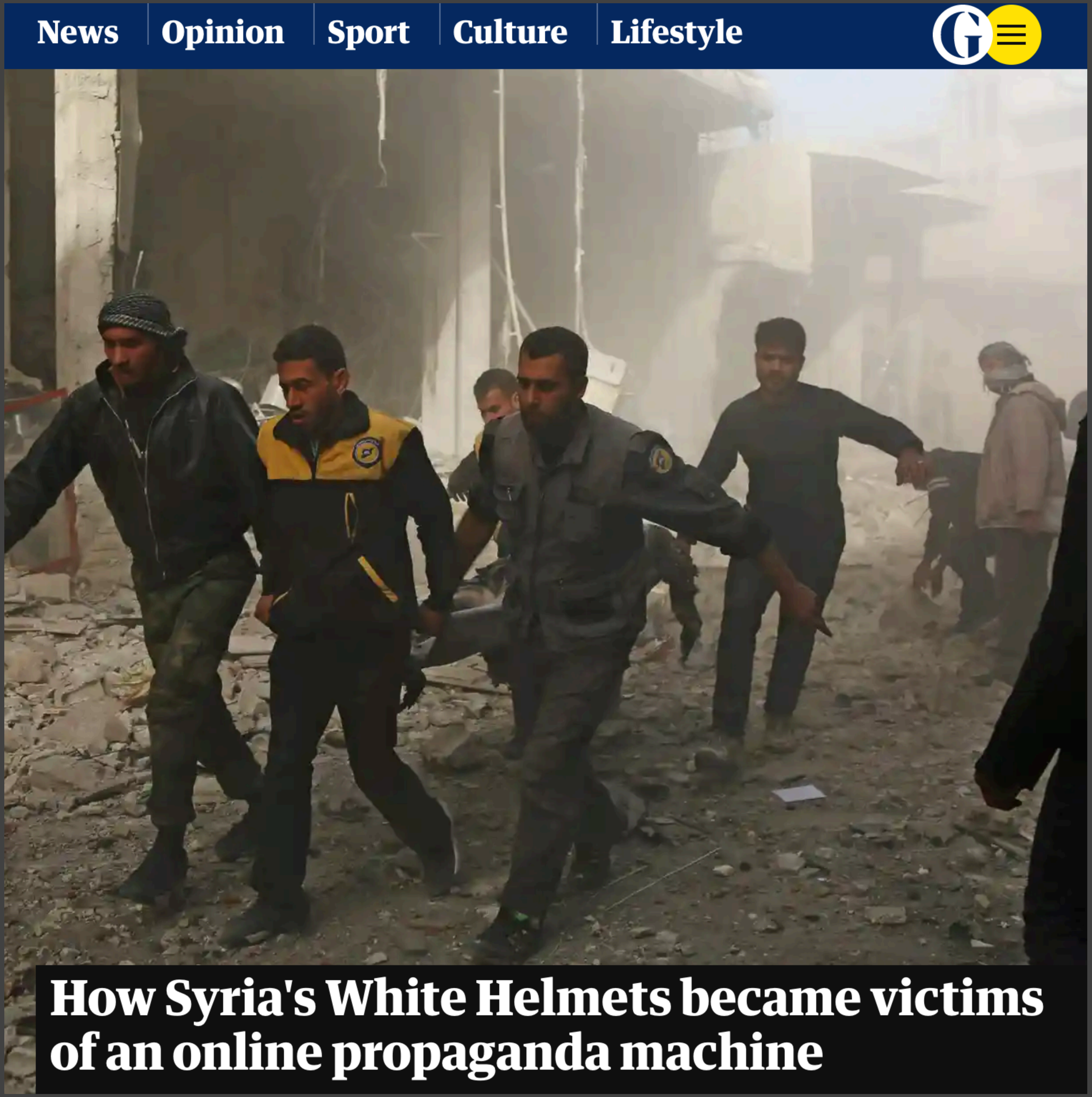
The Washington Post

Democracy Dies in Darkness

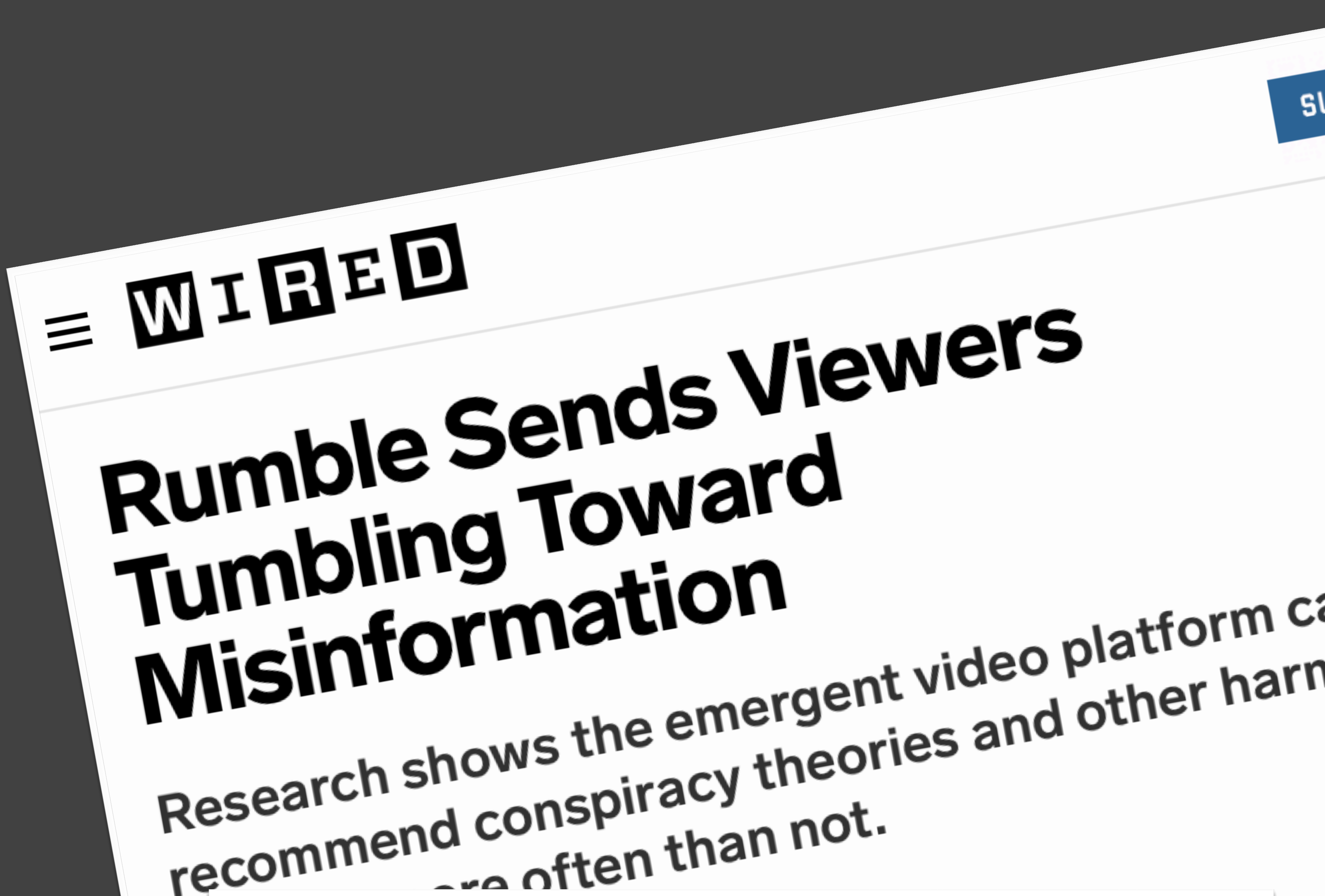
Business

Russian propaganda effort helped spread ‘fake news’ during election, experts say

Russian President Vladimir Putin, in an interview with RT in 2013, said that he wanted to “break the Anglo-Saxon monopoly on the global information streams.”



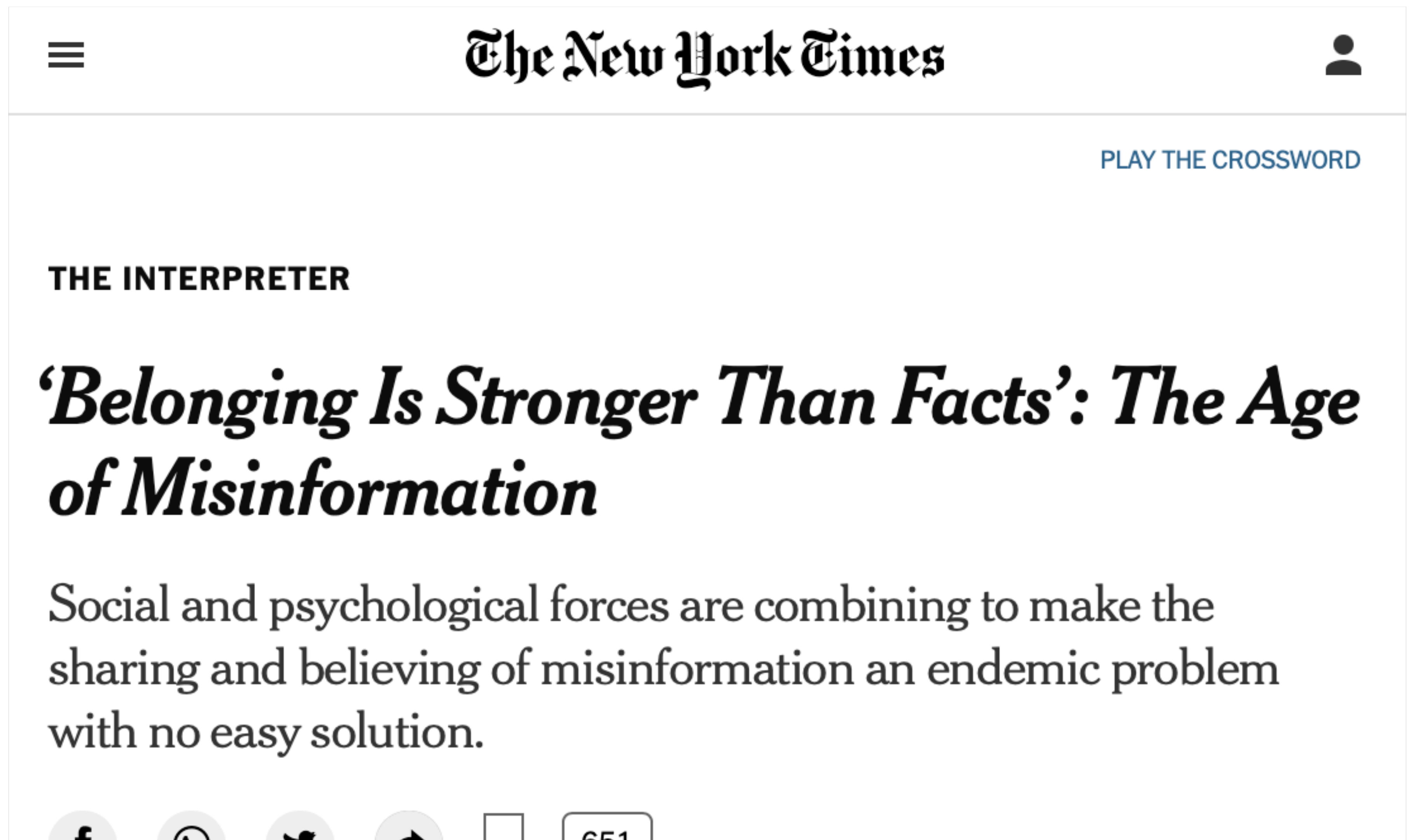
How Syria's White Helmets became victims of an online propaganda machine



Rumble Sends Viewers Tumbling Toward Misinformation

Research shows the emergent video platform can recommend conspiracy theories and other harmf …re often than not.

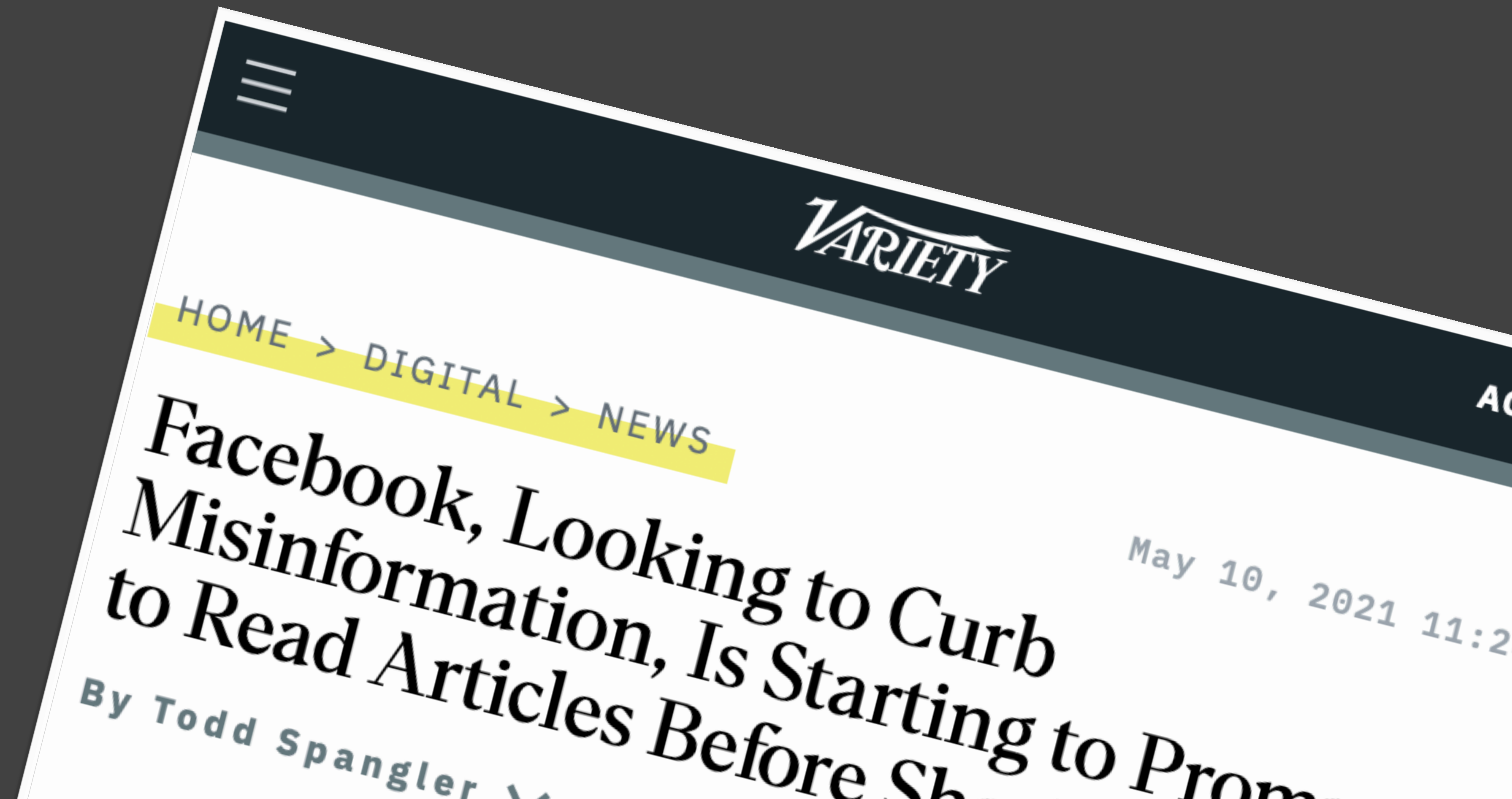
C



THE INTERPRETER

'Belonging Is Stronger Than Facts': The Age of Misinformation

Social and psychological forces are combining to make the sharing and believing of misinformation an endemic problem with no easy solution.

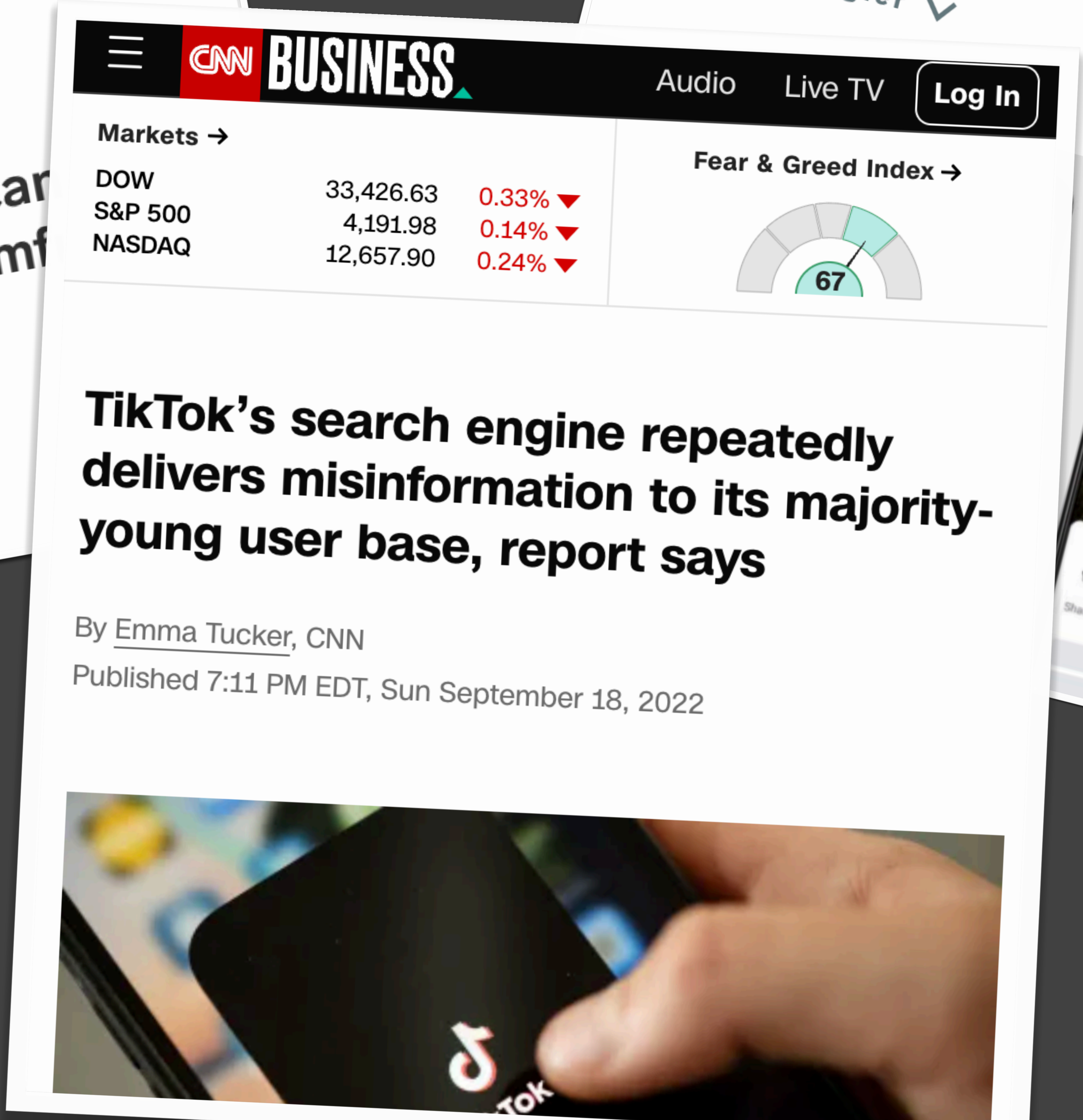


HOME > DIGITAL > NEWS

Facebook, Looking to Curb Misinformation, Is Starting to Prompt Use to Read Articles Before Sharing

May 10, 2021 11:2

By Todd Spangler



DOW	33,426.63	0.33% ▼
S&P 500	4,191.98	0.14% ▼
NASDAQ	12,657.90	0.24% ▼

TikTok's search engine repeatedly delivers misinformation to its majority-young user base, report says

By Emma Tucker, CNN

Published 7:11 PM EDT, Sun September 18, 2022

You're about to share this article without opening it

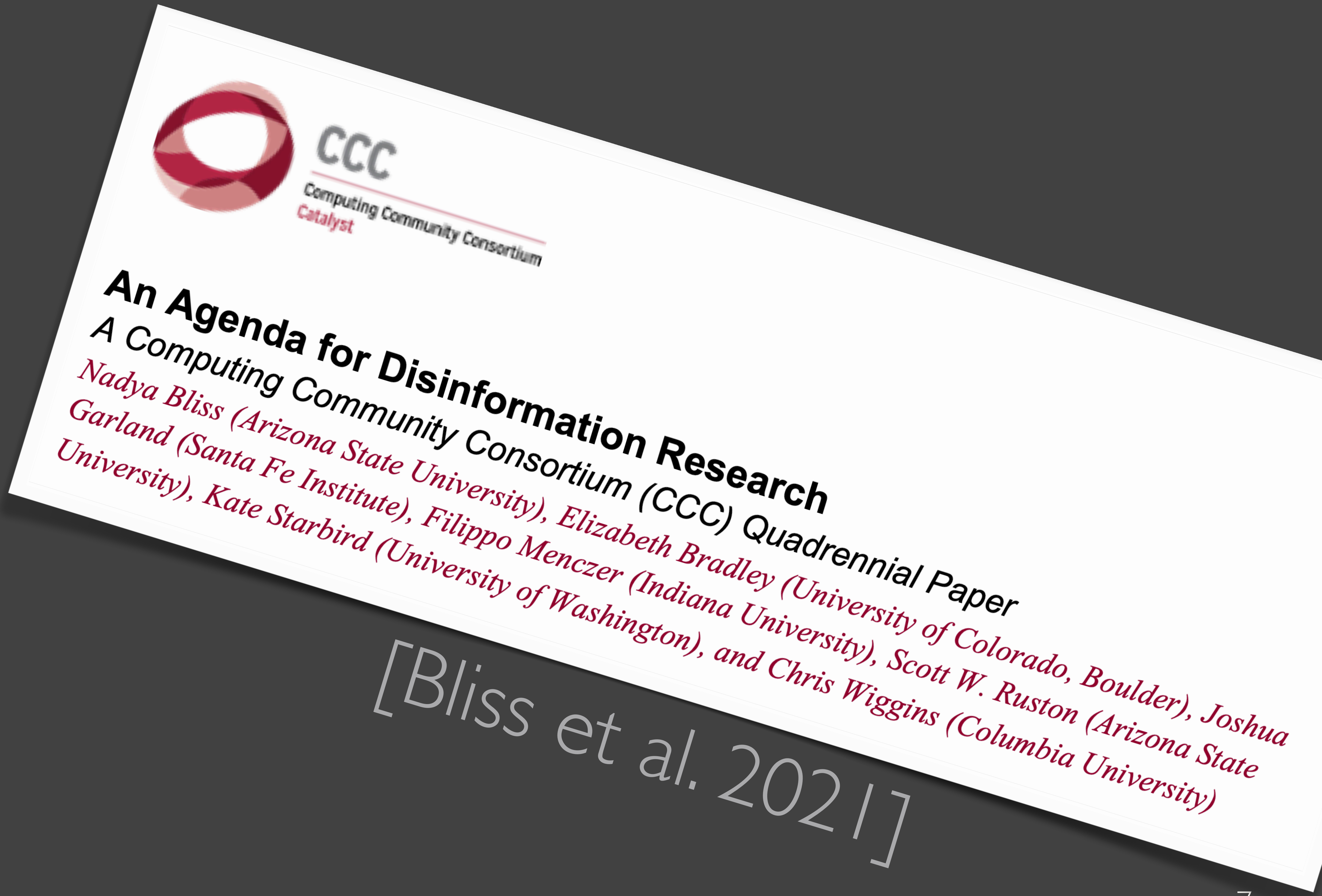
Sharing articles without reading them may mean missing key facts.

Open Article

Continue Sharing

The opportunity

"The digitization of information exchange, however, also makes the practices of disinformation detectable, the networks of influence discernable, and suspicious content characterizable."



CCC
Computing Community Consortium
Catalyst

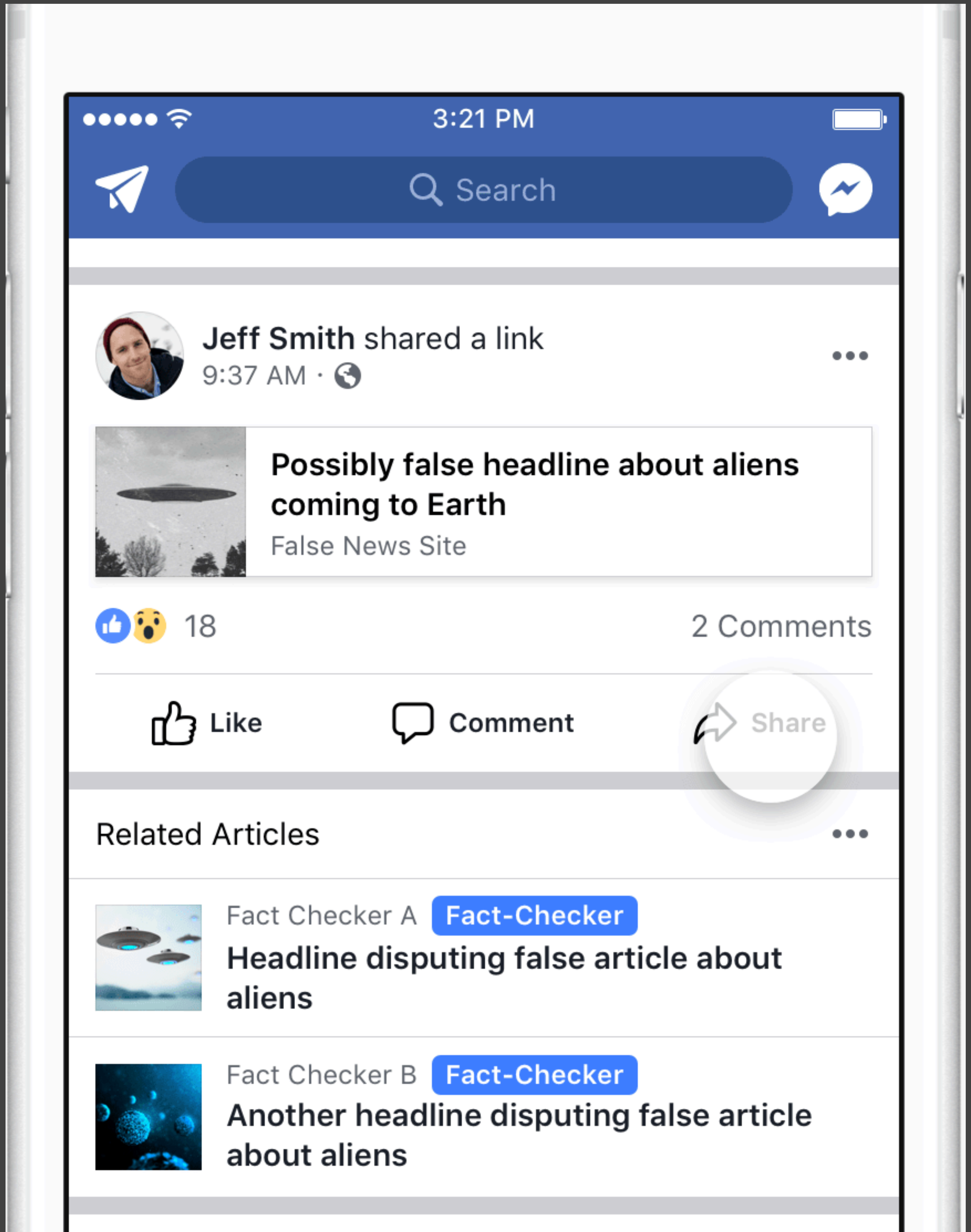
An Agenda for Disinformation Research

A Computing Community Consortium (CCC) Quadrennial Paper

Nadya Bliss (Arizona State University), Elizabeth Bradley (University of Colorado, Boulder), Joshua Garland (Santa Fe Institute), Filippo Menczer (Indiana University), Scott W. Ruston (Arizona State University), Kate Starbird (University of Washington), and Chris Wiggins (Columbia University)

[Bliss et al. 2021]

Poll: which design will better reduce the spread of disinformation?



3:21 PM

Search

Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

18 2 Comments

Like Comment Share

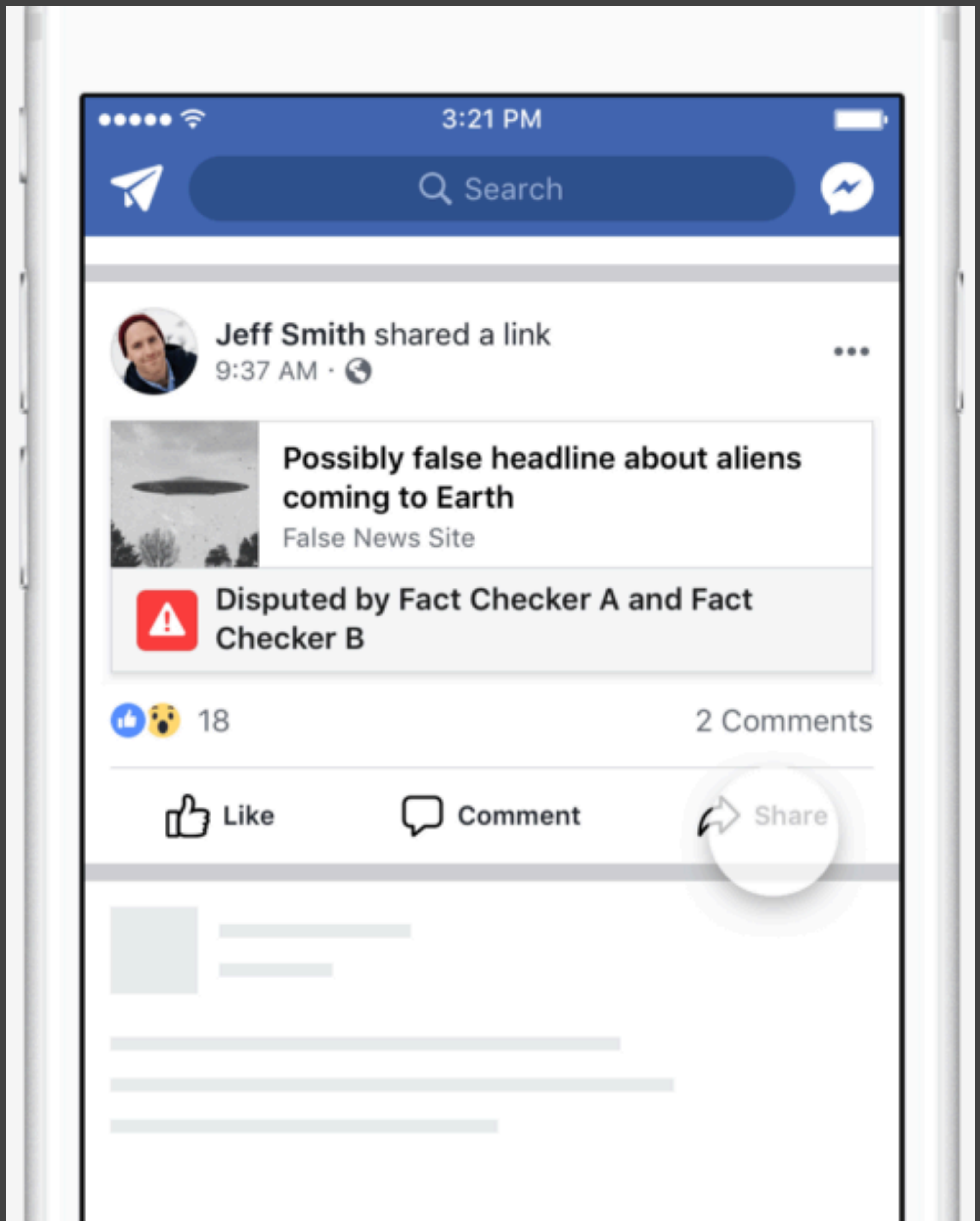
Related Articles

Fact Checker A Fact-Checker
Headline disputing false article about aliens

Fact Checker B Fact-Checker
Another headline disputing false article about aliens

related articles

fact check



3:21 PM

Search

Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

Disputed by Fact Checker A and Fact Checker B

18 2 Comments

Like Comment Share

From Whence Misinformation?

Let’s talk terms

When will I be referring to misinformation and when to disinformation throughout this lecture?

Misinformation = anything false

- Might be a rumor, or something not necessarily intentionally false

Disinformation = the specific intent is to deceive

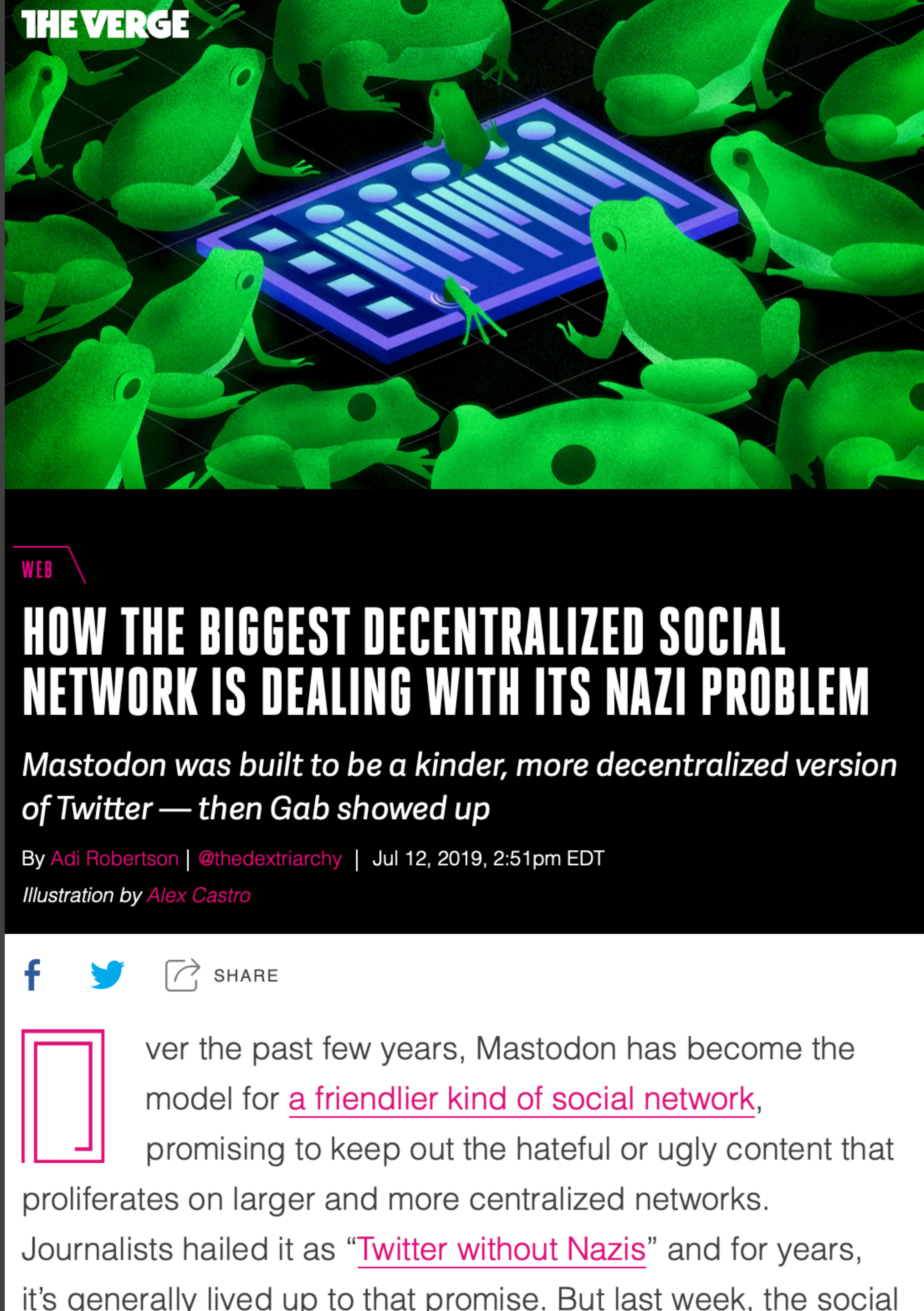
- Often built around a true or plausible core, wrapped up in a misleading way

Why now?

The effort required to connect groups together has lowered, making it possible for identity-based groups to connect that might otherwise have not:

Positive: social movements that are forced underground, such as LGBTQ military service members, can connect with each other online [Sheng 2020]

Negative: hate groups can also connect with each other online



THE VERGE

WEB

HOW THE BIGGEST DECENTRALIZED SOCIAL NETWORK IS DEALING WITH ITS NAZI PROBLEM

Mastodon was built to be a kinder, more decentralized version of Twitter — then Gab showed up

By Adi Robertson | @thedextriarchy | Jul 12, 2019, 2:51pm EDT

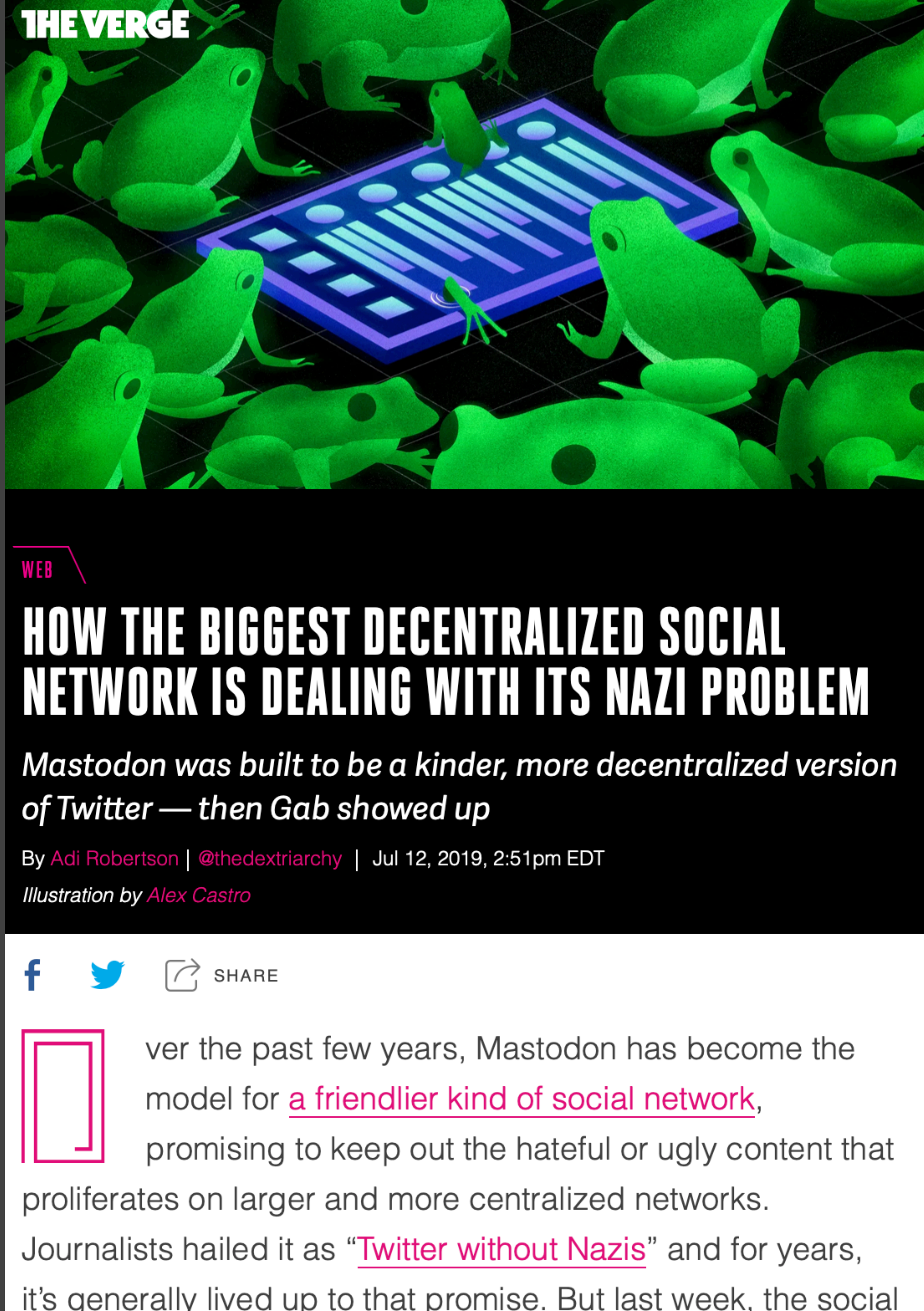
Illustration by Alex Castro

▯ver the past few years, Mastodon has become the model for a friendlier kind of social network, promising to keep out the hateful or ugly content that proliferates on larger and more centralized networks. Journalists hailed it as "Twitter without Nazis" and for years, it's generally lived up to that promise. But last week, the social

Why does it matter?

When groups can convene and push their own narrative, it enables "common knowledge attacks on democracy" [Farrell and Schneier 2018]

In other words, it can destabilize democracy by flooding public debate and confusing our shared understandings and expectations, which are required for democracy to function



THE VERGE

WEB

HOW THE BIGGEST DECENTRALIZED SOCIAL NETWORK IS DEALING WITH ITS NAZI PROBLEM

Mastodon was built to be a kinder, more decentralized version of Twitter — then Gab showed up

By Adi Robertson | @thedextriarchy | Jul 12, 2019, 2:51pm EDT

Illustration by Alex Castro

ver the past few years, Mastodon has become the model for a friendlier kind of social network, promising to keep out the hateful or ugly content that proliferates on larger and more centralized networks. Journalists hailed it as "Twitter without Nazis" and for years, it's generally lived up to that promise. But last week, the social

Fingers pointed 👉

#1, "It's trolls": disinformation factories generate disinformation to harm us [Bail et al. 2020]

#2, "Post truth": people default to motivated reasoning, which means that we are inclined to believe information that is consistent with our political views, and disinclined to believe information that contradicts our political views [Kahan 2017]. We are more loyal to political party than loyal to truth [Van Bavel and Pereira 2018]

Which, neither, or both? [1 min]

While these explanations are not wrong, they are also not the explanations with the strongest evidence

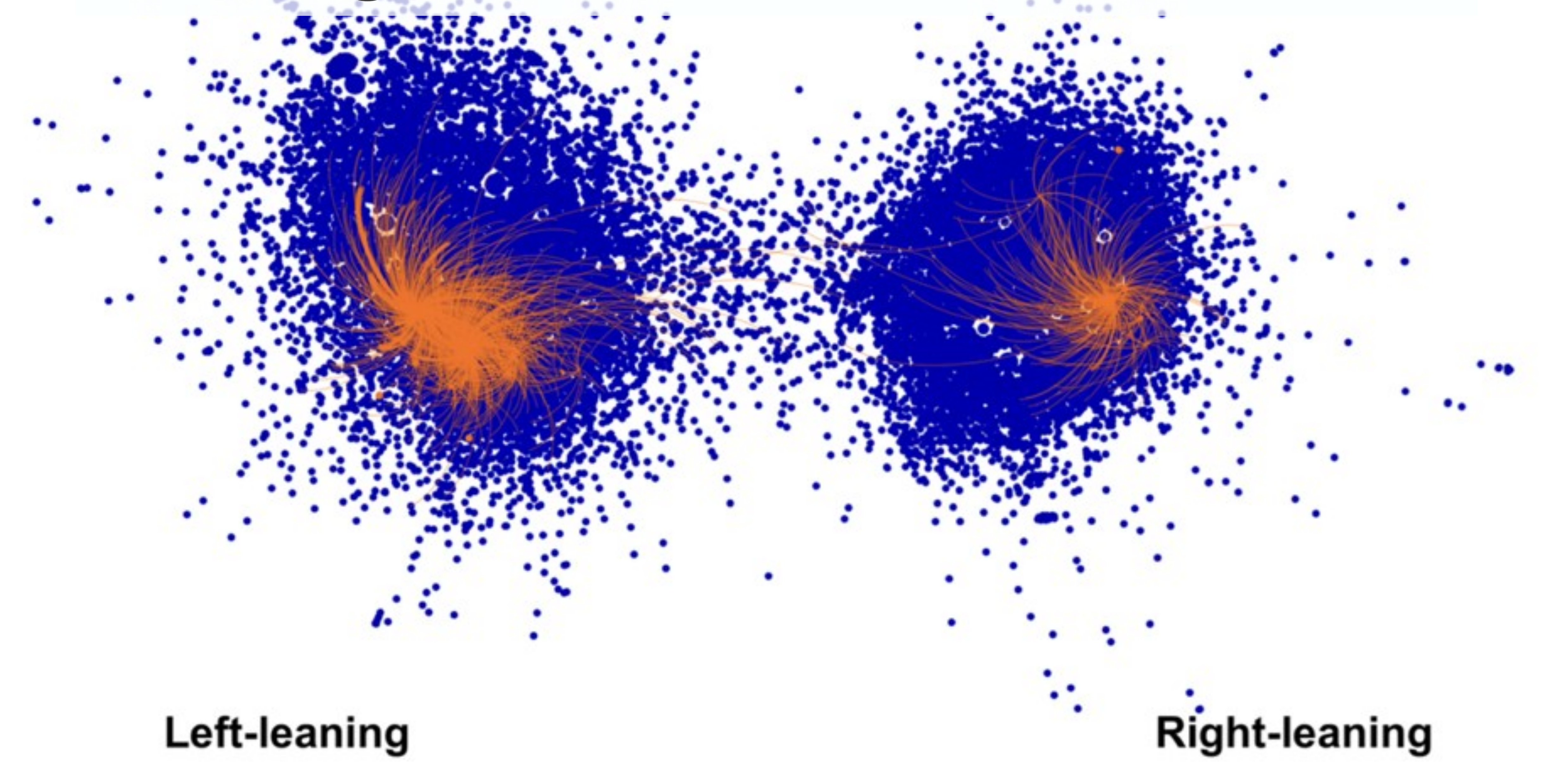
Finger #1 👉 state actors

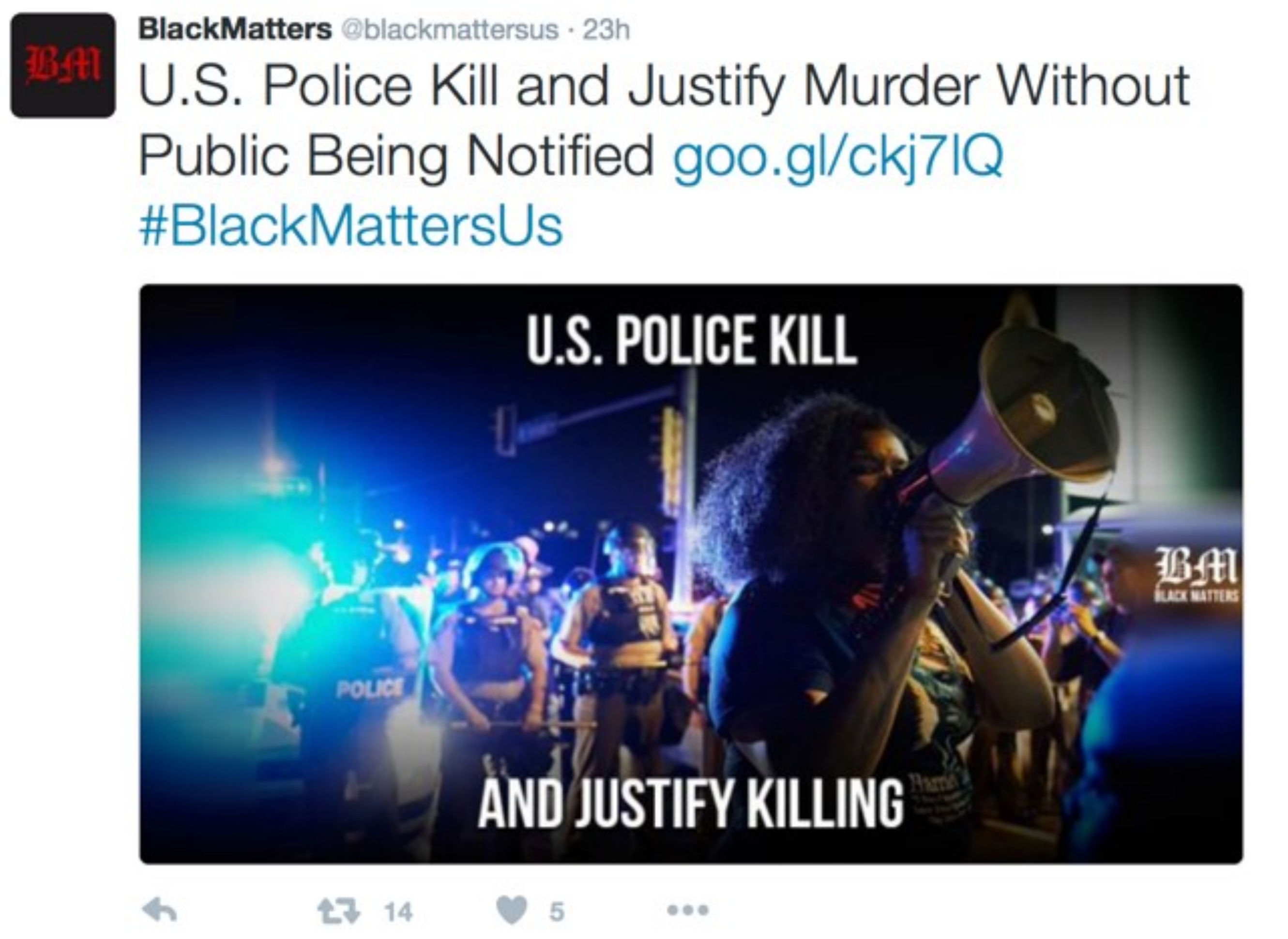
Yes, state actors exist.

Twitter retweet network for Black Lives Matter in 2016: Russian IRA (orange) both posed as BLM activists on the left, and infiltrated anti-BLM communities on the right

[Starbird, Arif, and Wilson 2019]

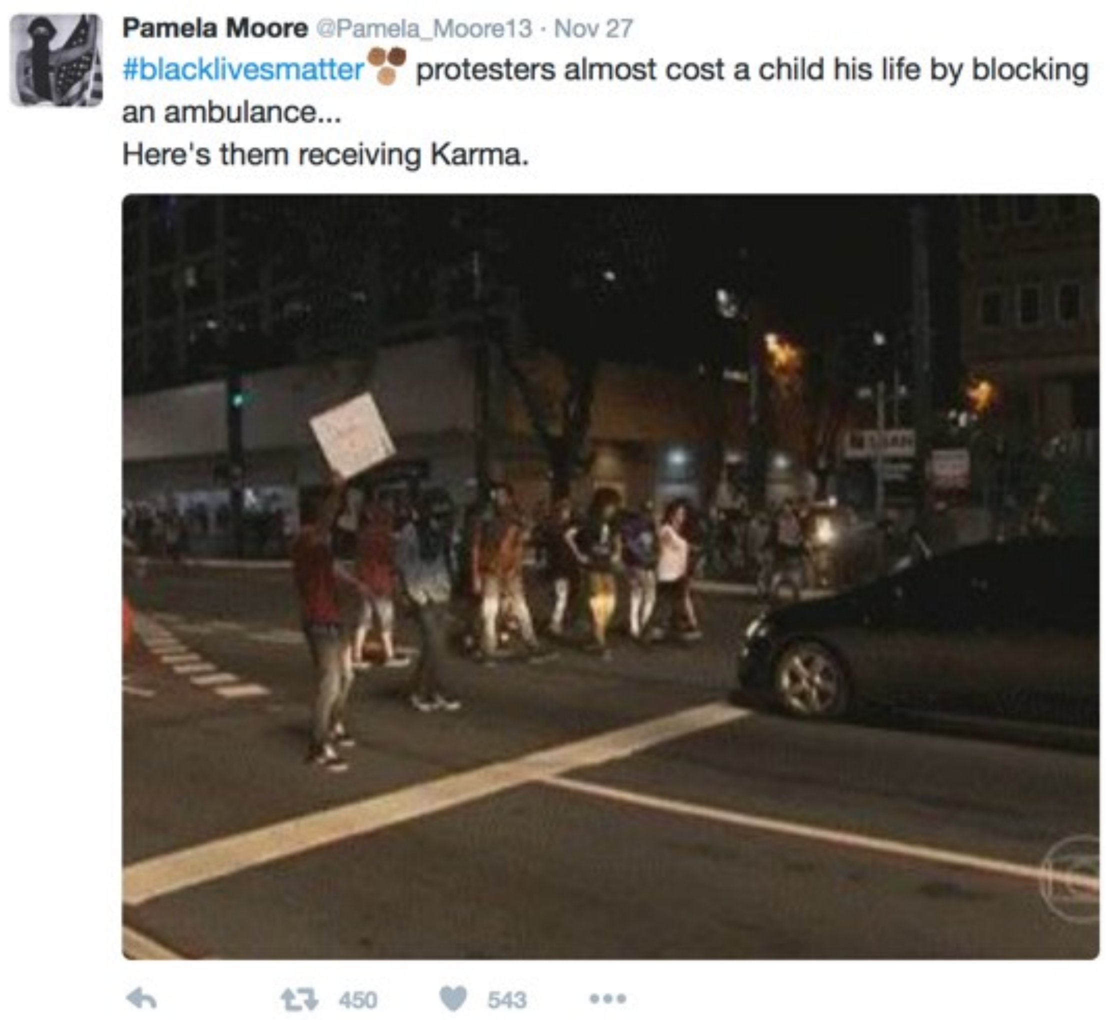
Orange: accounts Twitter removed as being Russian IRA





BlackMatters @blackmattersus · 23h

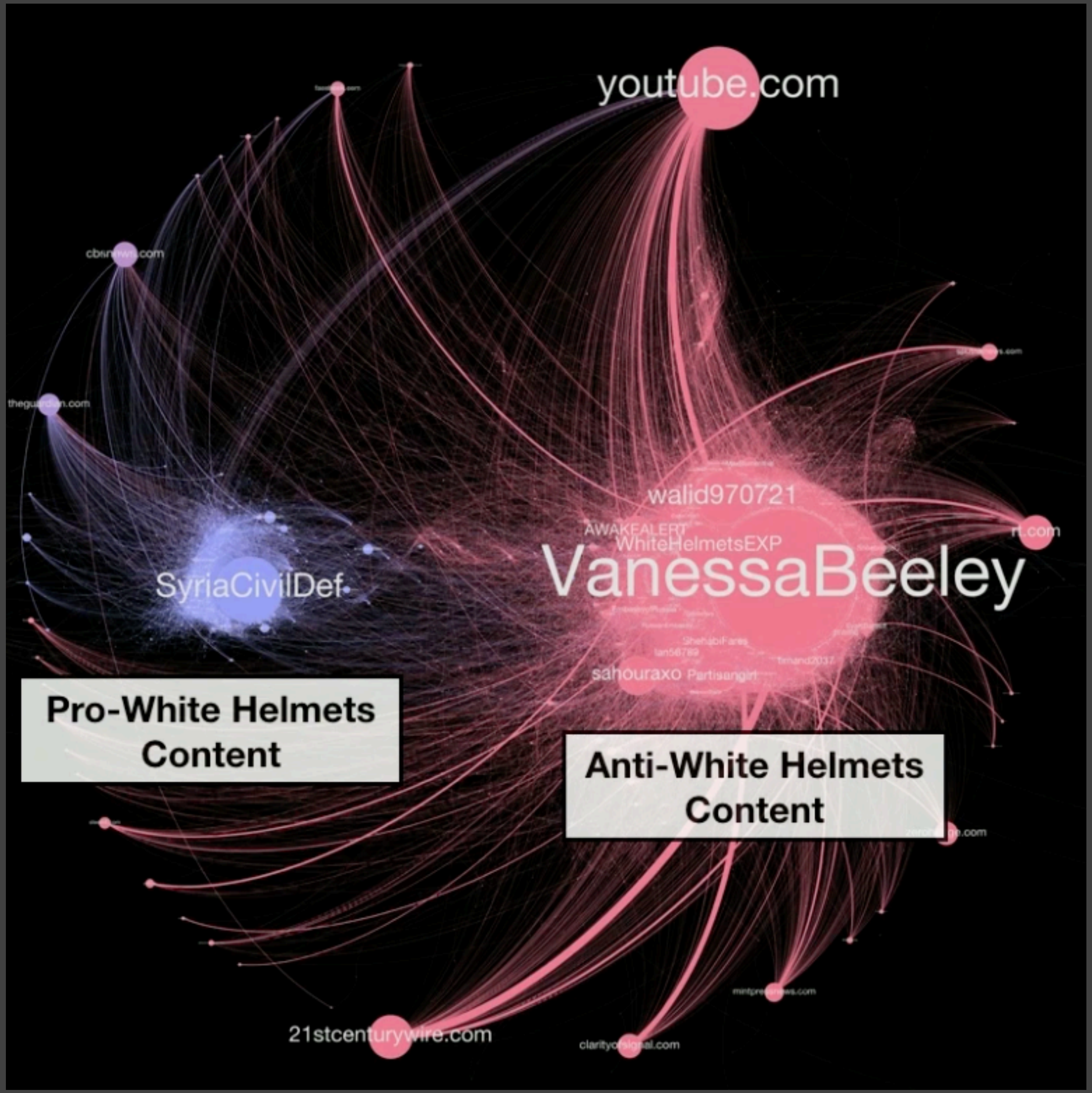
U.S. Police Kill and Justify Murder Without Public Being Notified goo.gl/ckj7lQ #BlackMattersUs



Pamela Moore @Pamela_Moore13 · Nov 27

#blacklivesmatter protesters almost cost a child his life by blocking an ambulance...
Here's them receiving Karma.

But is it just state actors?



[Starbird @ CS 547, 2019]

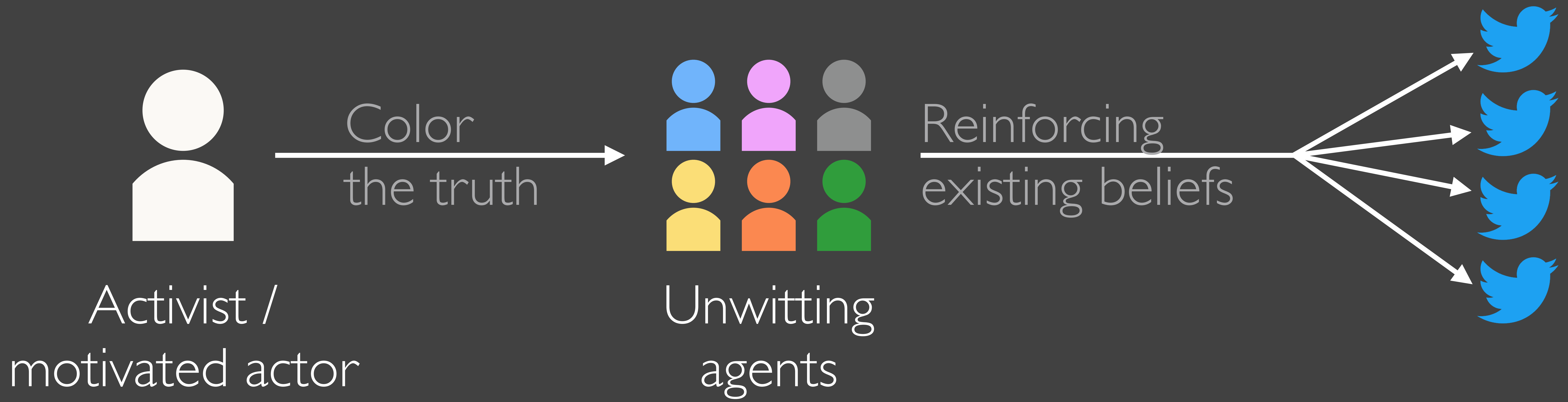
Context is the Syrian Civil War and the White Helmets, a humanitarian response group. Anti-White Helmet accounts — pink — are dominant in volume, delegitimizing the White Helmets' claims

Not just bots and trolls: lots are journalists aligned with Syrian and Russian government interests, Syrian and Russian government members, and alt. media

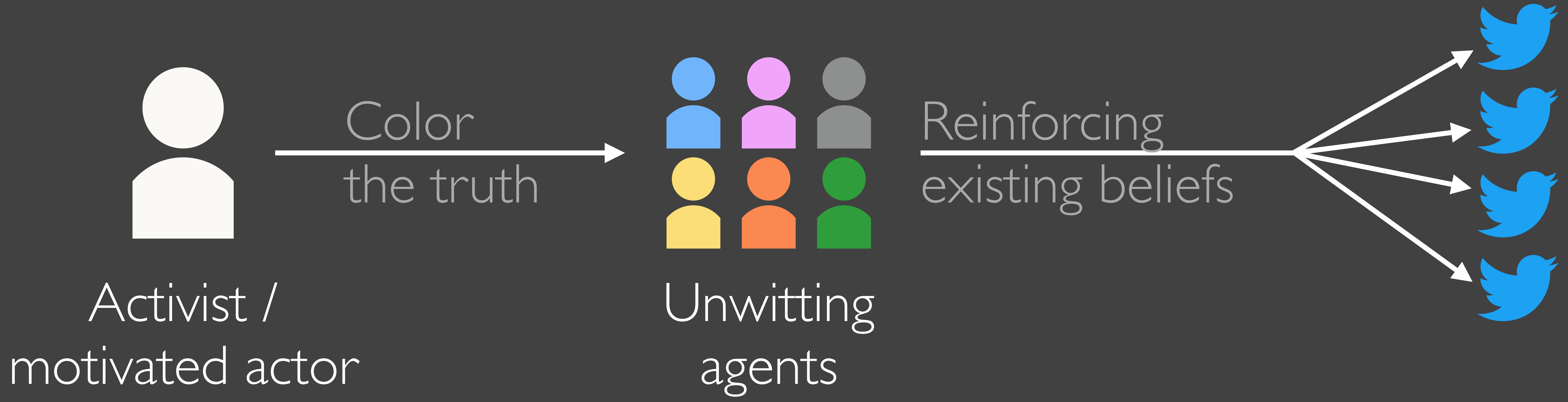
It looks more like activism than "just" disinformation

But is it just state actors?

Disinformation campaigns often involve many unwitting agents who are unaware of their role and whose views and behaviors have been shaped by motivated actors [Bittman 1985, Starbird, Arif, and Wilson 2019]



Cold War-era Soviet technique: sell journalists on anonymous tips aligned with their beliefs. Once one journalist took the bait, others became interested.

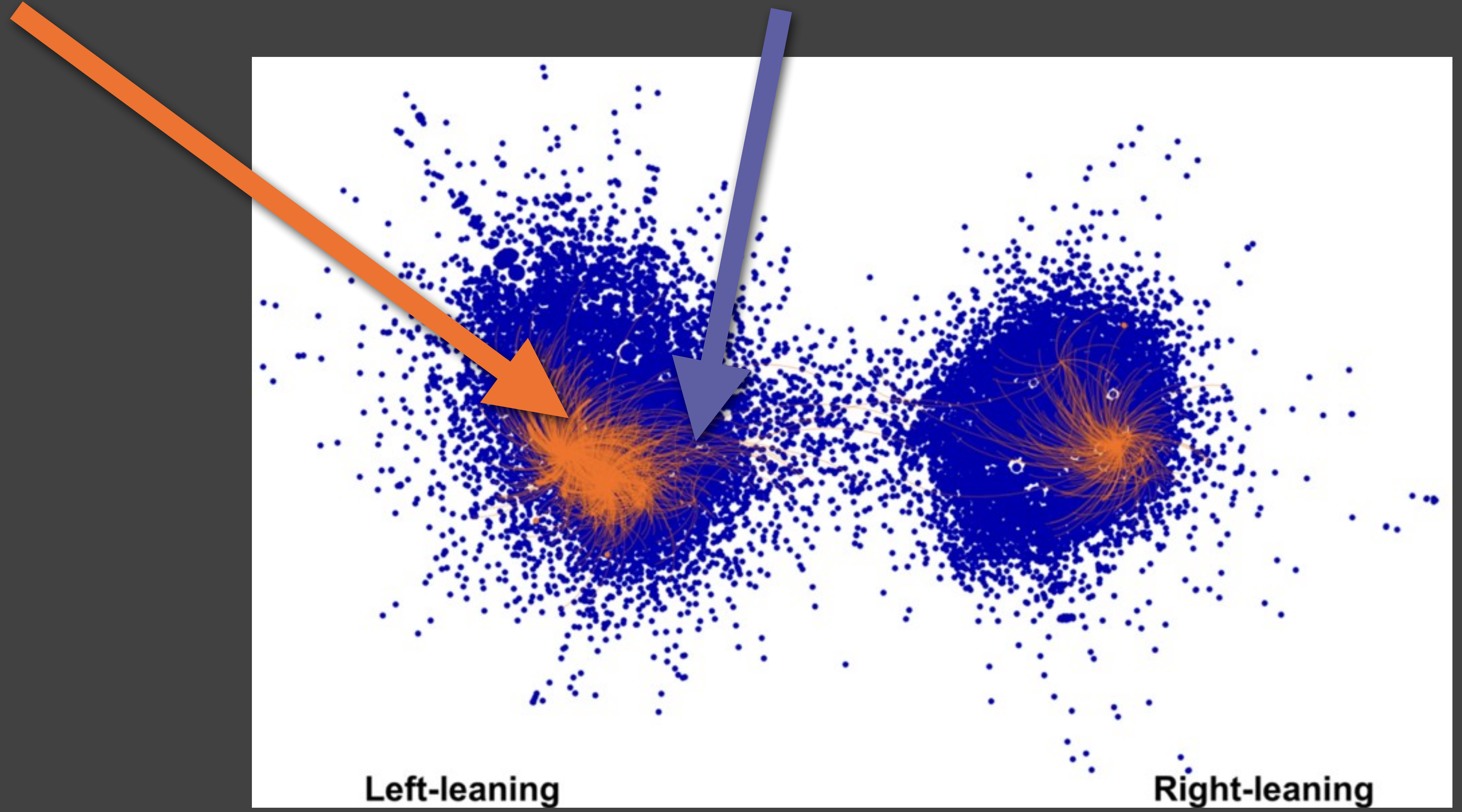


Activist / motivated actor

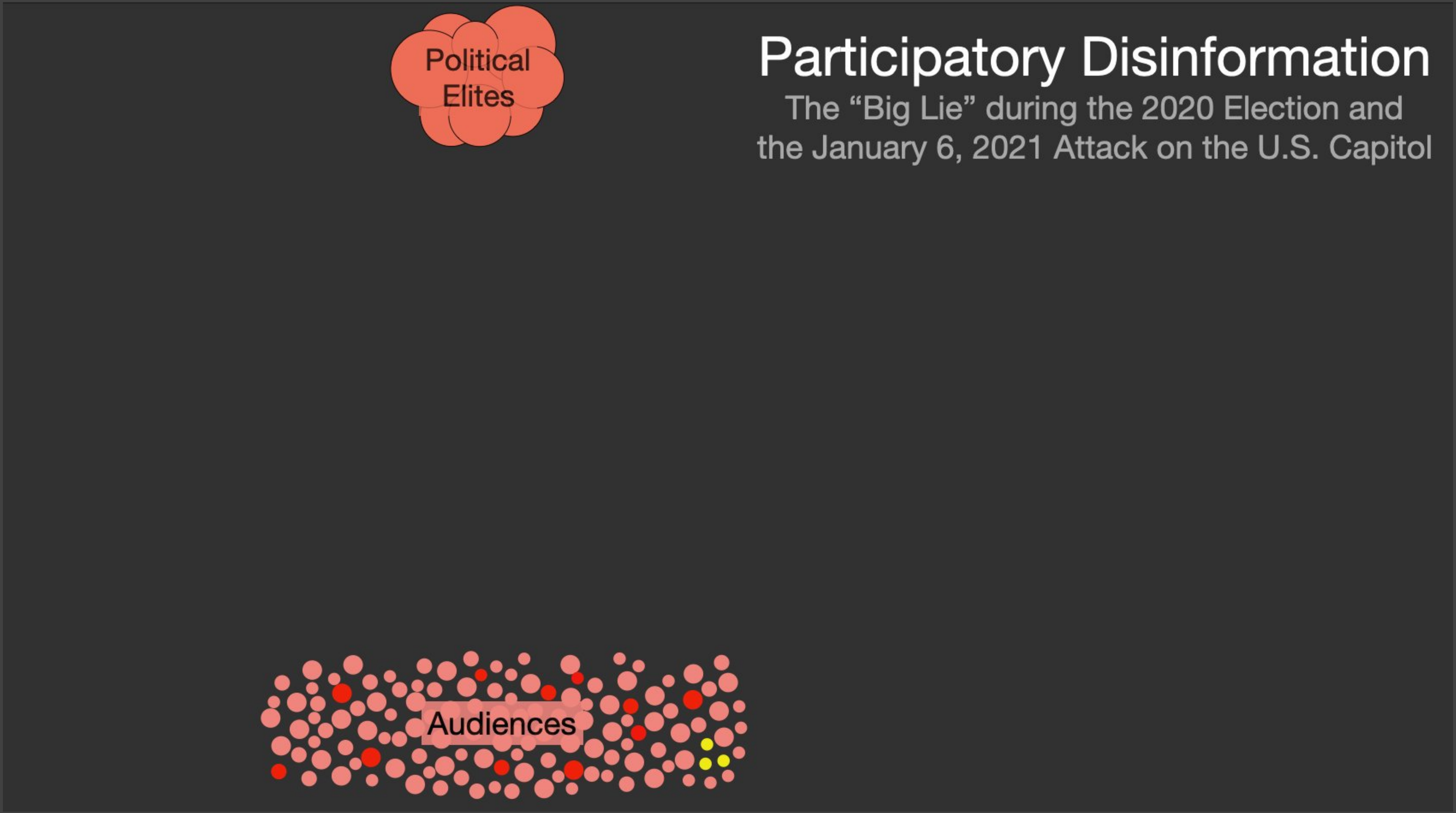
Color the truth

Unwitting agents

Reinforcing existing beliefs



Orange arcs = blue accounts retweeting content from orange accounts = unwitting agents



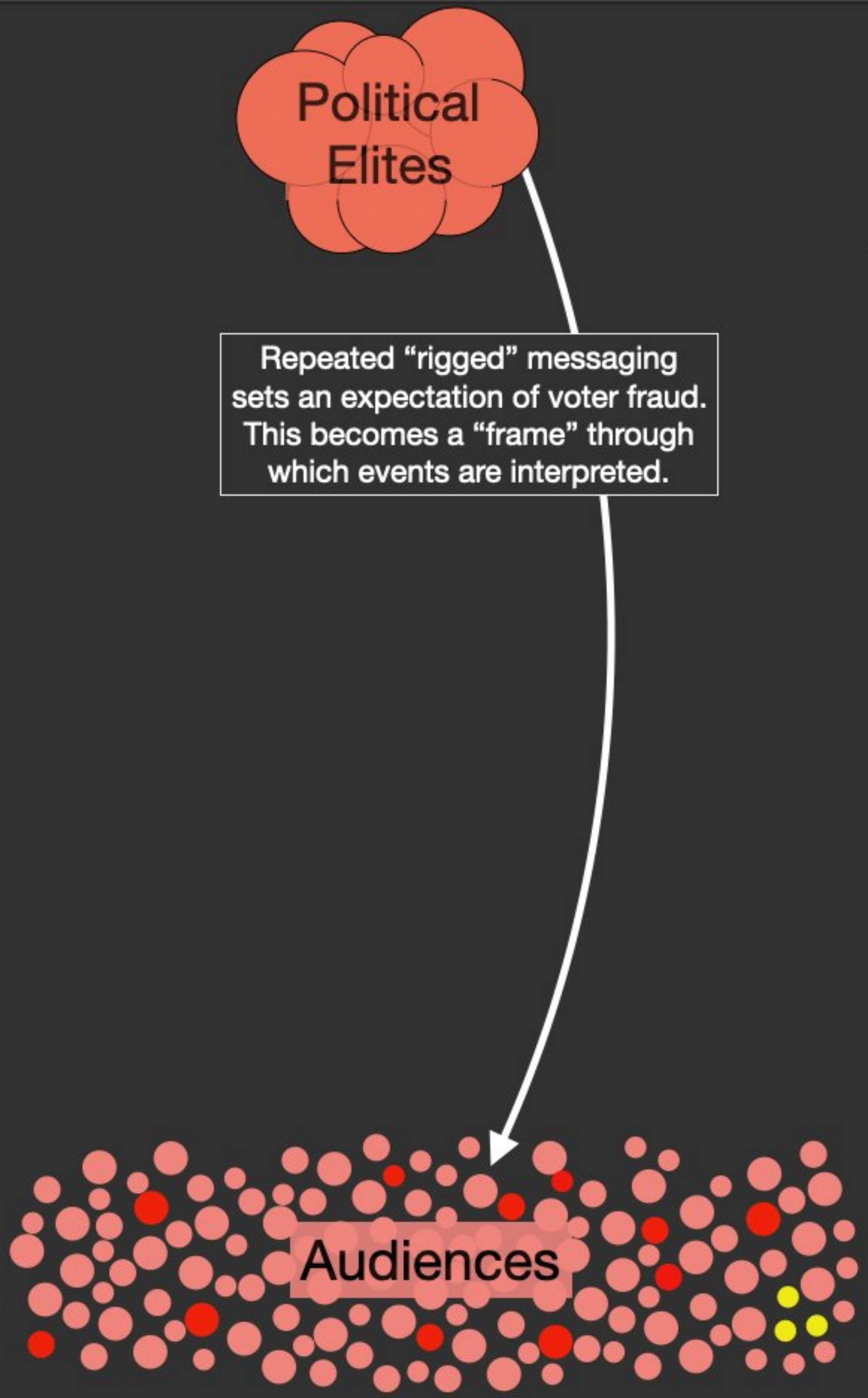
Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

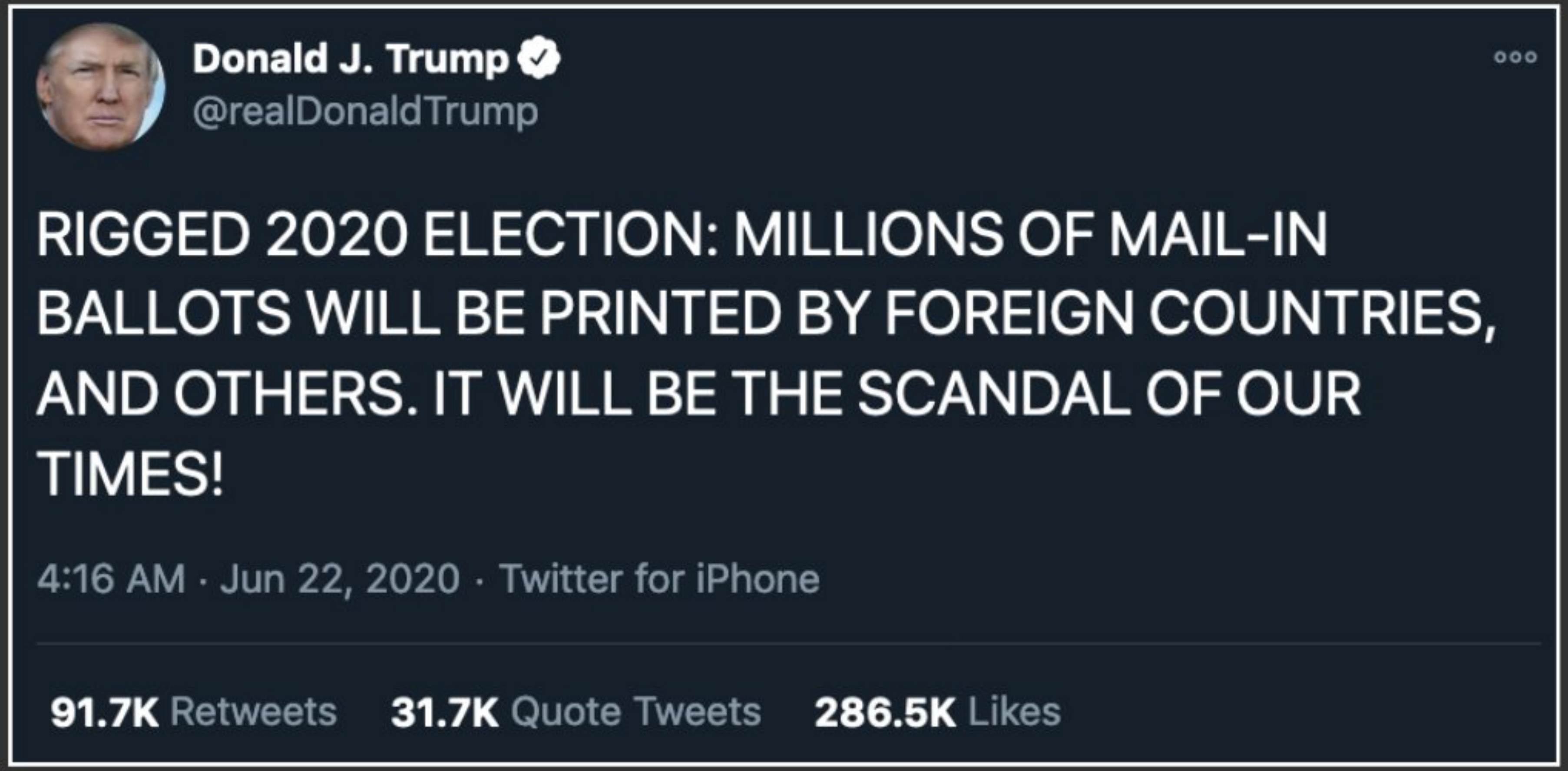
[Starbird 2021; Prochaska 2023]

Participatory Disinformation

The "Big Lie" during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol



Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.



Donald J. Trump
@realDonaldTrump

RIGGED 2020 ELECTION: MILLIONS OF MAIL-IN BALLOTS WILL BE PRINTED BY FOREIGN COUNTRIES, AND OTHERS. IT WILL BE THE SCANDAL OF OUR TIMES!

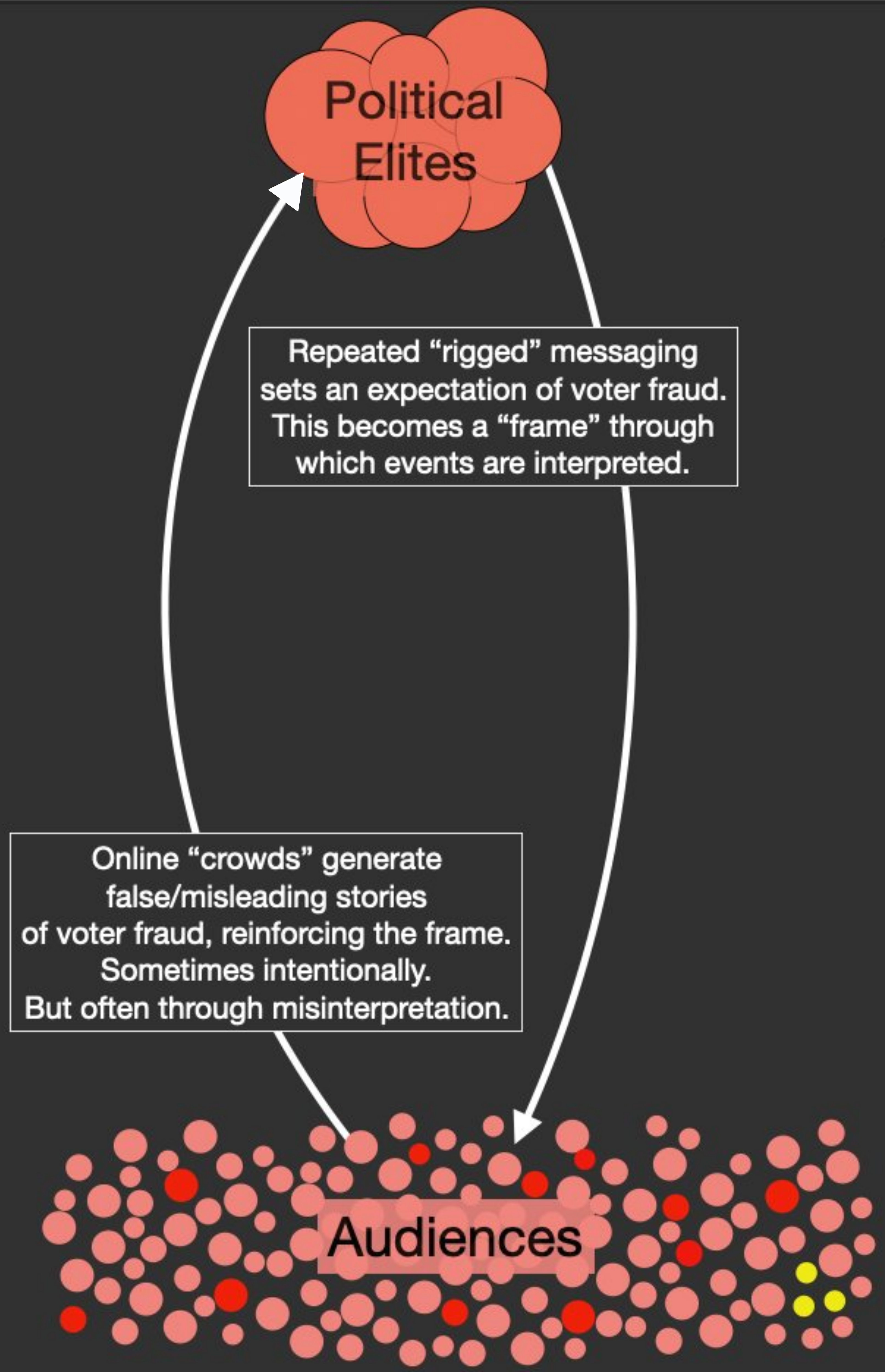
4:16 AM · Jun 22, 2020 · Twitter for iPhone

91.7K Retweets **31.7K** Quote Tweets **286.5K** Likes

[Starbird 2021; Prochaska 2023]

Participatory Disinformation

The "Big Lie" during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol



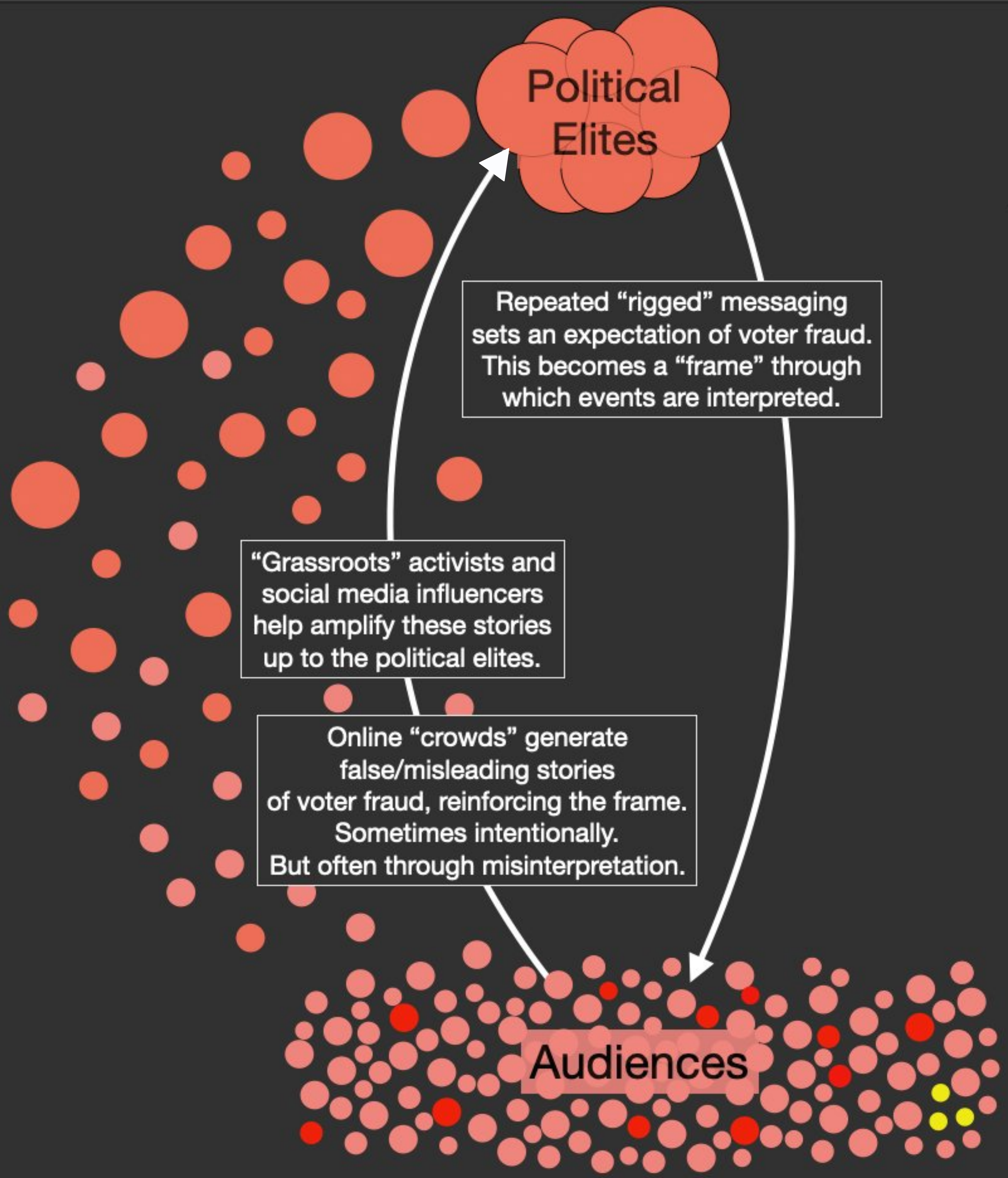
Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.

Online "crowds" generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

[Starbird 2021; Prochaska 2023]

Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol



Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.

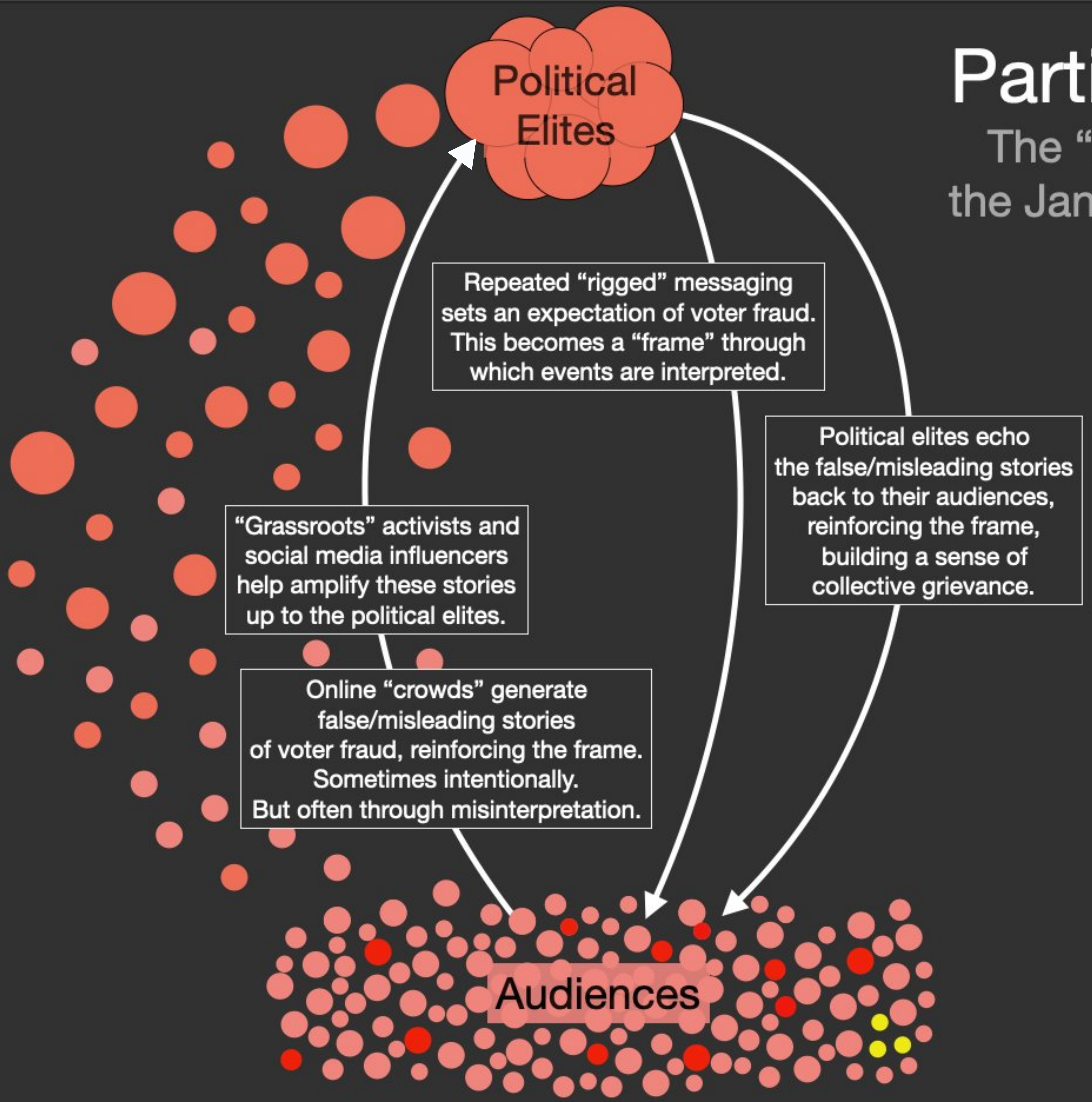
“Grassroots” activists and social media influencers help amplify these stories up to the political elites.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

[Starbird 2021; Prochaska 2023]

Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol



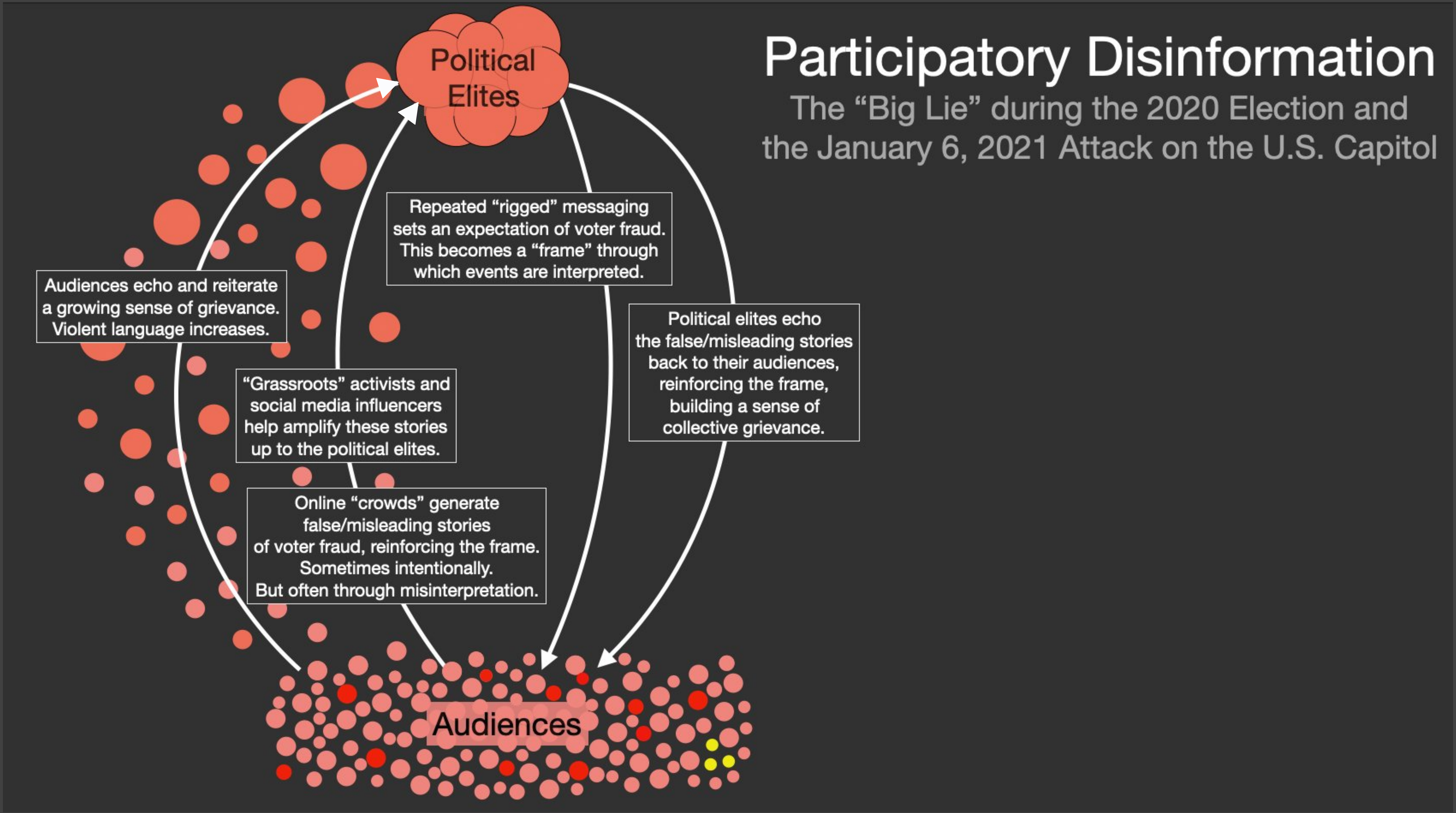
Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.

Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.

“Grassroots” activists and social media influencers help amplify these stories up to the political elites.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

[Starbird 2021; Prochaska 2023]



Participatory Disinformation

The "Big Lie" during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.

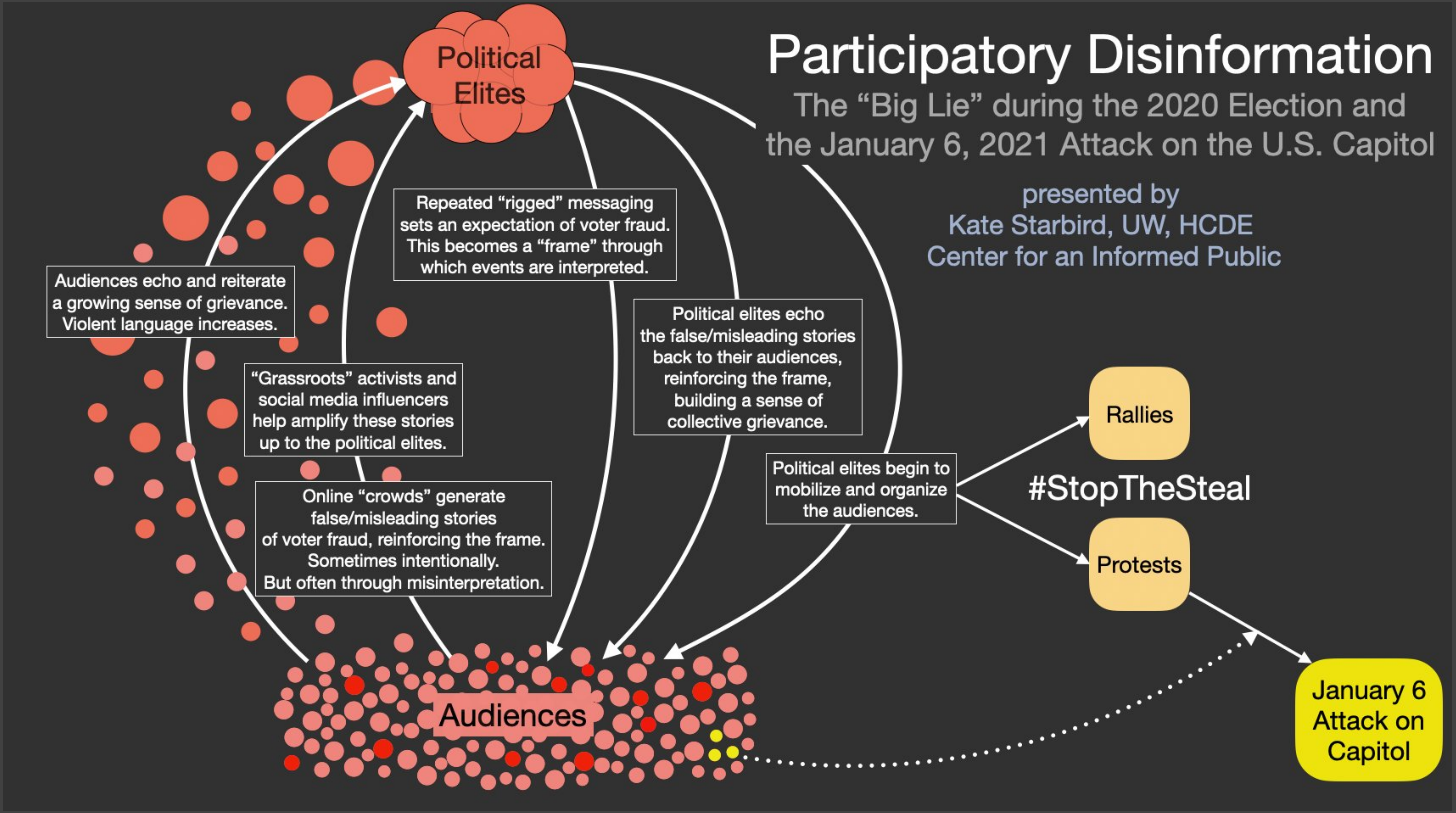
Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.

Online "crowds" generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

"Grassroots" activists and social media influencers help amplify these stories up to the political elites.

Audiences echo and reiterate a growing sense of grievance. Violent language increases.

[Starbird 2021; Prochaska 2023]



Participatory Disinformation

The "Big Lie" during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

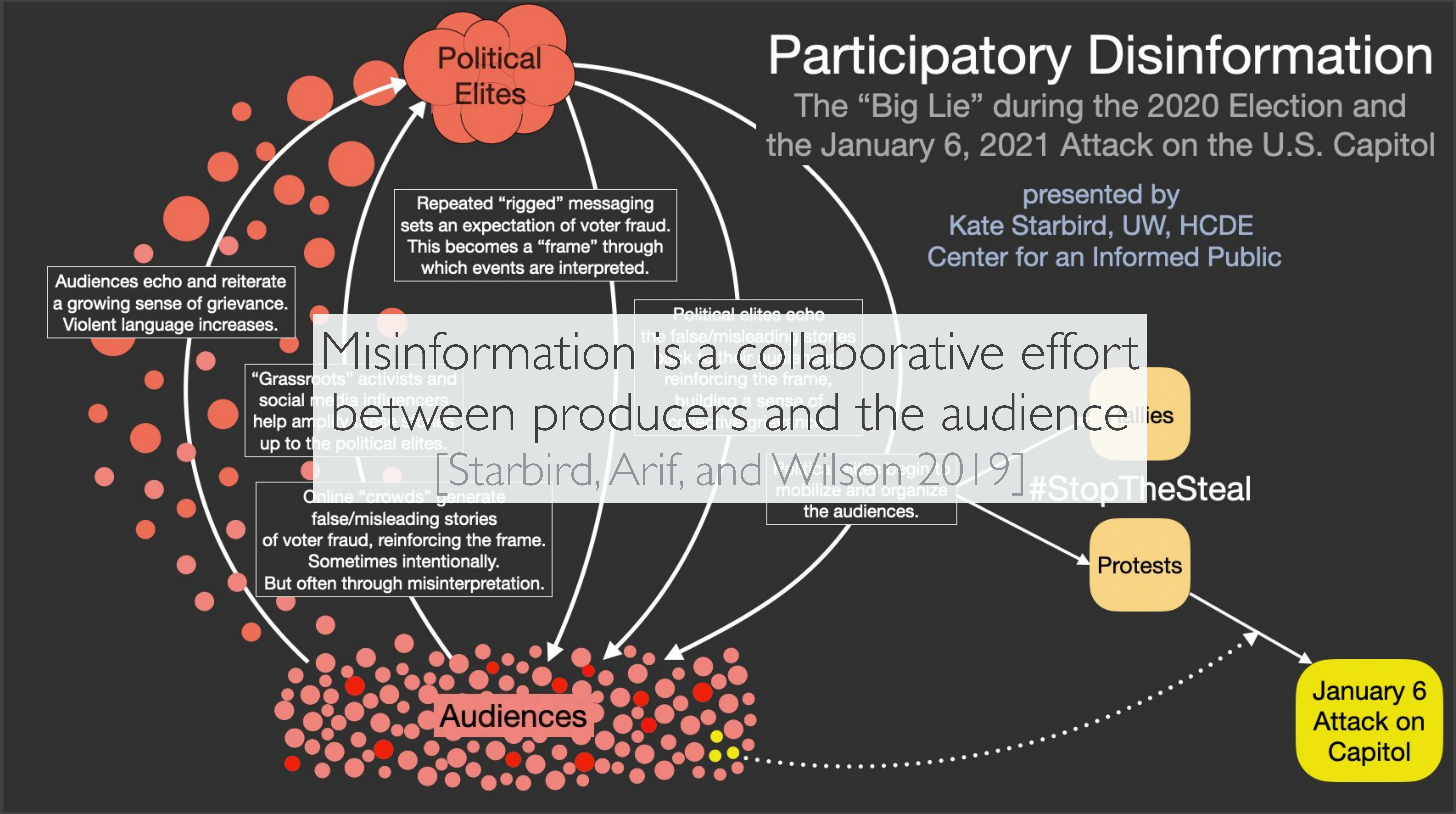
presented by
Kate Starbird, UW, HCDE
Center for an Informed Public

Political Elites
Repeated "rigged" messaging sets an expectation of voter fraud. This becomes a "frame" through which events are interpreted.
Audiences echo and reiterate a growing sense of grievance. Violent language increases.
"Grassroots" activists and social media influencers help amplify these stories up to the political elites.
Online "crowds" generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.
Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.
Political elites begin to mobilize and organize the audiences.
Audiences
Rallies
#StopTheSteal
Protests
January 6 Attack on Capitol

[Starbird 2021; Prochaska 2023]

Misinformation is a collaborative effort between producers and the audience

[Starbird, Arif, and Wilson 2019]



Participatory Disinformation

The “Big Lie” during the 2020 Election and the January 6, 2021 Attack on the U.S. Capitol

presented by
Kate Starbird, UW, HCDE
Center for an Informed Public

Political Elites

Repeated “rigged” messaging sets an expectation of voter fraud. This becomes a “frame” through which events are interpreted.

Political elites echo the false/misleading stories back to their audiences, reinforcing the frame, building a sense of collective grievance.

Audiences echo and reiterate a growing sense of grievance. Violent language increases.

“Grassroots” activists and social media influencers help amplify these stories up to the political elites.

Online “crowds” generate false/misleading stories of voter fraud, reinforcing the frame. Sometimes intentionally. But often through misinterpretation.

Political elites begin to mobilize and organize the audiences.

Rallies

#StopTheSteal

Protests

January 6 Attack on Capitol

Audiences

[Starbird 2021; Prochaska 2023]

Finger #2 👉 post-truth

While we are more likely to believe news that is concordant with our beliefs, the larger effect is whether we engage with higher-level reasoning instead of automatic reasoning [Pennycook and Rand 2021]

More like post-attention…

[Pennycook et al. 2021]

People rate accuracy as the single most important factor when deciding whether to share

However, whether a headline is politically concordant has a much larger effect on sharing intention than the headline's accuracy

So what gives?

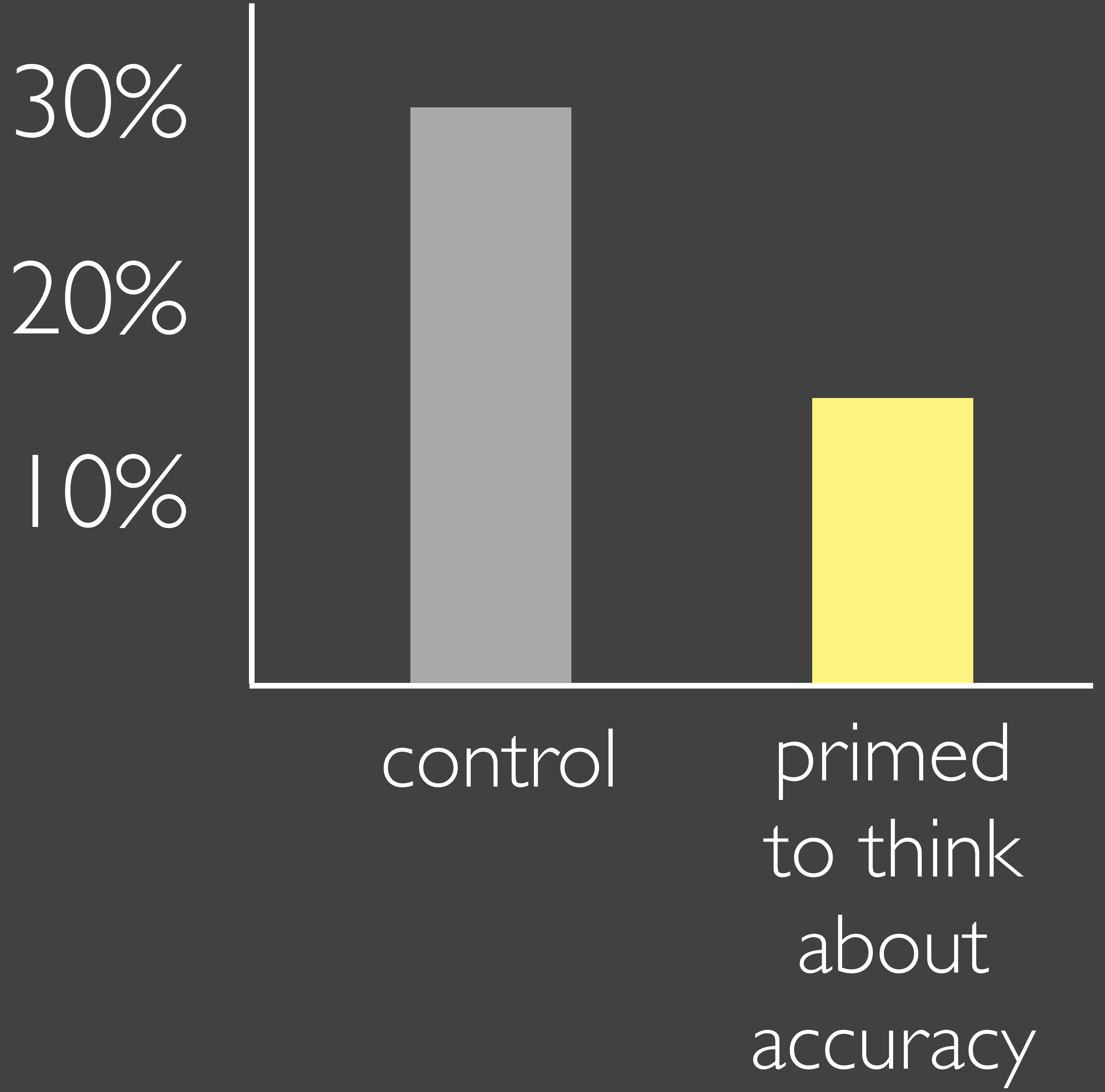
More like post-attention…

[Pennycook et al. 2021; Pennycook and Rand 2022]

Theory: we don't pay attention to accuracy, and are more focused on pleasing followers or signaling group membership

Evidence: focusing participants' attention on accuracy before seeing a headline reduces sharing of false headlines by half

% of people likely to share false headline



How much do we consume?

Most people rarely see misinformation.

National sample of mobile, desktop, and TV consumption: misinformation is 0.15% of Americans' media diet [Allen et al. 2020]

The average US adult saw ~1 misinformation story in the 2016 election [Hunt and Gentzkow 2017]

Exposure to misinformation is highly concentrated [Guess, Nyhan, and Reifler 2020]: 1% of people account for 80% of exposures to misinformation [Grinberg et al. 2019]

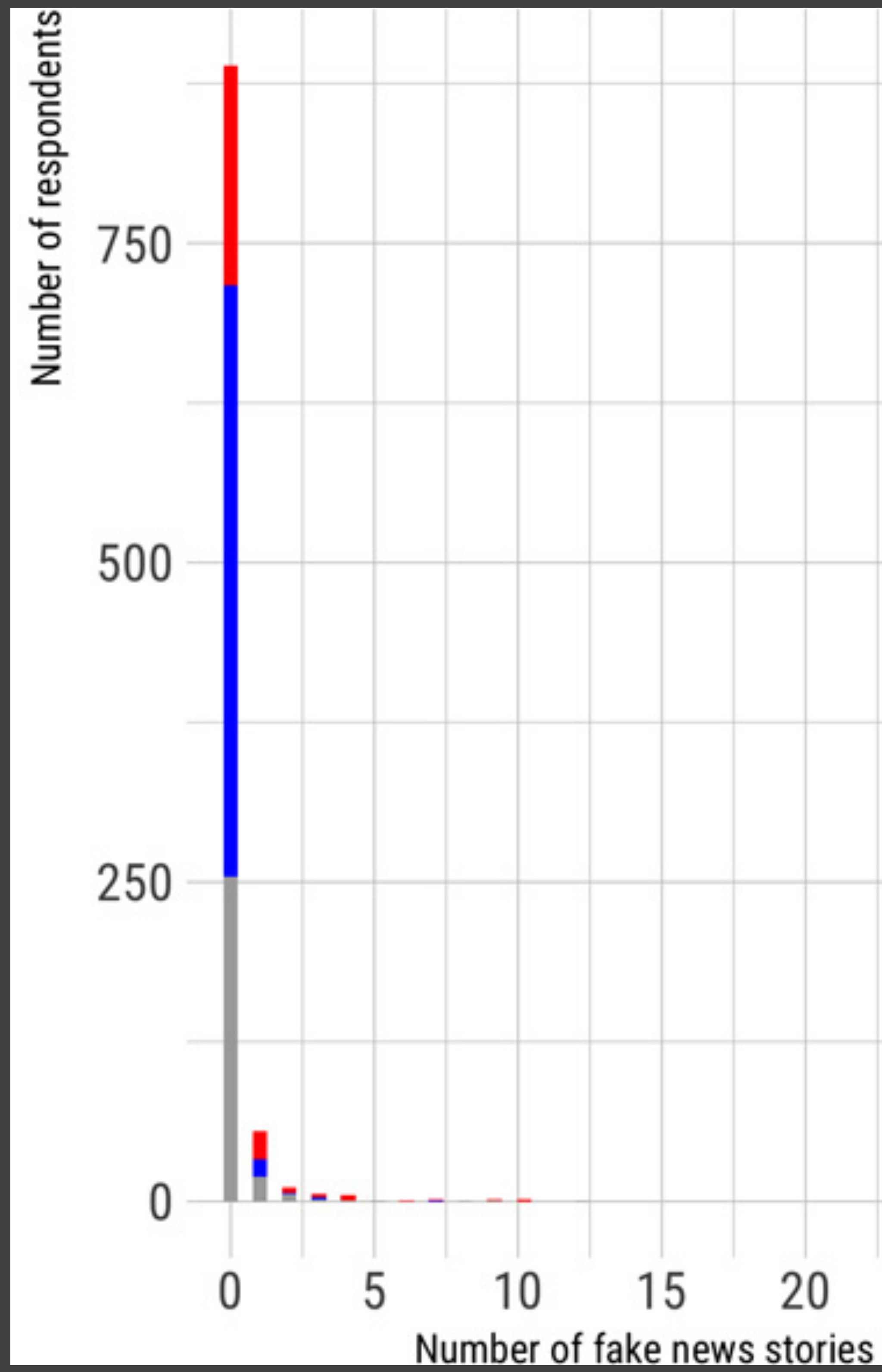
This exposure is typically pro-attitudinal [Guess, Nyhan and Reifler 2020]

How much do we share?

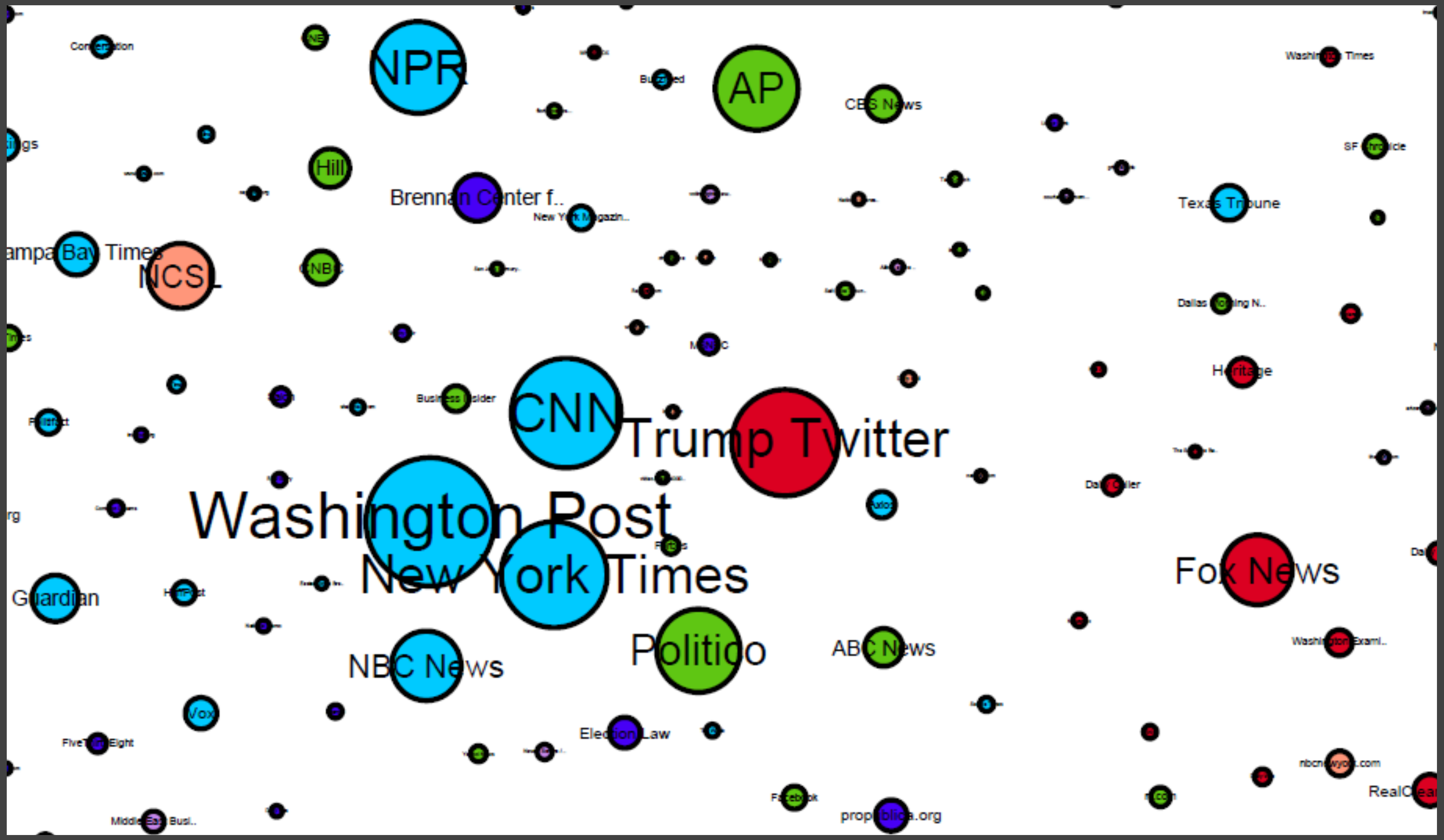
It’s rare: most never share disinformation

In the 2016 election, >65 year olds were 7x more likely than teenagers to share disinformation, and conservatives were more likely to share than liberals [Guess, Nagler, and Tucker 2019]

Misinformation supersharers who accounted for 80% of shares in 2020 were dispropoprtionately Republican, middle-aged white women living in low educated but high income neighborhoods [Barbiri-Bartov, Swire-Thompson, and Grinberg 2024]



Mainstream media amplify the message



Analysis of mail-in voter fraud disinformation suggested that social media played a smaller role than mainstream media in 2020:

1) Trump tweets

2) Mainstream media, trying to be neutral and avoid claims of anti-conservative bias, cover Trump's claims and thereby spread them

[Benkler et al. 2020]

So, from whence misinformation?

Finger #1: "It's trolls."

Actually: "It's motivated actors, who activate unwitting agents."

Finger #2: "Post-truth"

Actually: "People care about being accurate, but generally aren't paying attention to accuracy when they share."

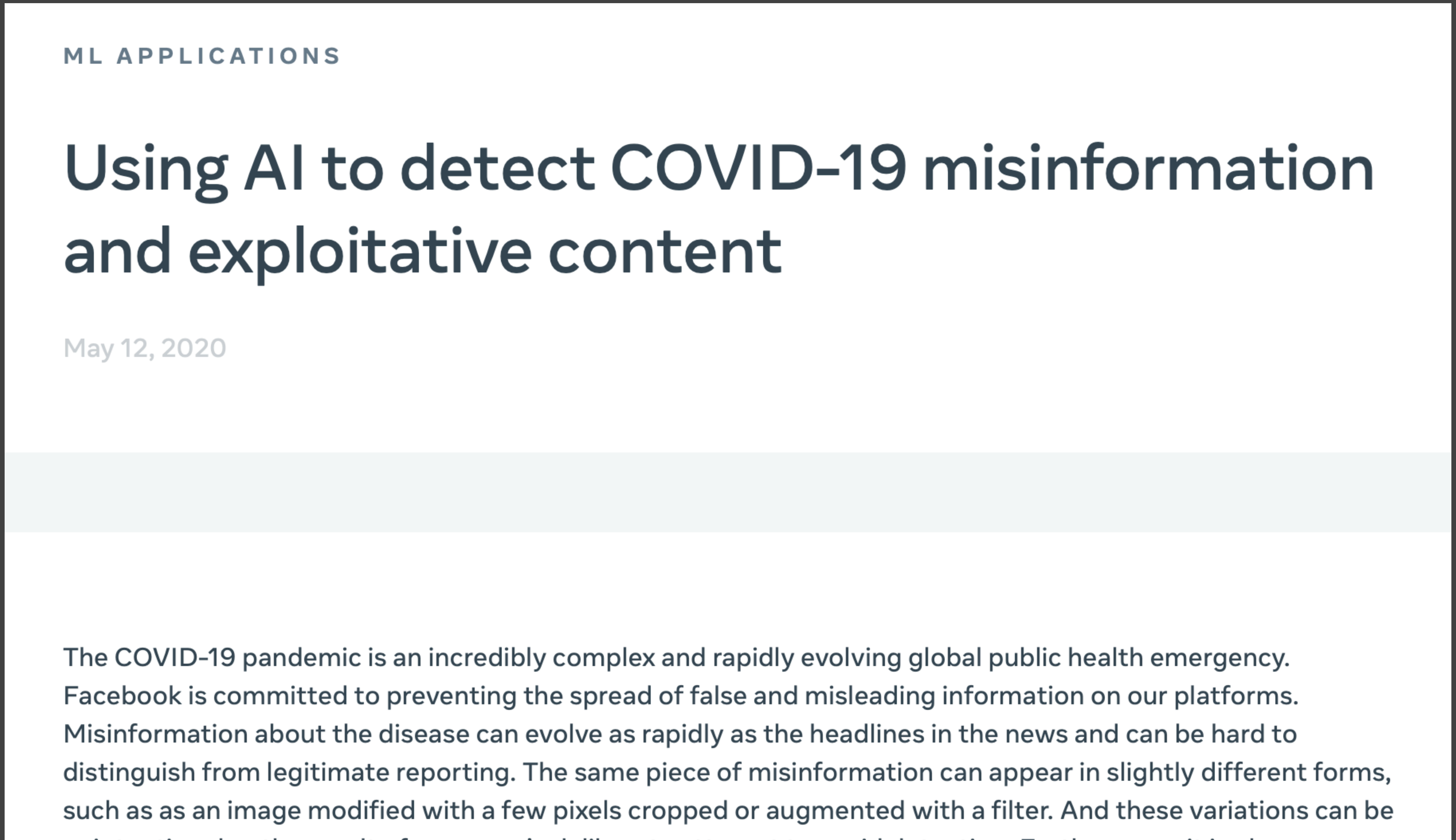
"It's everywhere"

Actually: "Exposure and sharing is rare, but very concentrated."

Classification

Machine learning

Some categories of misinformation (e.g., near copies of flagged articles) can be flagged automatically



ML APPLICATIONS

Using AI to detect COVID-19 misinformation and exploitative content

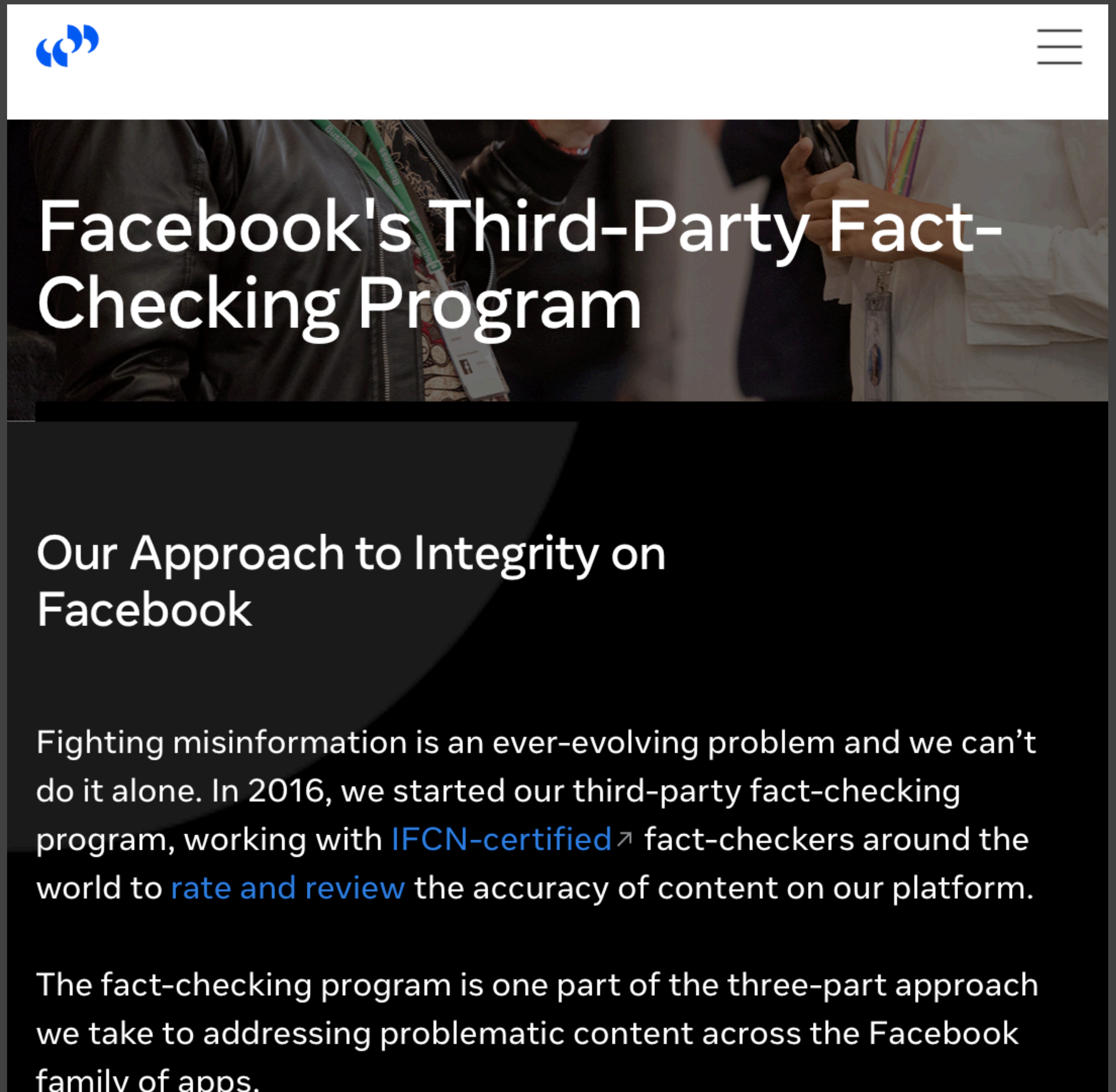
May 12, 2020

The COVID-19 pandemic is an incredibly complex and rapidly evolving global public health emergency. Facebook is committed to preventing the spread of false and misleading information on our platforms. Misinformation about the disease can evolve as rapidly as the headlines in the news and can be hard to distinguish from legitimate reporting. The same piece of misinformation can appear in slightly different forms, such as as an image modified with a few pixels cropped or augmented with a filter. And these variations can be

https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/

Factcheckers

Twitter and Facebook have historically relied on third party fact checkers to decide whether an article is misinformation

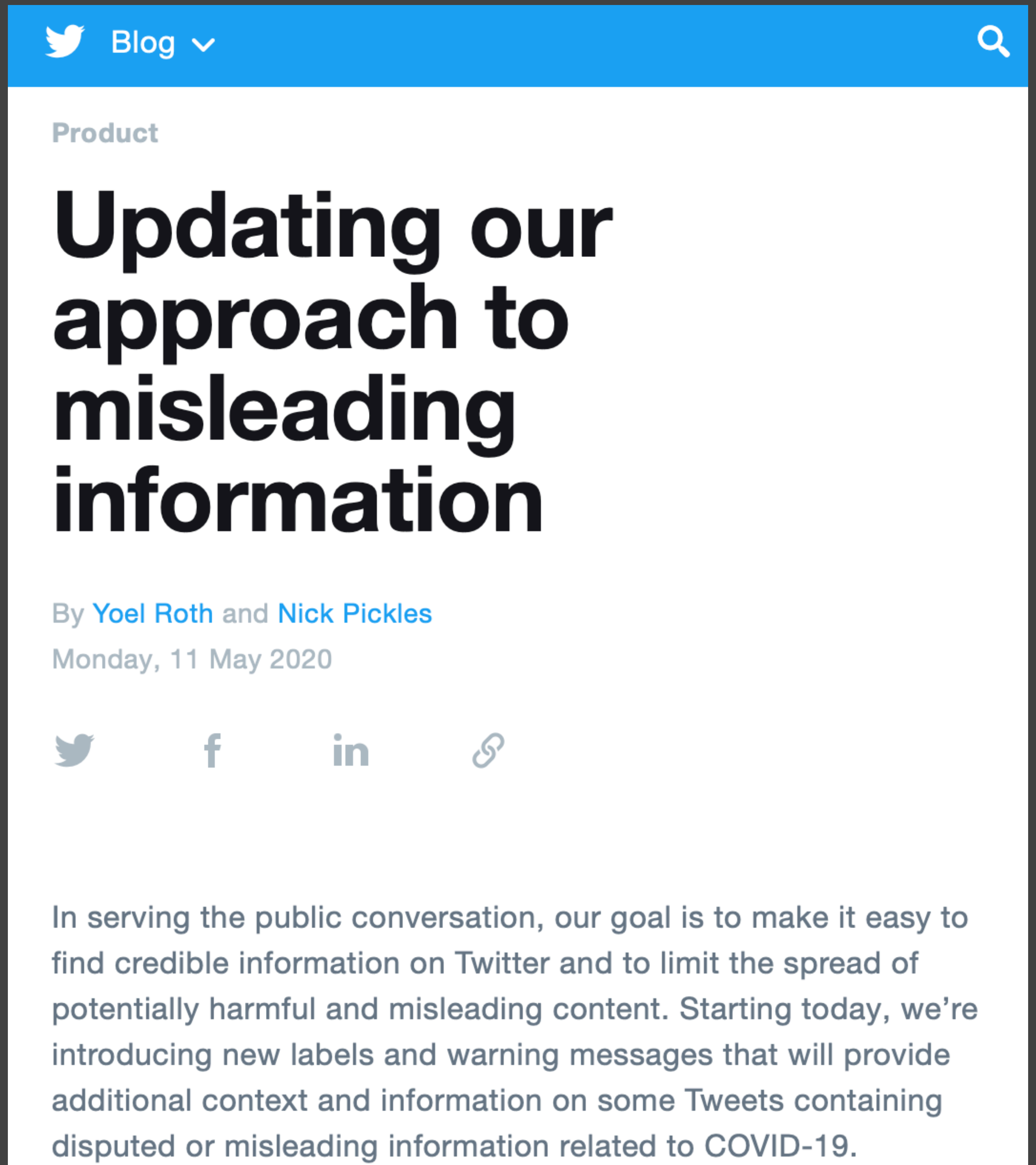


Facebook's Third-Party Fact-Checking Program

Our Approach to Integrity on Facebook

Fighting misinformation is an ever-evolving problem and we can't do it alone. In 2016, we started our third-party fact-checking program, working with IFCN-certified fact-checkers around the world to rate and review the accuracy of content on our platform.

The fact-checking program is one part of the three-part approach we take to addressing problematic content across the Facebook family of apps.



Product

Updating our approach to misleading information

By Yoel Roth and Nick Pickles
Monday, 11 May 2020

In serving the public conversation, our goal is to make it easy to find credible information on Twitter and to limit the spread of potentially harmful and misleading content. Starting today, we're introducing new labels and warning messages that will provide additional context and information on some Tweets containing disputed or misleading information related to COVID-19.

Factcheckers

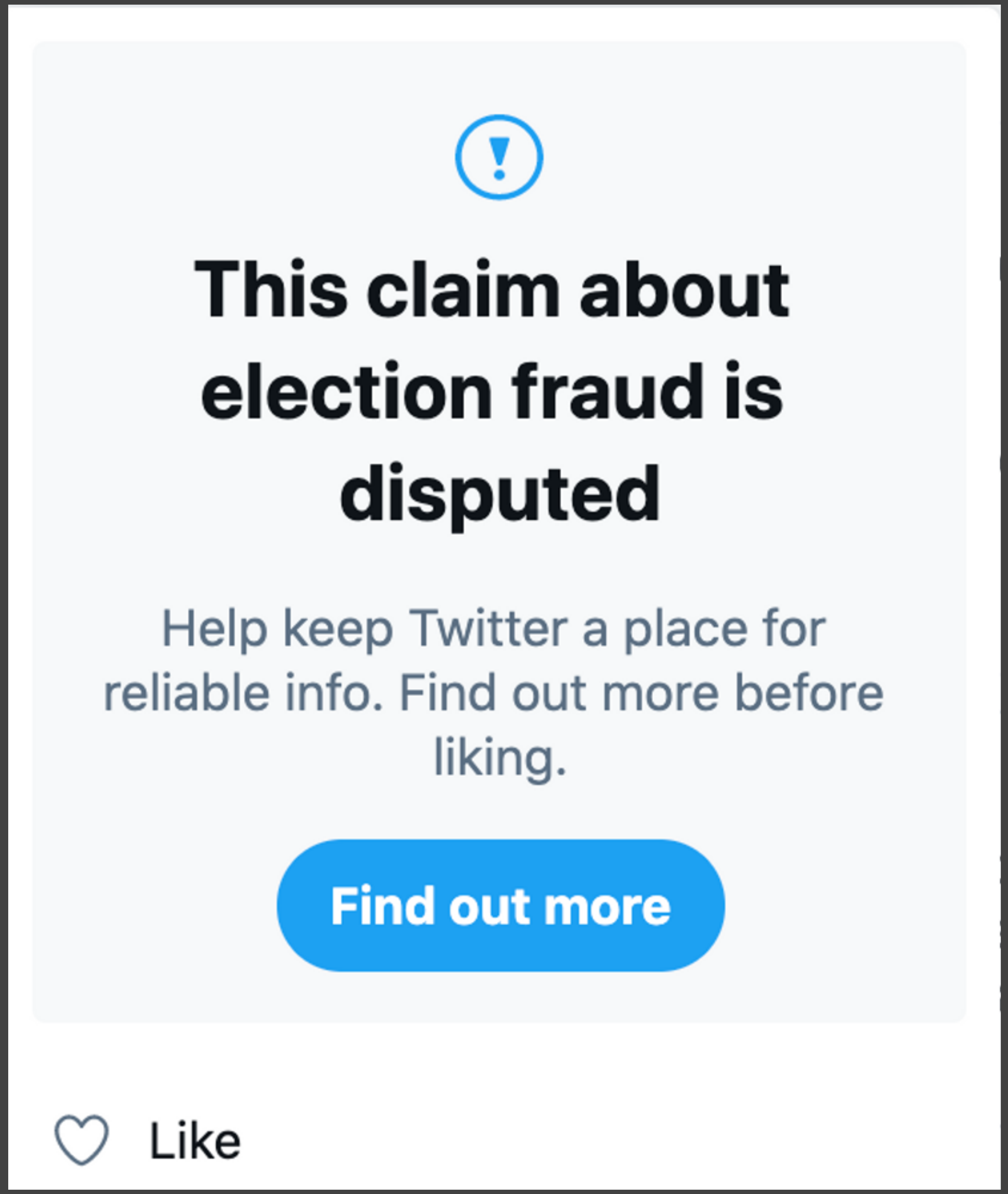
Twitter and Facebook have historically relied on third party fact checkers to decide whether an article is misinformation

Factchecking works! Fact checker labels reduce belief in the article even for those who distrust fact checkers [Martel and Rand 2024]

However, this does not cover the long tail: Facebook's partners comprise 26 fact checkers who collectively review 200 articles per month [Rodrigo 2020]

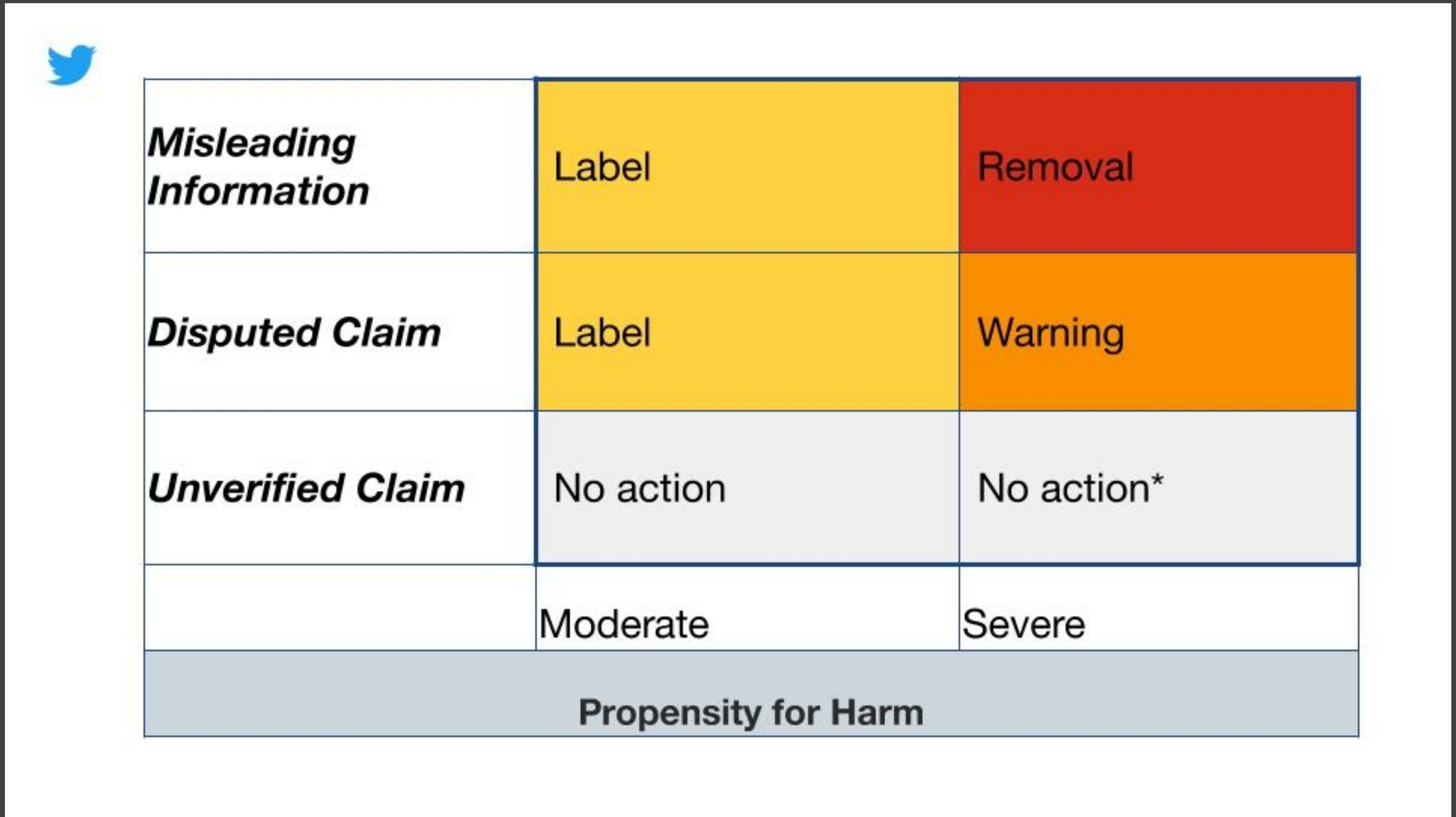
Fact checkers can also take days to do the research, by which time the article or video has spread widely

Twitter's pre-Elon criteria



This claim about election fraud is disputed

Help keep Twitter a place for reliable info. Find out more before liking.



Misleading Information	Label	Removal
Disputed Claim	Label	Warning
Unverified Claim	No action	No action*
	Moderate	Severe
	Propensity for Harm	

See [Atreja, Hemphill, and Resnick 2023] for more

Disinformation campaigns

[Starbird, Arif, and Wilson 2019; Allen, Watts, and Rand 2024]

Instead of classifying individual pieces of content, we can study and classify disinformation campaigns — a collection of information actions

1) Is this campaign pushing a false narrative? Then, classify:
2) Is this article a part of this disinformation campaign?

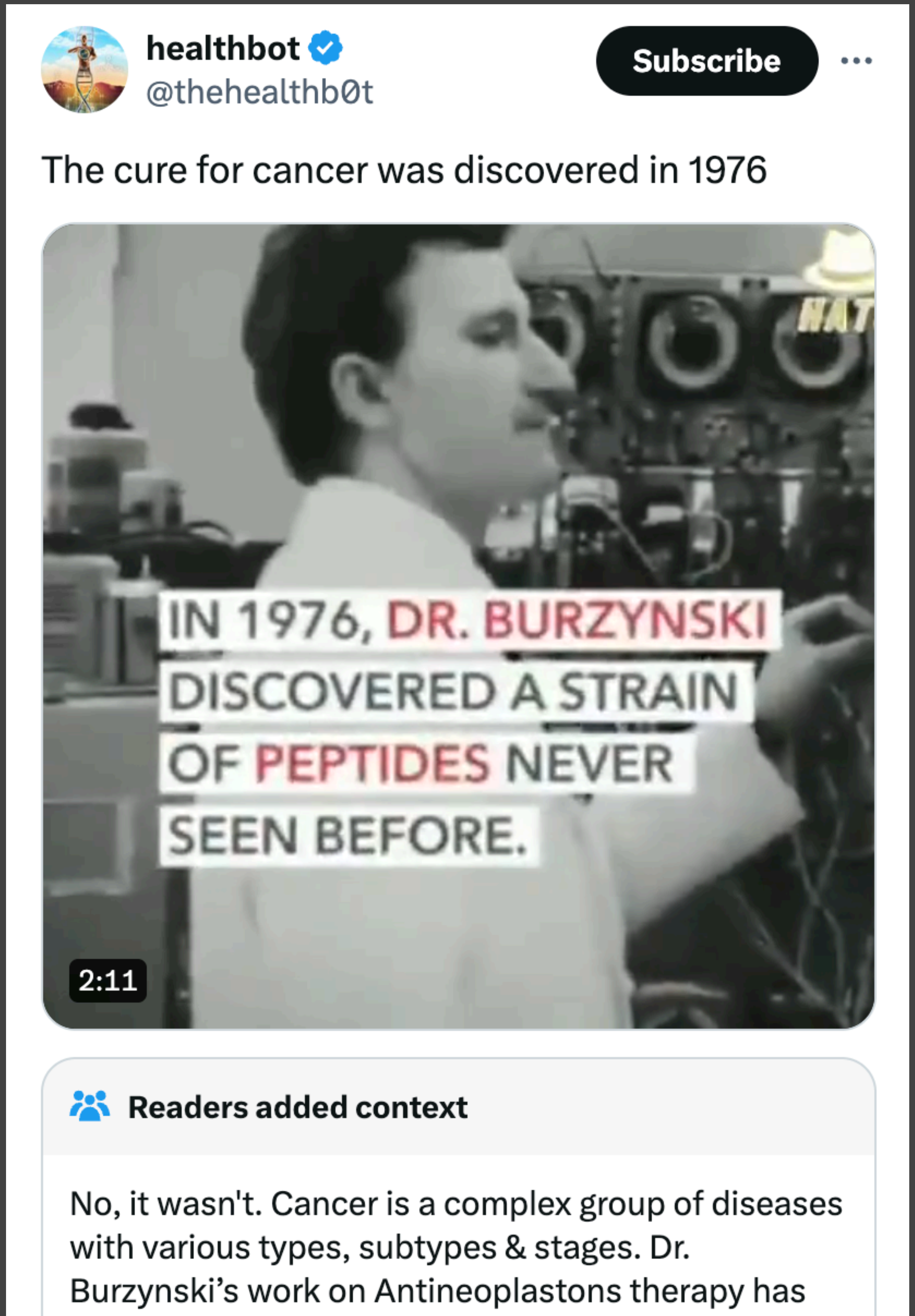
Otherwise, much is missed: vaccine-skeptical posts had an effect 46x greater than the actual misinformation that the platforms flagged. ("Unflagged stories highlighting rare deaths after vaccination were among Facebook's most-viewed stories")

Community Notes

Representative samples of "crowd jurors" can be as accurate as fact checkers and much faster [Allen et al. 2021]

Result via a difference-in-difference study: community notes reduce retweets by half [Renault, Amariles, and Troussel 2024]

But, they are still too slow: tweets get half of their total views within 80 minutes, but community notes don't appear on average until after two days [Chuai 2024]



healthbot
@thehealthb0t

Subscribe

The cure for cancer was discovered in 1976

IN 1976, DR. BURZYNSKI DISCOVERED A STRAIN OF PEPTIDES NEVER SEEN BEFORE.

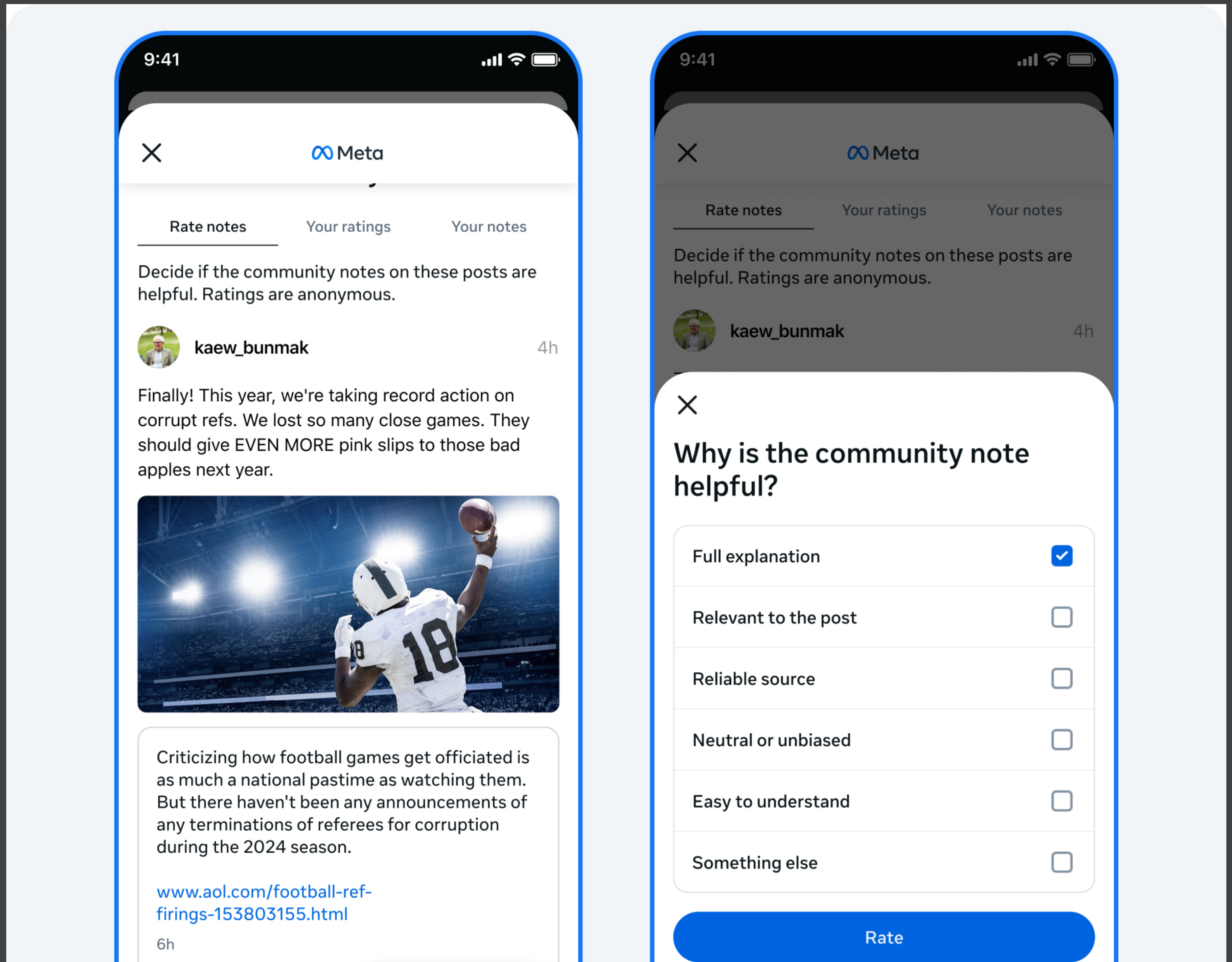
2:11

Readers added context

No, it wasn't. Cancer is a complex group of diseases with various types, subtypes & stages. Dr. Burzynski's work on Antineoplastons therapy has

Community Notes

Post-Trump's reelection, Meta is now pushing crowdsourced post annotation as "less biased than the third party fact checking program" (their quote, not mine)



Decide if the community notes on these posts are helpful. Ratings are anonymous.

kaew_bunmak 4h

Finally! This year, we're taking record action on corrupt refs. We lost so many close games. They should give EVEN MORE pink slips to those bad apples next year.

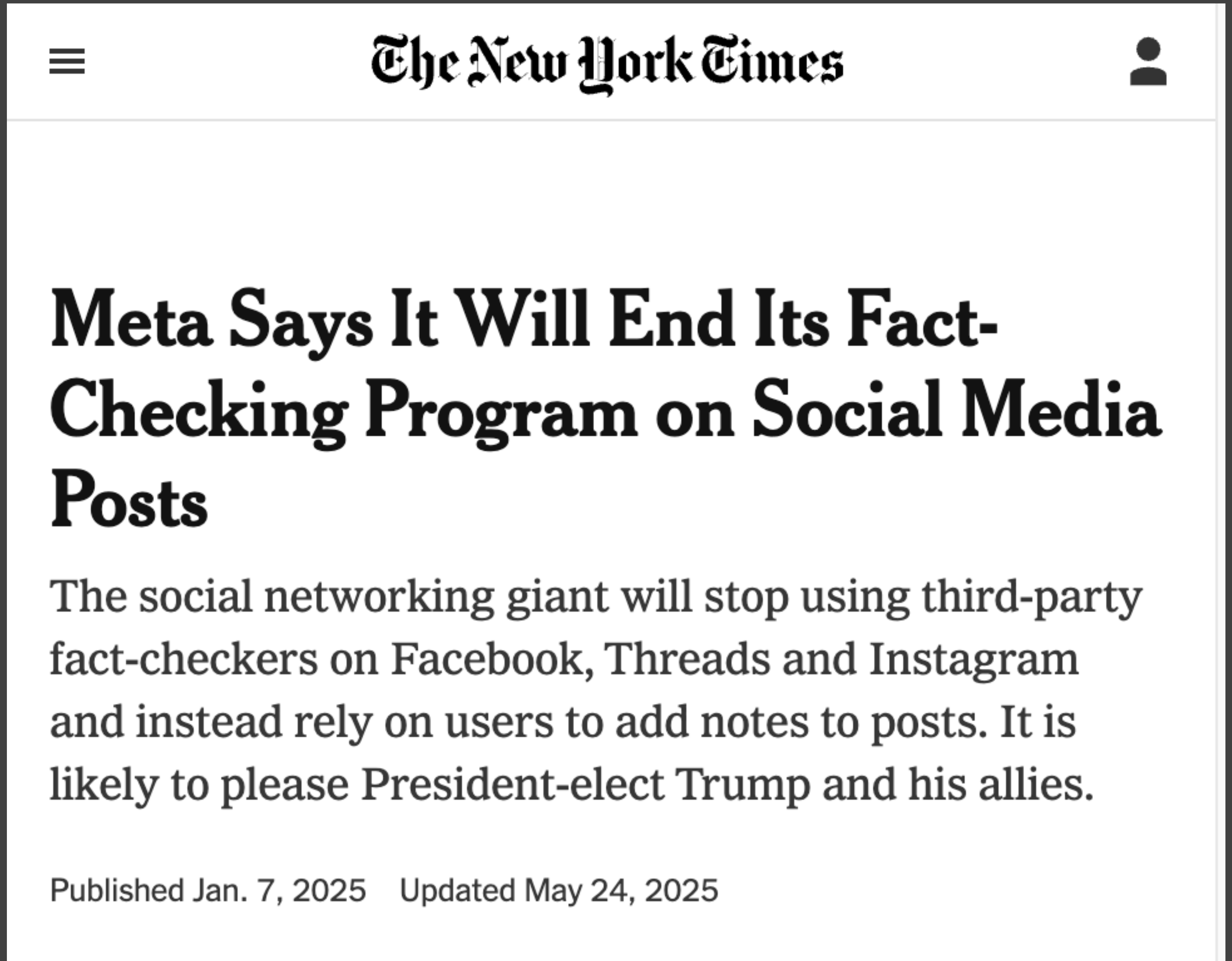
Criticizing how football games get officiated is as much a national pastime as watching them. But there haven't been any announcements of any terminations of referees for corruption during the 2024 season.

www.aol.com/football-ref-firings-153803155.html

6h

Why is the community note helpful?

- Full explanation ☑
- Relevant to the post □
- Reliable source □
- Neutral or unbiased □
- Easy to understand □
- Something else □



The New York Times

Meta Says It Will End Its Fact-Checking Program on Social Media Posts

The social networking giant will stop using third-party fact-checkers on Facebook, Threads and Instagram and instead rely on users to add notes to posts. It is likely to please President-elect Trump and his allies.

Published Jan. 7, 2025 Updated May 24, 2025

Interventions

Reduce feed ranking

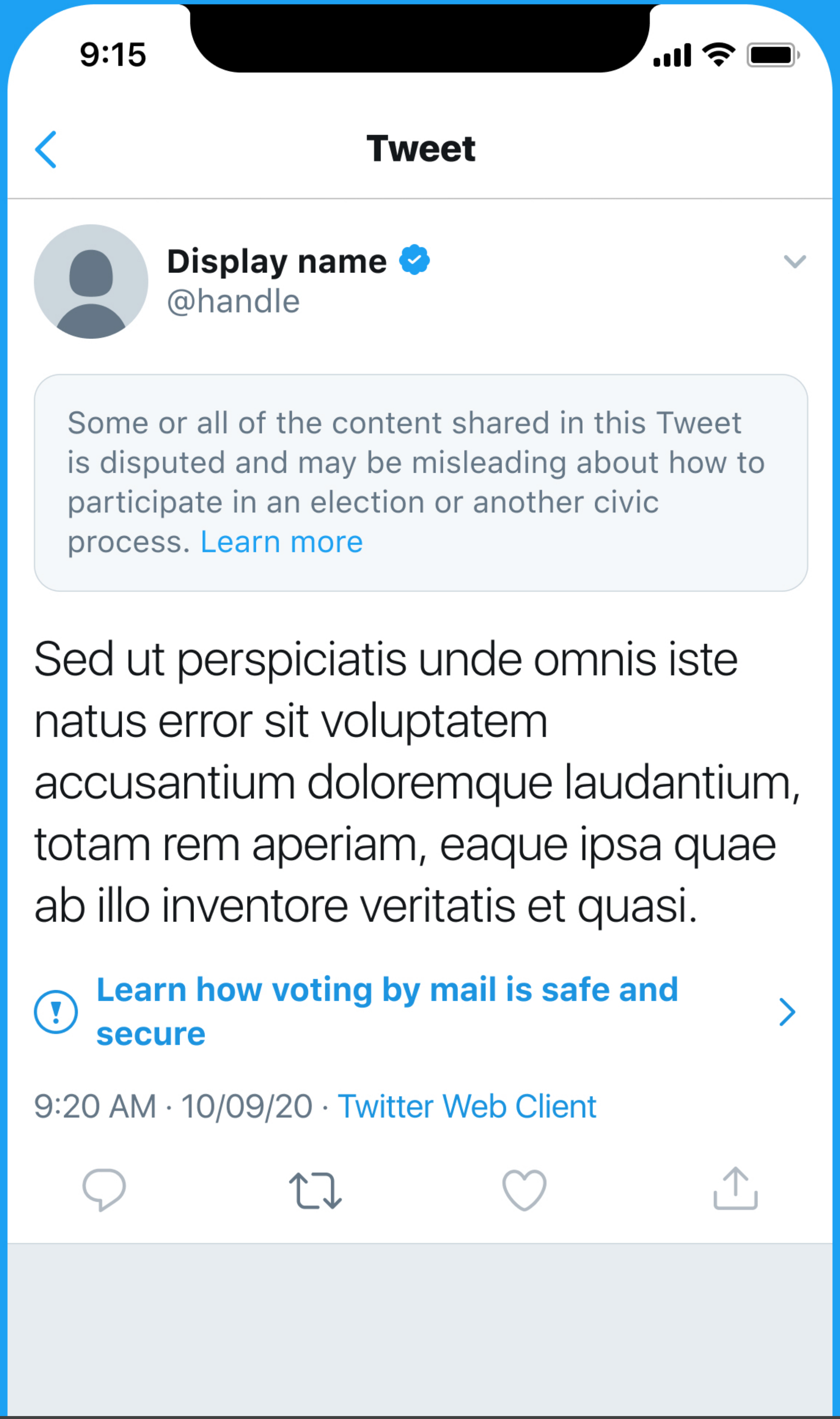
Platforms can (temporarily) reduce the feed ranking of links that might be disinformation, slowing their spread while fact checkers review it

Ex: Article is lower in your Instagram feed, video is recommended less often on YouTube

Pros: walks a line between removal and unconstrained spread

Cons: opaque, unclear when it's happening, likely too late once other media start reporting on it

An intervention gallery



9:15

Tweet

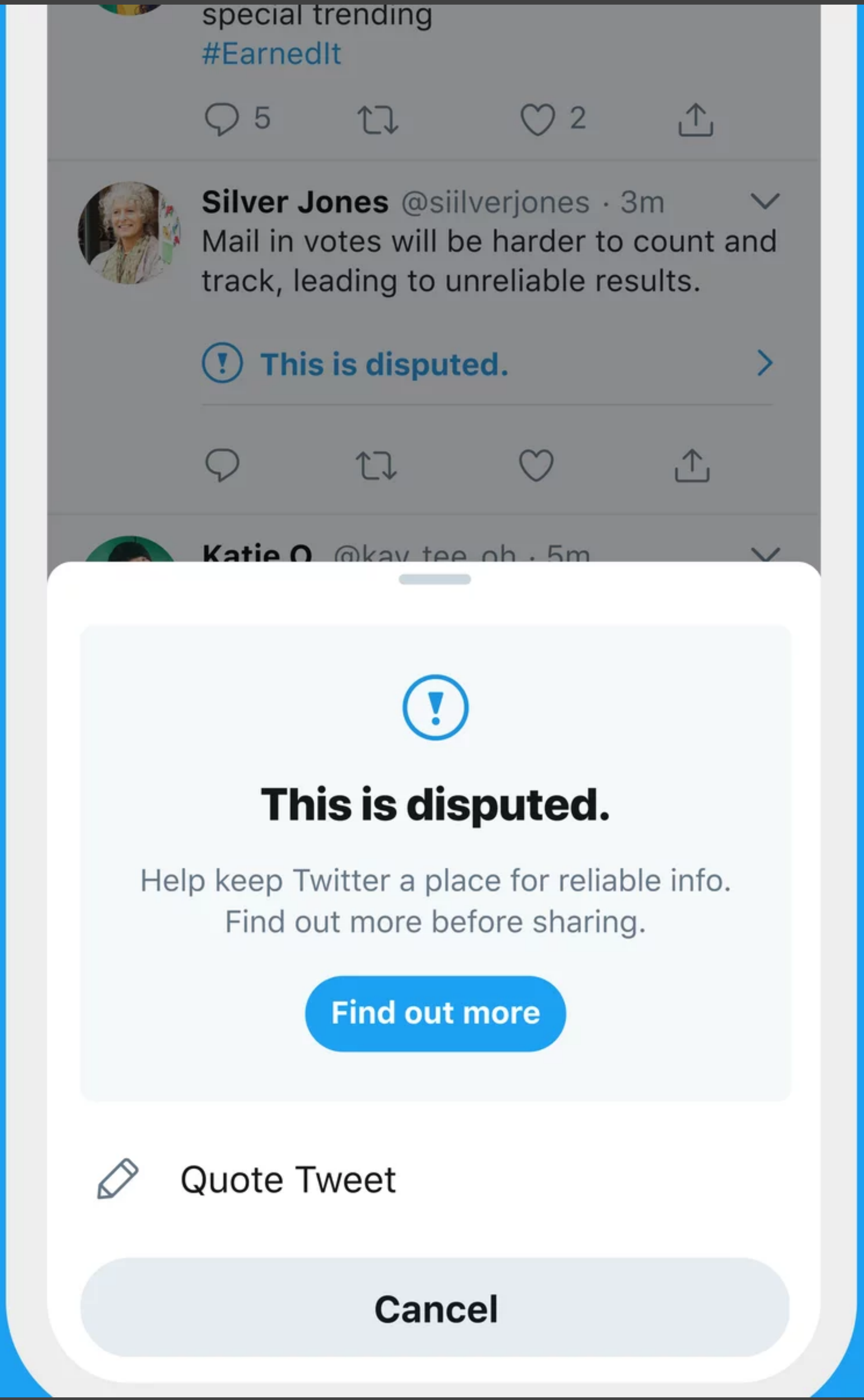
Display name
@handle

Some or all of the content shared in this Tweet is disputed and may be misleading about how to participate in an election or another civic process. Learn more

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi.

Learn how voting by mail is safe and secure

9:20 AM · 10/09/20 · Twitter Web Client



special trending
#EarnedIt

5 2

Silver Jones @siilverjones · 3m
Mail in votes will be harder to count and track, leading to unreliable results.

This is disputed.

Katie O @kay_tee_oh · 5m

This is disputed.

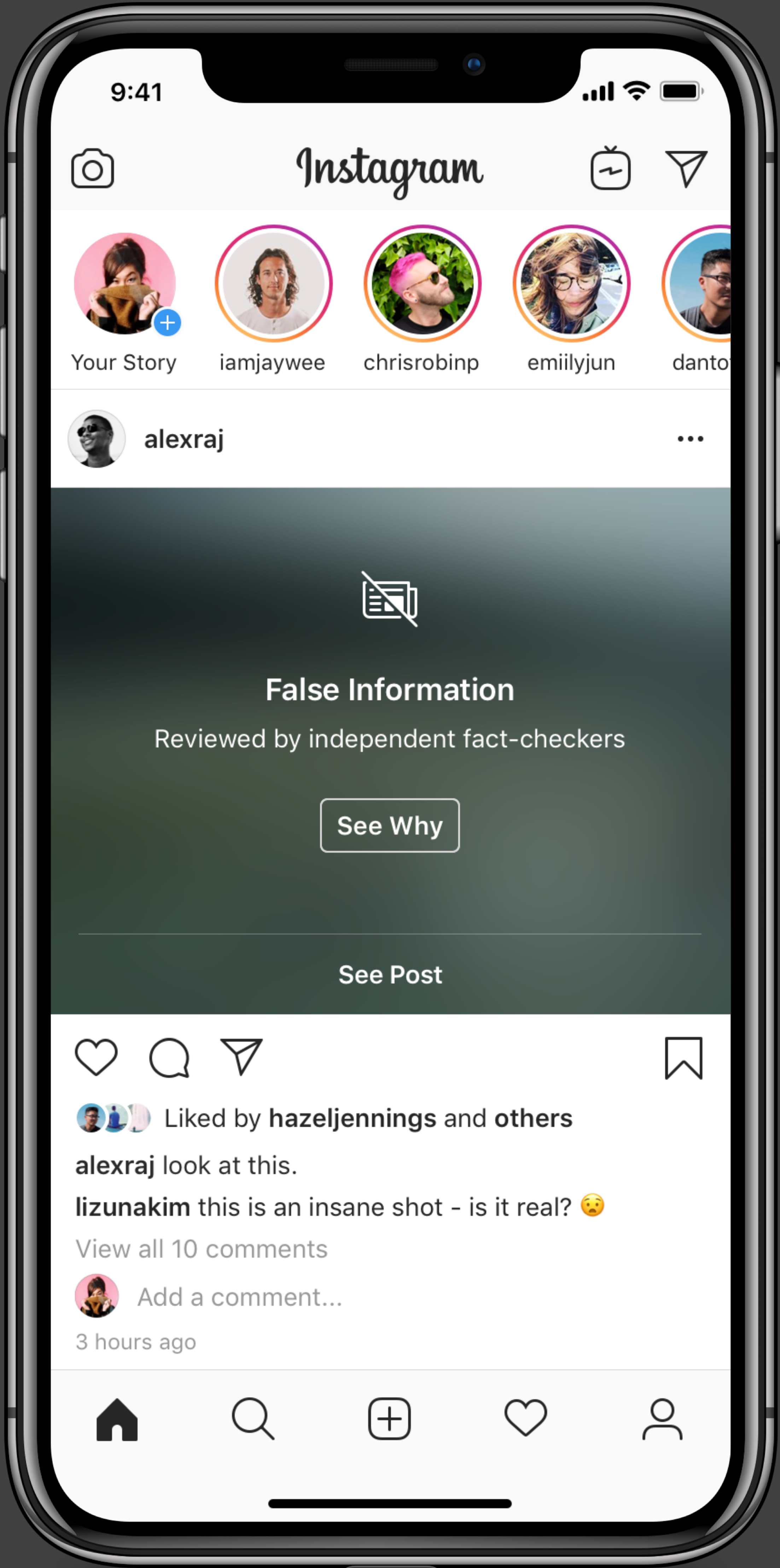
Help keep Twitter a place for reliable info. Find out more before sharing.

Find out more

Quote Tweet

Cancel

An intervention gallery



False Information

Reviewed by independent fact-checkers

Liked by **hazeljennings** and **others**

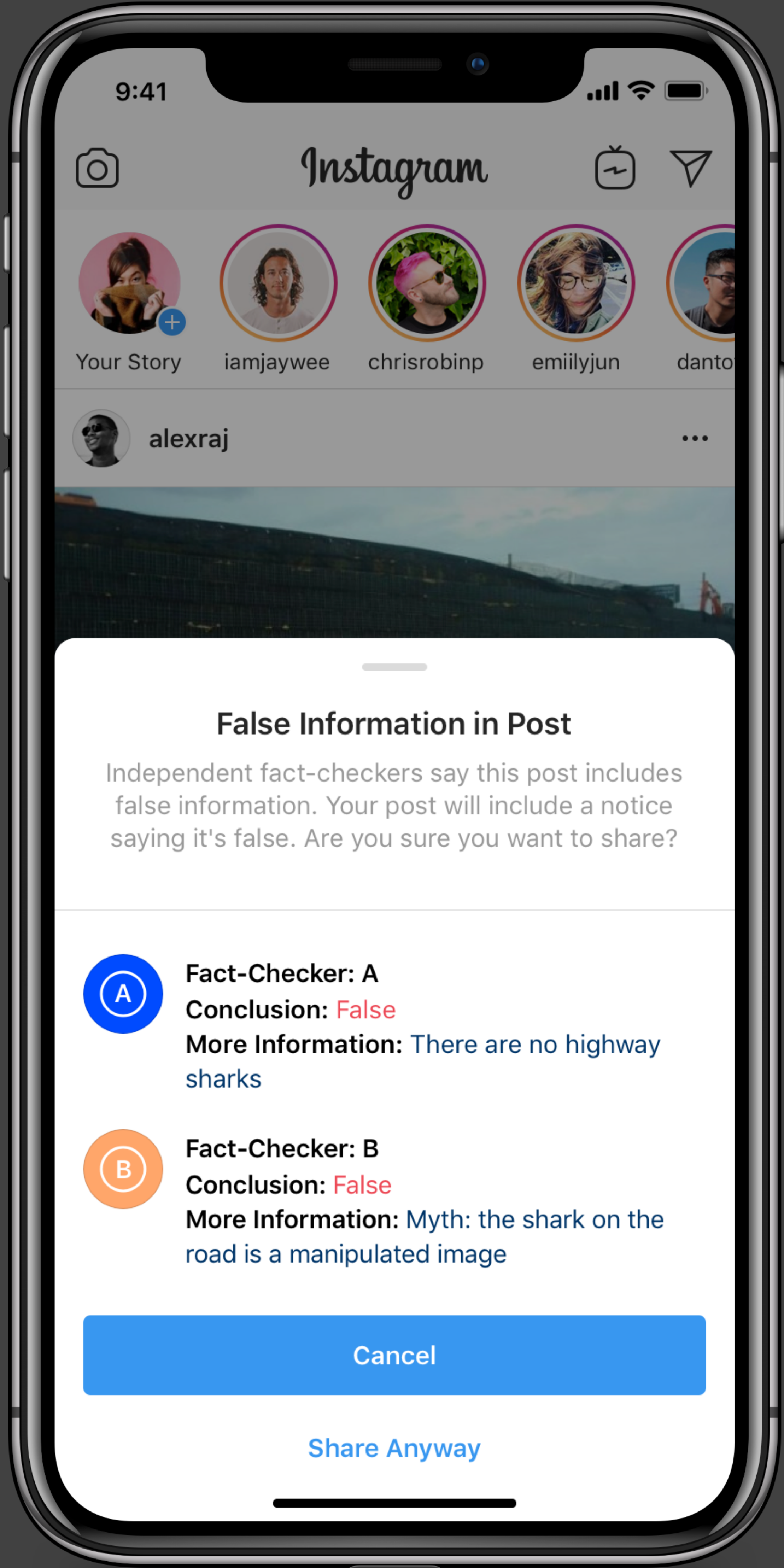
alexraj look at this.

lizunakim this is an insane shot - is it real? 😮

View all 10 comments

Add a comment...

3 hours ago

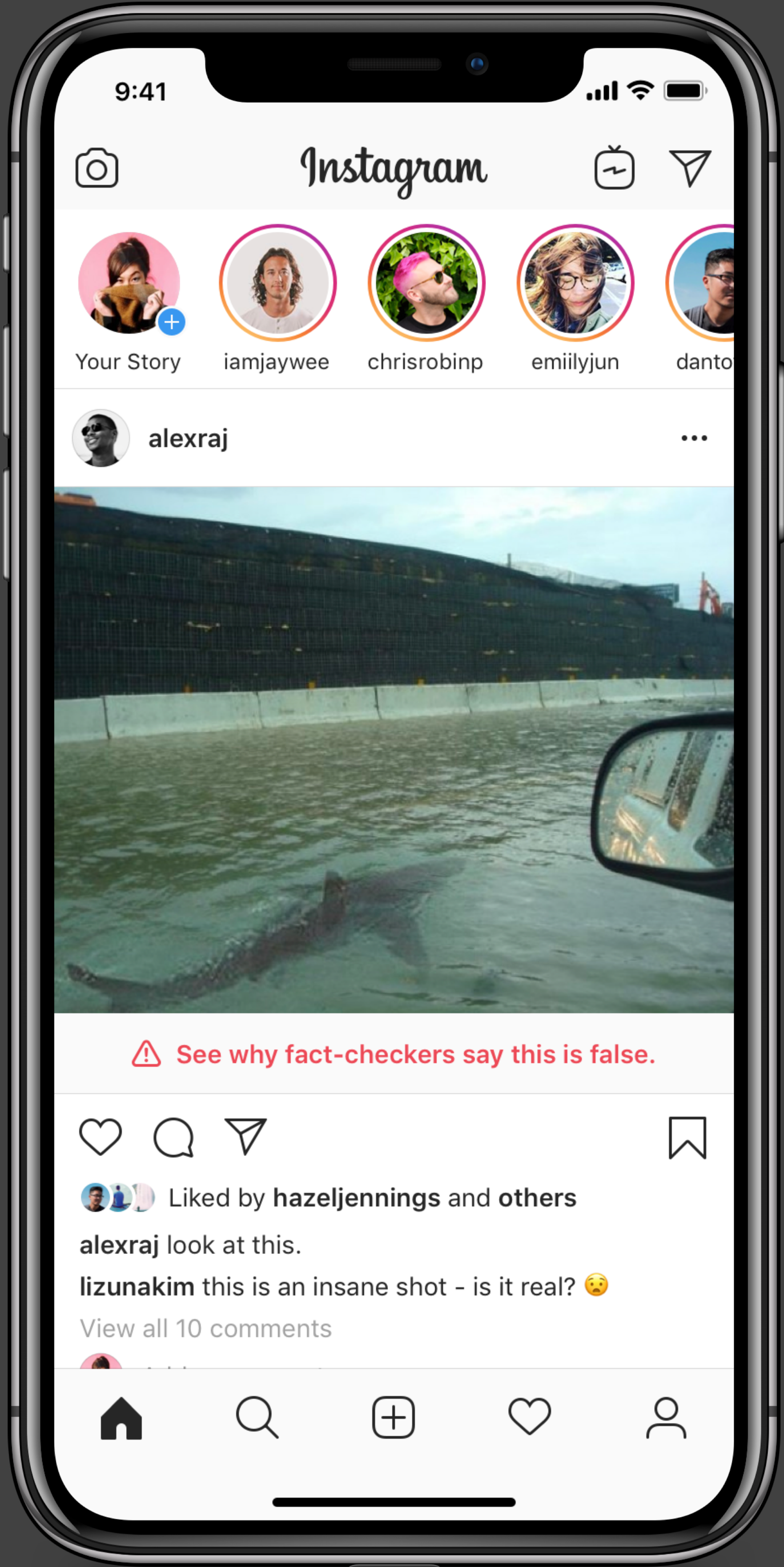


False Information in Post

Independent fact-checkers say this post includes false information. Your post will include a notice saying it's false. Are you sure you want to share?

Fact-Checker: A
Conclusion: False
More Information: There are no highway sharks

Fact-Checker: B
Conclusion: False
More Information: Myth: the shark on the road is a manipulated image



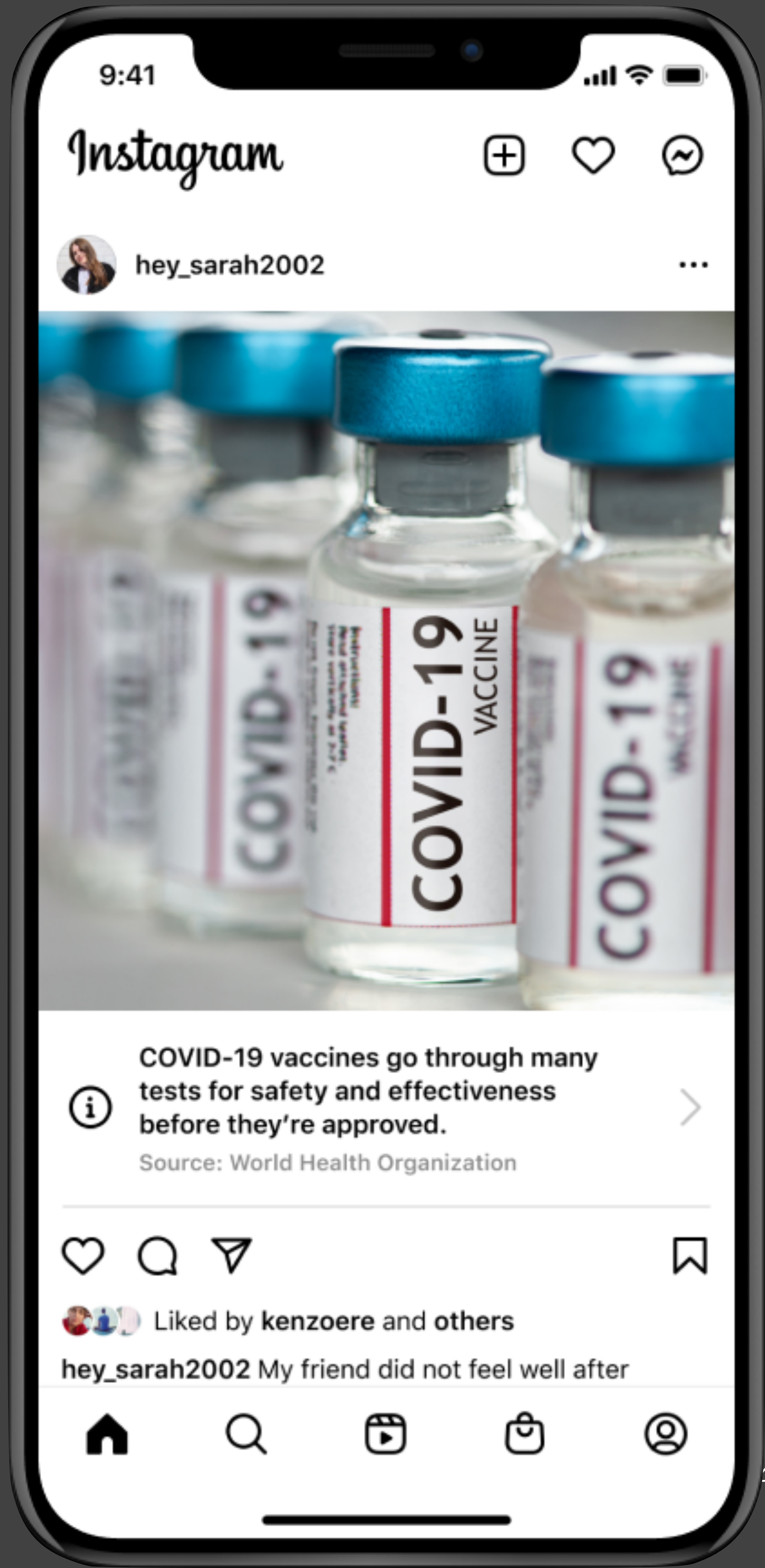
⚠ See why fact-checkers say this is false.

Liked by **hazeljennings** and **others**

alexraj look at this.

lizunakim this is an insane shot - is it real? 😮

View all 10 comments



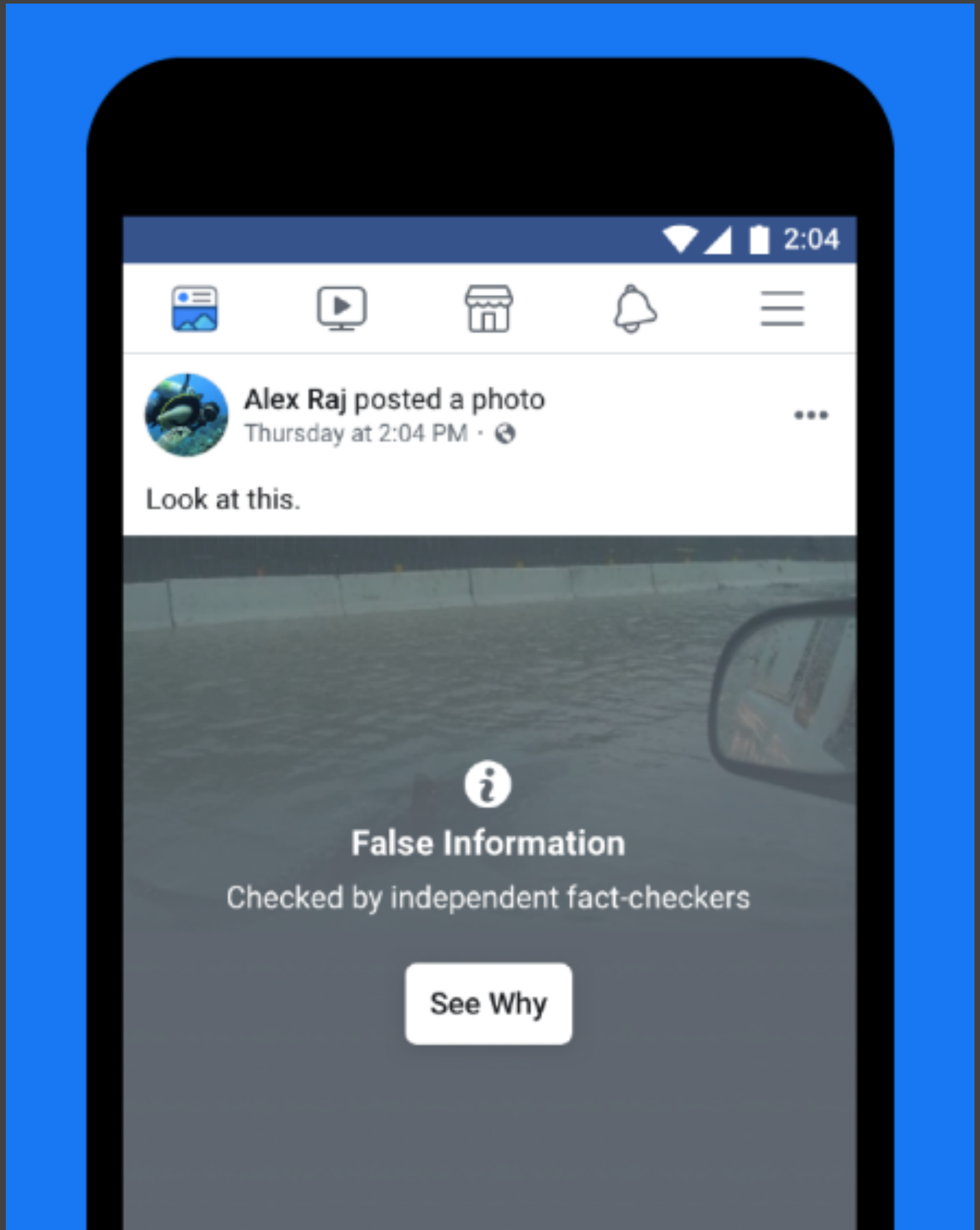
COVID-19 vaccines go through many tests for safety and effectiveness before they're approved.

Source: World Health Organization

Liked by **kenzoere** and **others**

hey_sarah2002 My friend did not feel well after

An intervention gallery



2:04

Alex Raj posted a photo
Thursday at 2:04 PM ·

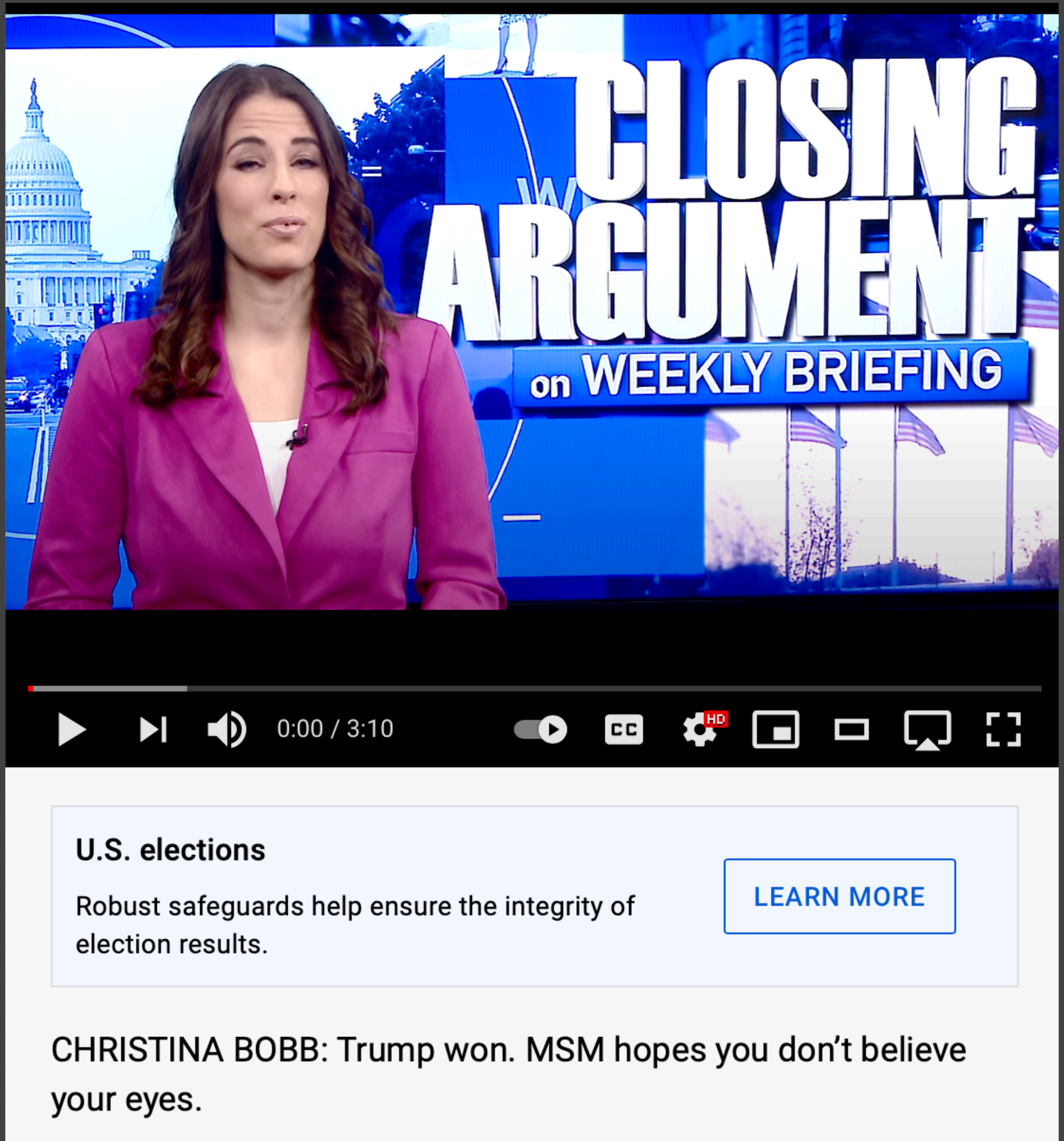
Look at this.

False Information

Checked by independent fact-checkers

See Why

An intervention gallery



U.S. elections

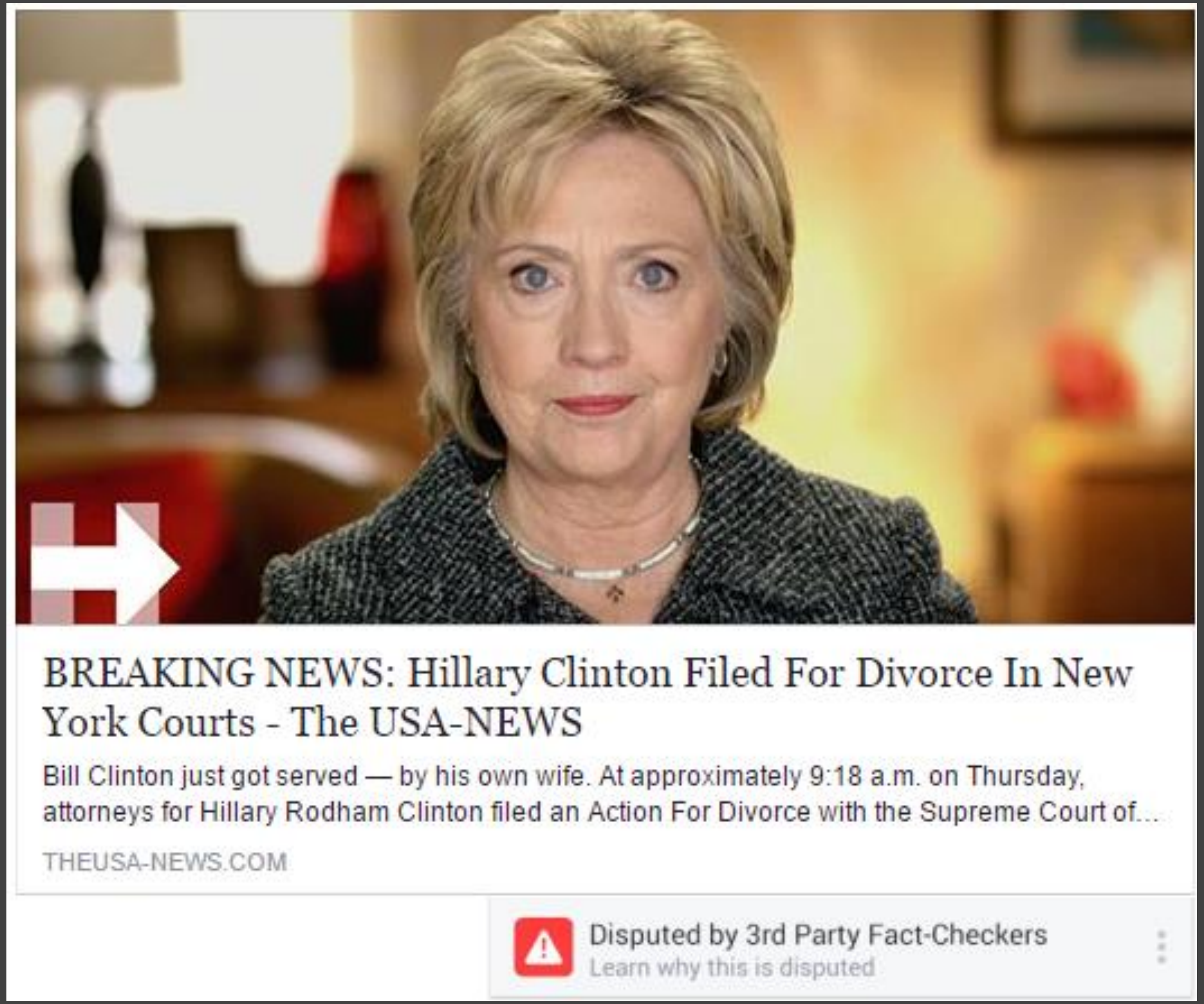
Robust safeguards help ensure the integrity of election results.

CHRISTINA BOBB: Trump won. MSM hopes you don't believe your eyes.

Implied truth effect

Labeling some stories as false leads people to believe that everything not explicitly labeled as false…is true. [Pennycook et al. 2020]

This is problematic when fact checkers can only check a tiny percentage of all content on the site.



BREAKING NEWS: Hillary Clinton Filed For Divorce In New York Courts - The USA-NEWS

Bill Clinton just got served — by his own wife. At approximately 9:18 a.m. on Thursday, attorneys for Hillary Rodham Clinton filed an Action For Divorce with the Supreme Court of...

THEUSA-NEWS.COM

Disputed by 3rd Party Fact-Checkers

Learn why this is disputed

Priming accuracy

Bringing attention to the accuracy of information shared on FB+Twitter moderately improves the quality of news shared later [Pennycook et al. 2021; Lin, Garro, et al. 2024]

Why? Recall: we're not in a post-truth world, where people don't care about accuracy. We instead tend to be more focused on other motivators, like pleasing our followers.

Since they rely on redirecting attention, these are not long-lasting effects (yet).

TWO TIPS FOR THINKING CRITICALLY BEFORE SHARING AN ARTICLE.

Look closely at the web address. Read carefully.

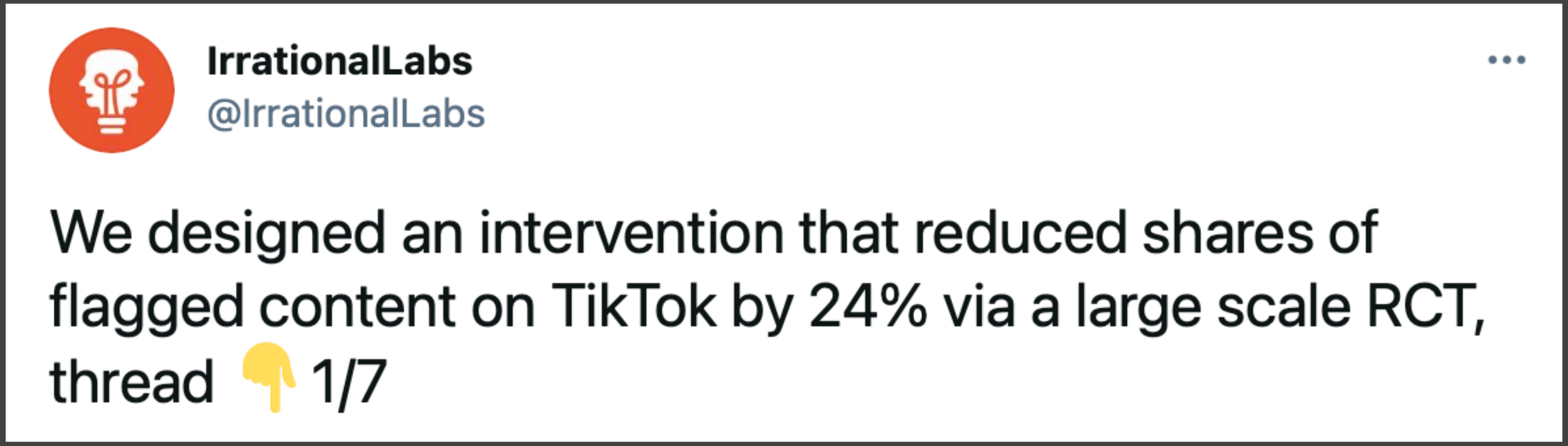
Meta

Think about

ACCURACY

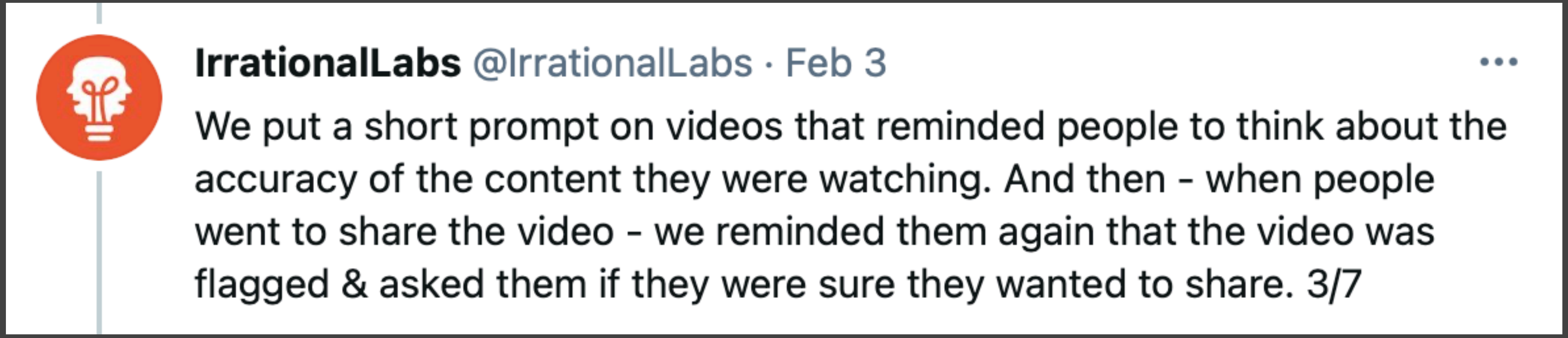
Before you share

Priming accuracy in practice



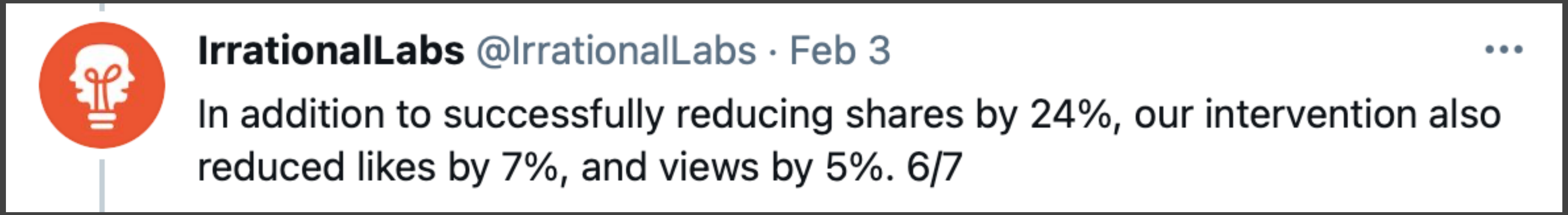
IrrationalLabs
@IrrationalLabs

We designed an intervention that reduced shares of flagged content on TikTok by 24% via a large scale RCT, thread 👇 1/7



IrrationalLabs @IrrationalLabs · Feb 3

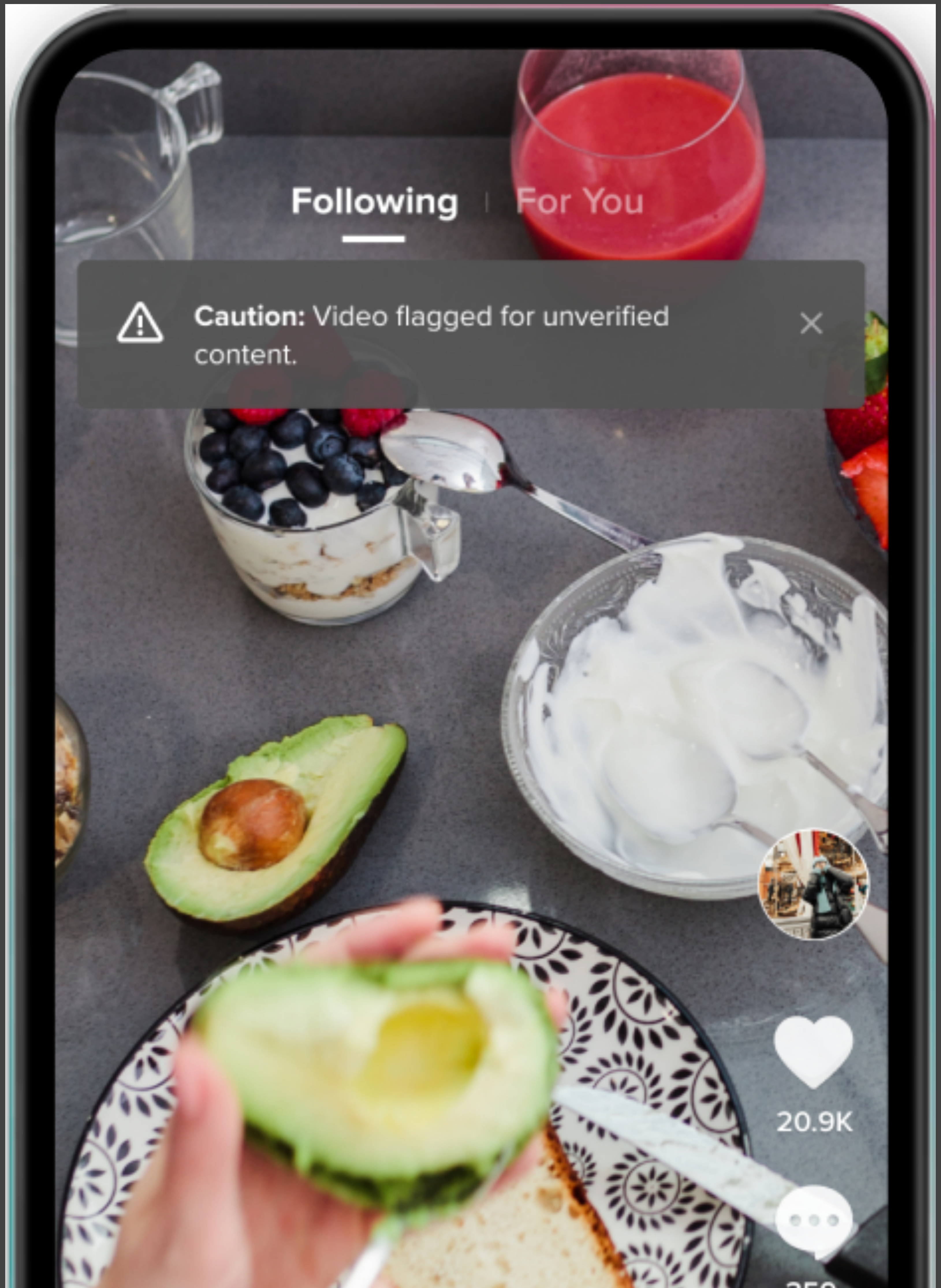
We put a short prompt on videos that reminded people to think about the accuracy of the content they were watching. And then - when people went to share the video - we reminded them again that the video was flagged & asked them if they were sure they wanted to share. 3/7



IrrationalLabs @IrrationalLabs · Feb 3

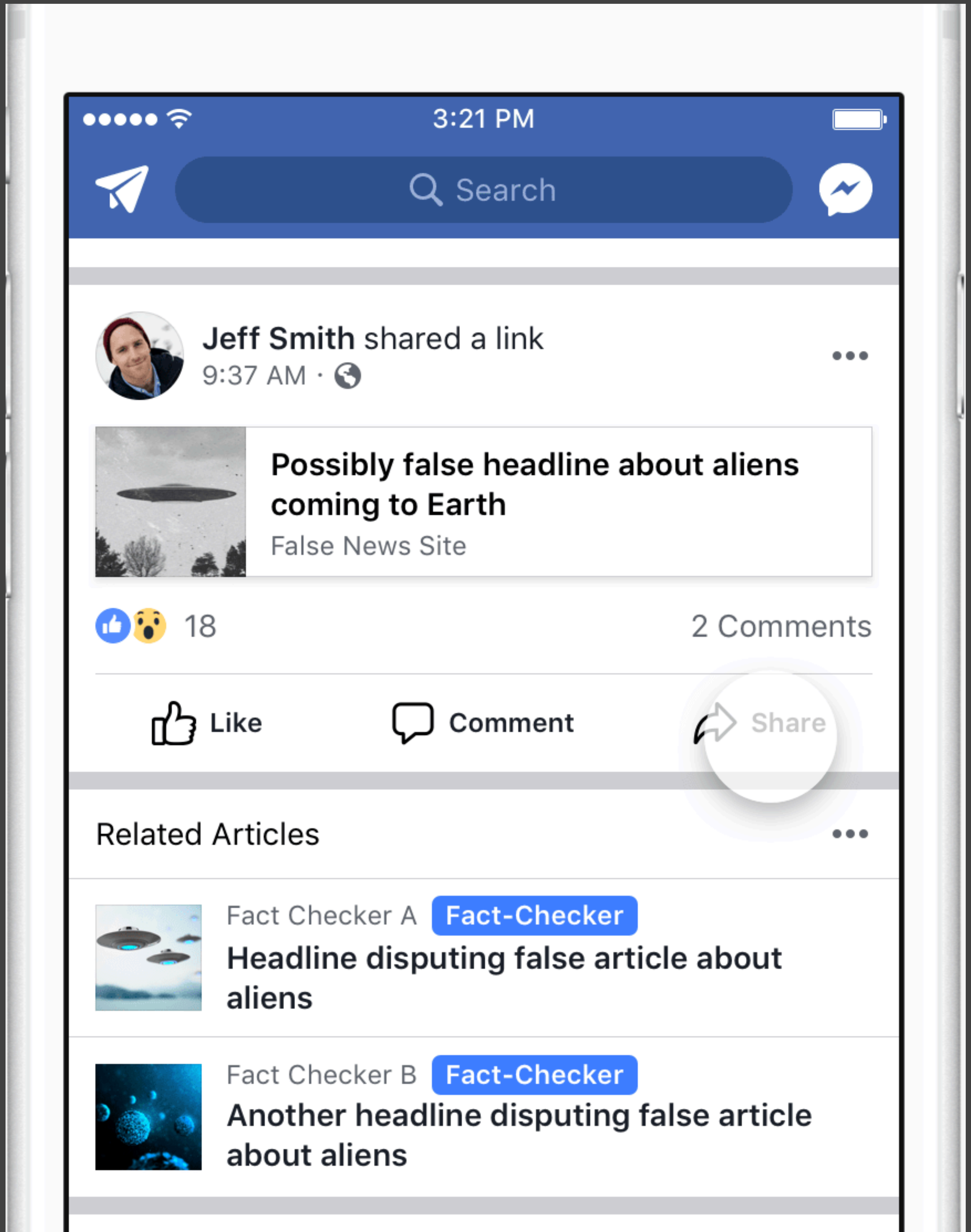
In addition to successfully reducing shares by 24%, our intervention also reduced likes by 7%, and views by 5%. 6/7

[https://twitter.com/IrrationalLabs/status/1357033901311451140]



⚠ **Caution:** Video flagged for unverified content.

Back to our question: which design will better reduce the spread of disinformation?



3:21 PM

Search

Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

18 2 Comments

Like Comment Share

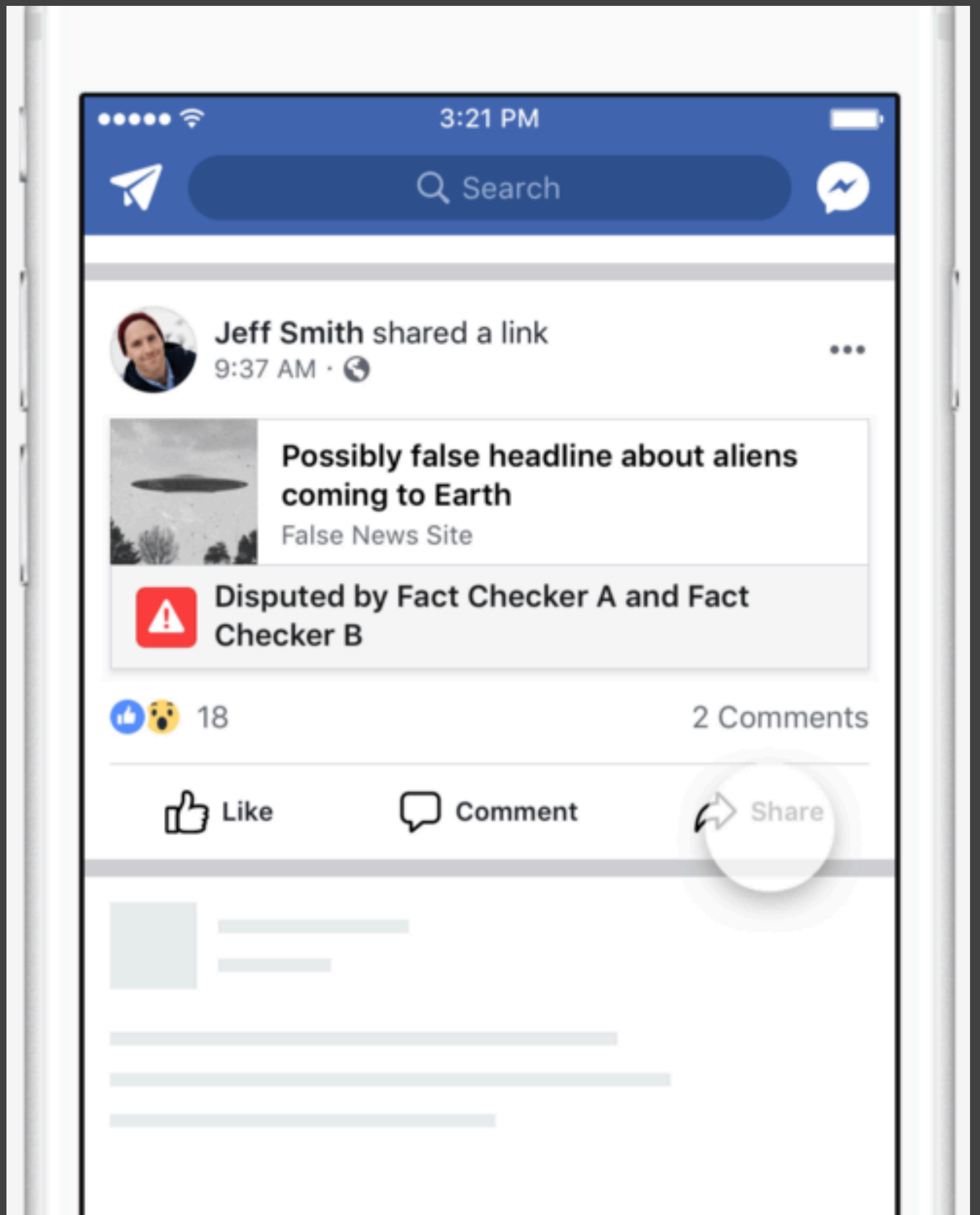
Related Articles

Fact Checker A Fact-Checker
Headline disputing false article about aliens

Fact Checker B Fact-Checker
Another headline disputing false article about aliens

related articles

fact check



3:21 PM

Search

Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

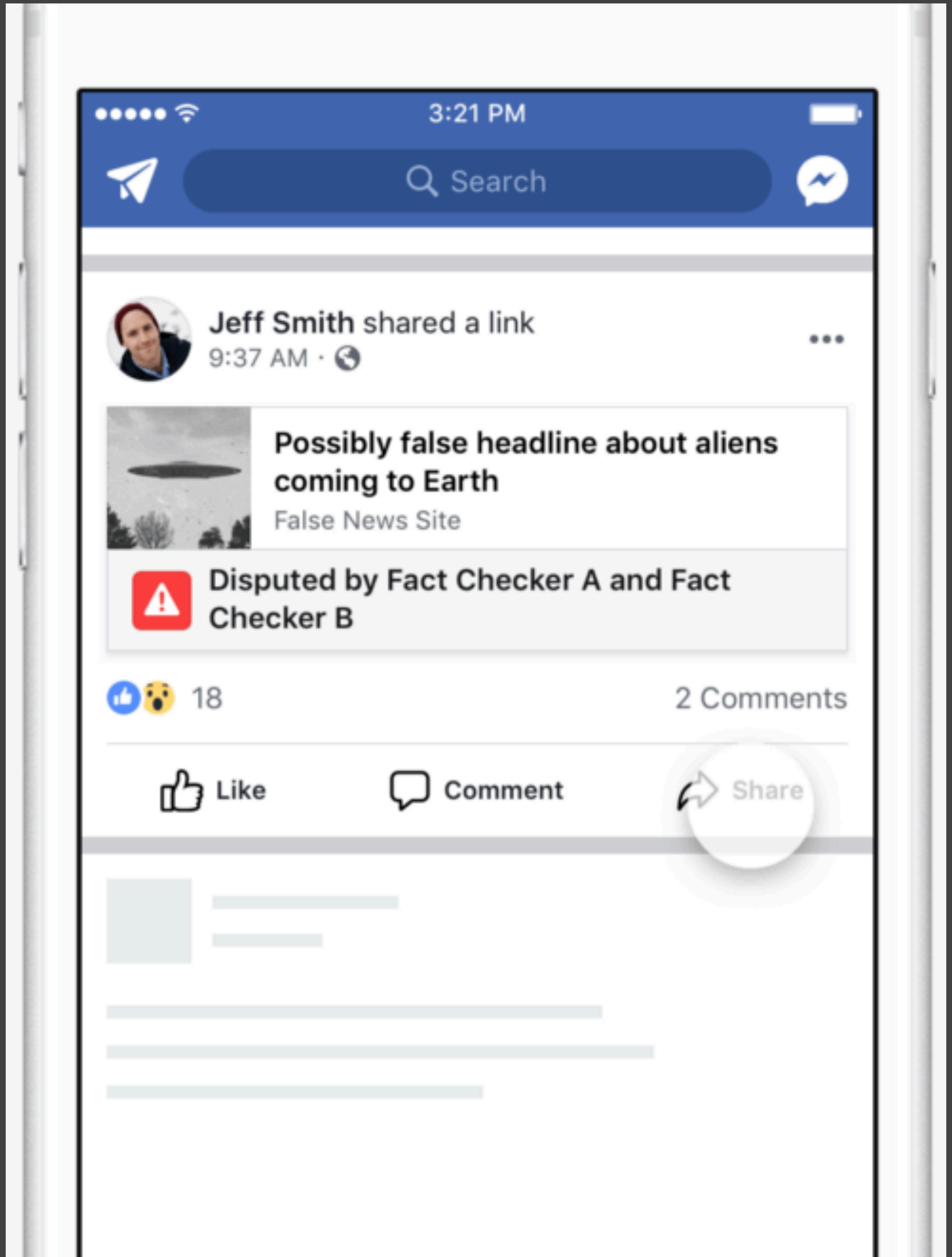
Disputed by Fact Checker A and Fact Checker B

18 2 Comments

Like Comment Share

Facebook's arc

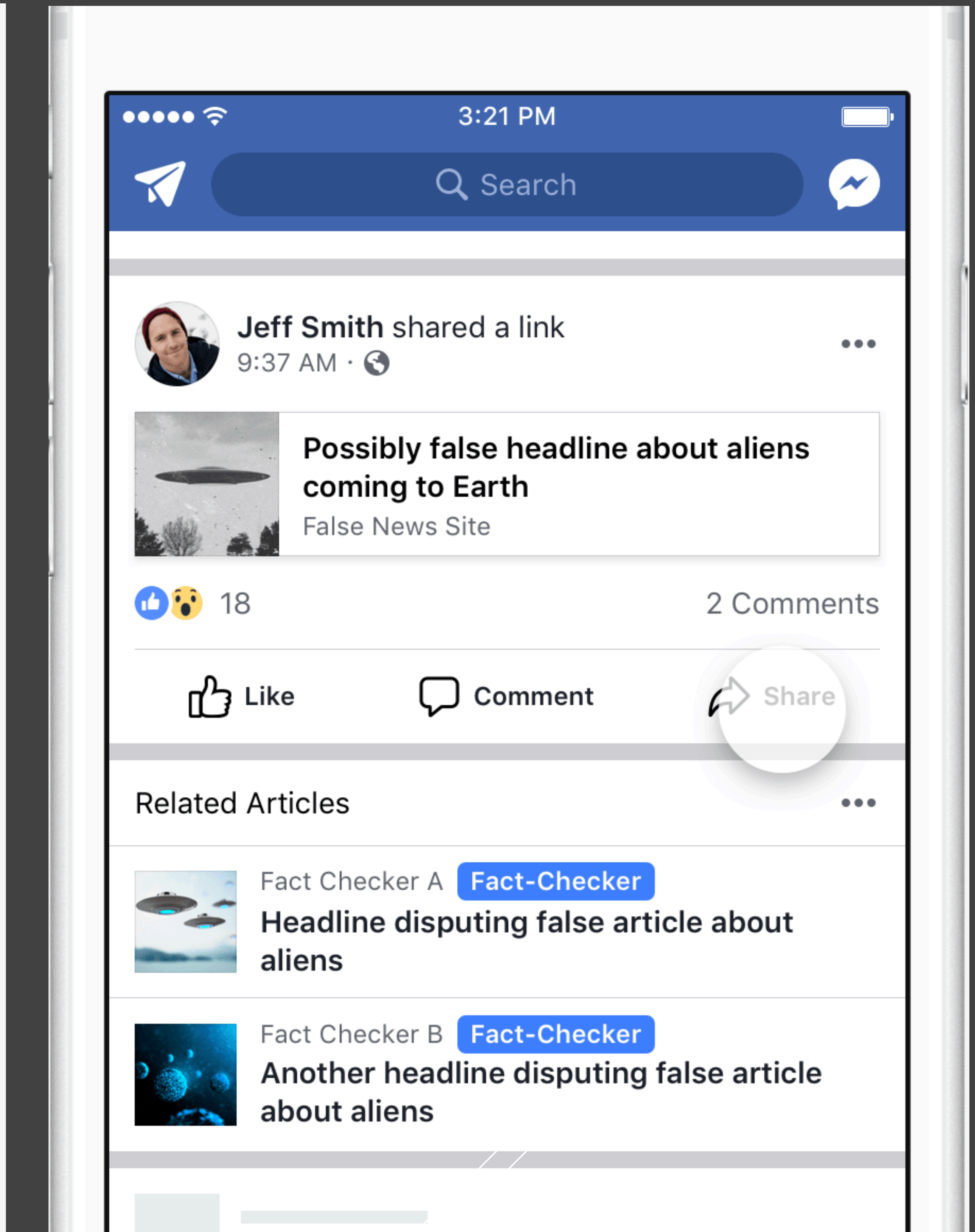
fact check → related articles → fact check



Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

Disputed by Fact Checker A and Fact Checker B



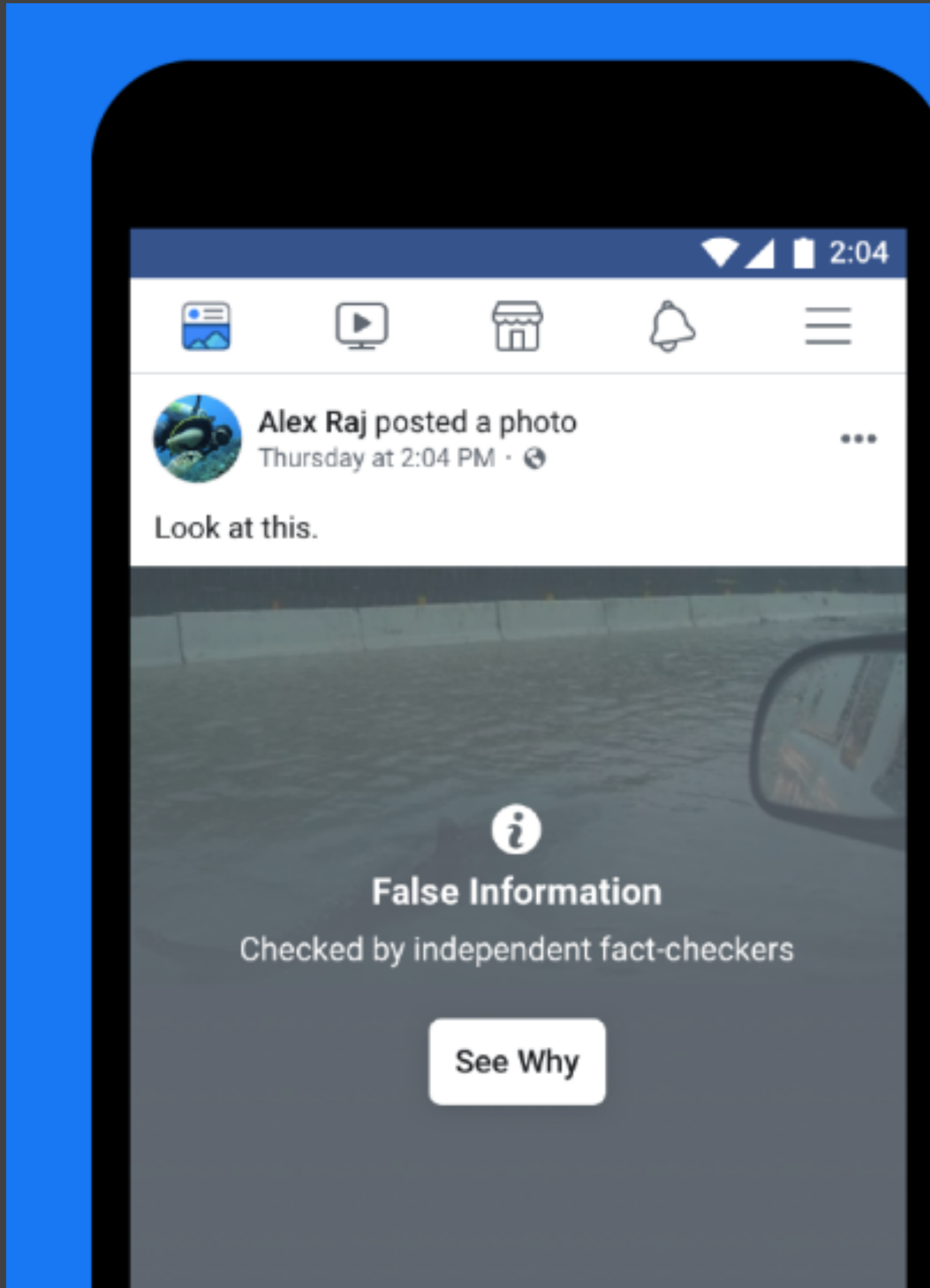
Jeff Smith shared a link
9:37 AM ·

Possibly false headline about aliens coming to Earth
False News Site

Related Articles

Fact Checker A **Fact-Checker**
Headline disputing false article about aliens

Fact Checker B **Fact-Checker**
Another headline disputing false article about aliens



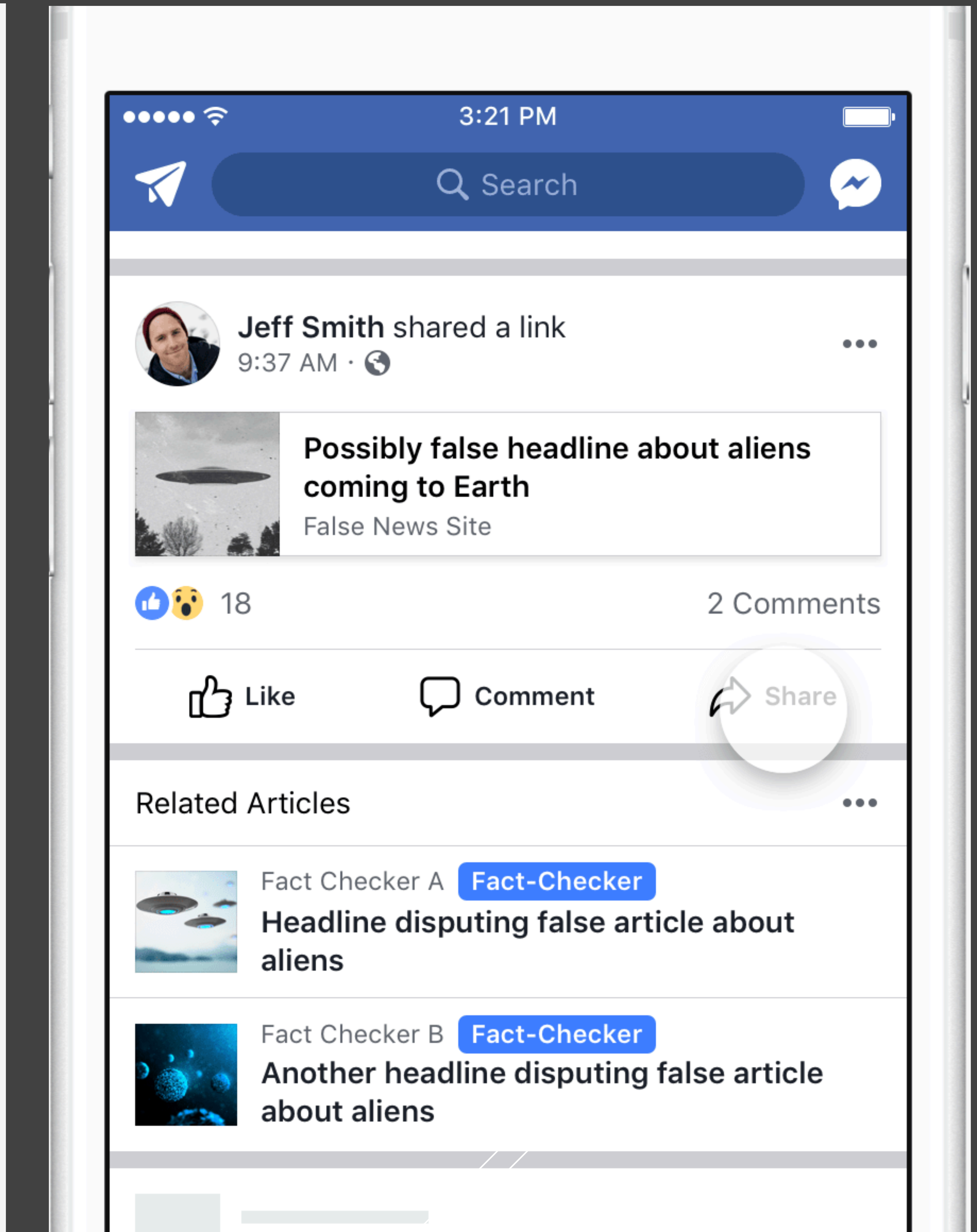
Alex Raj posted a photo
Thursday at 2:04 PM ·

Look at this.

False Information
Checked by independent fact-checkers

Facebook's arc

related articles → fact check → community notes



Jeff Smith shared a link
9:37 AM ·

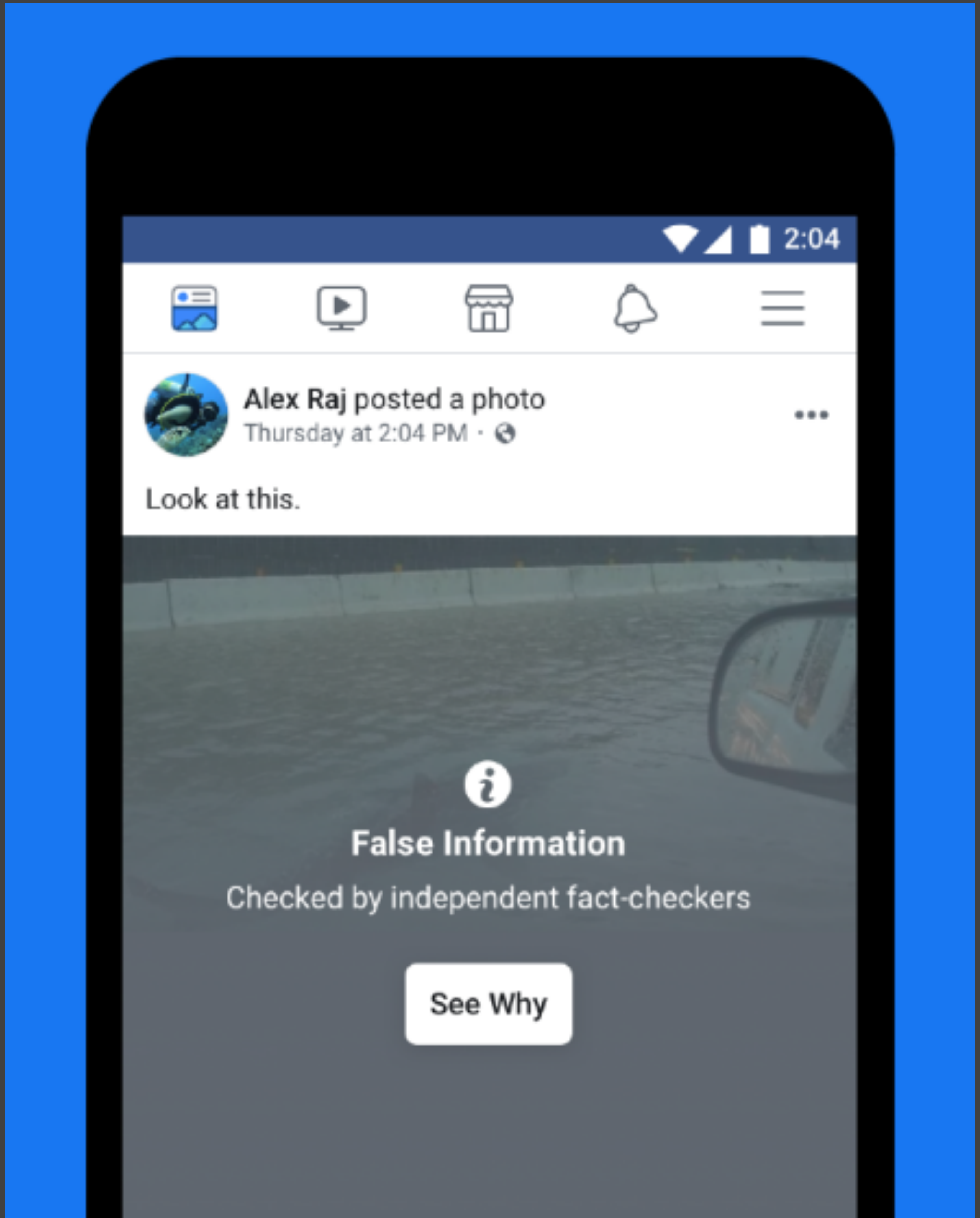
Possibly false headline about aliens coming to Earth
False News Site

18 2 Comments

Related Articles

Fact Checker A Fact-Checker
Headline disputing false article about aliens

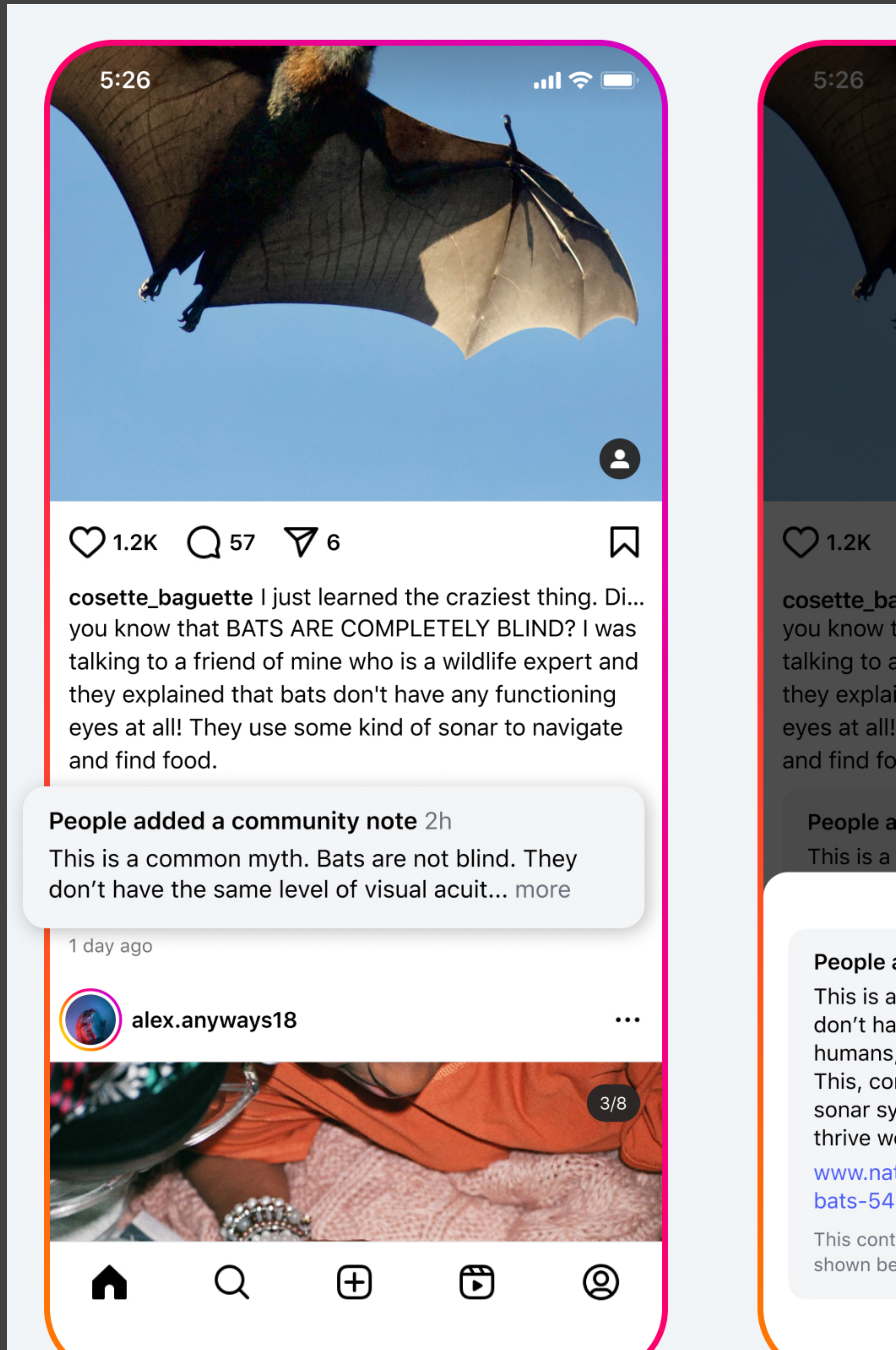
Fact Checker B Fact-Checker
Another headline disputing false article about aliens



Alex Raj posted a photo
Thursday at 2:04 PM ·

Look at this.

False Information
Checked by independent fact-checkers



cosette_baguette I just learned the craziest thing. Di... you know that BATS ARE COMPLETELY BLIND? I was talking to a friend of mine who is a wildlife expert and they explained that bats don't have any functioning eyes at all! They use some kind of sonar to navigate and find food.

People added a community note 2h
This is a common myth. Bats are not blind. They don't have the same level of visual acuit... more

1 day ago

alex.anyways18

cosette_ba
you know t
talking to a
they explai
eyes at all!
and find fo

People a
This is a

People a
This is a
don't ha
humans,
This, co
sonar sy
thrive w
www.nat
bats-54
This cont
shown be

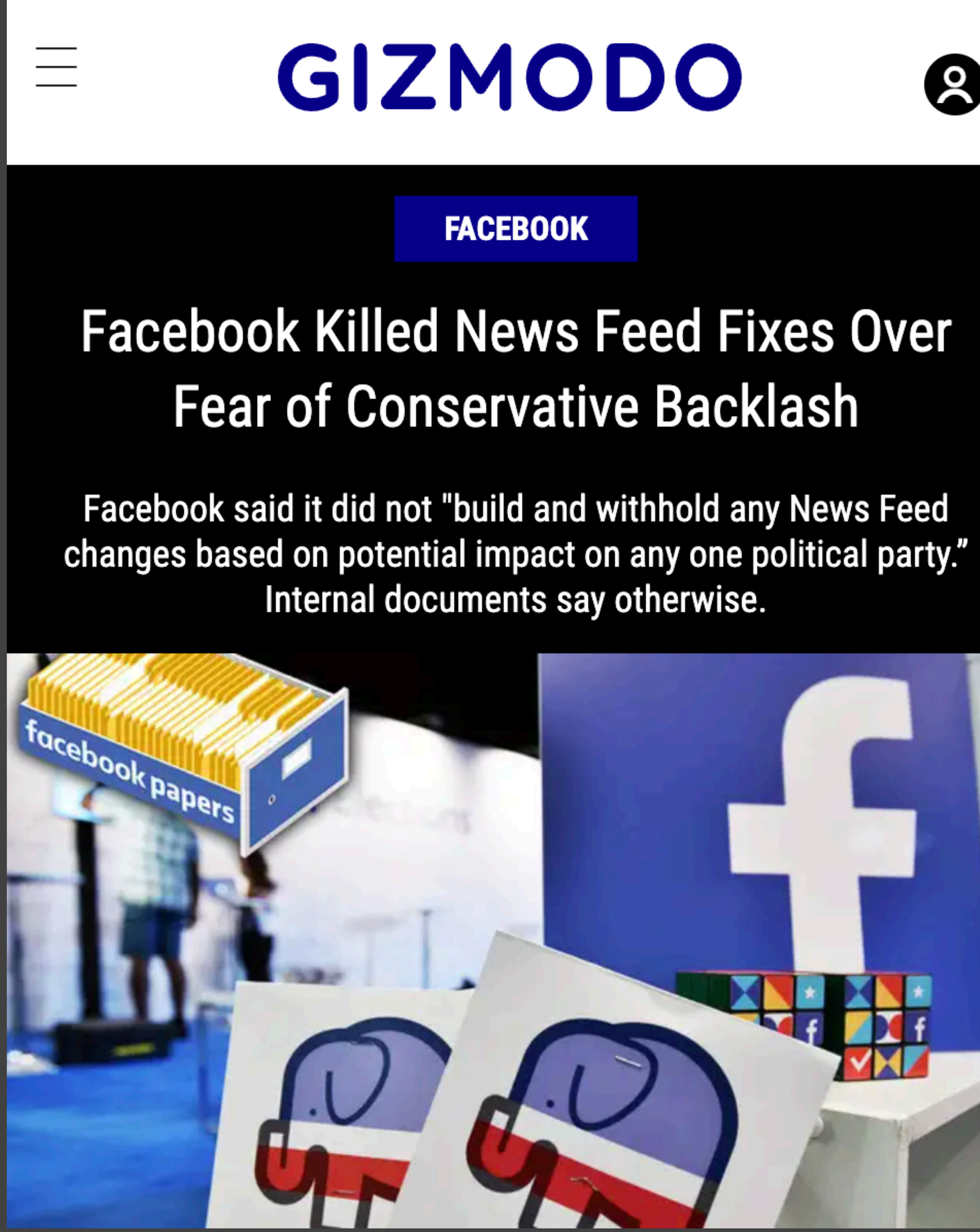
No politically-neutral option

There exists vastly more conservative-leaning disinformation than liberal-leaning disinformation [Hunt and Gentzkow 2017; Törnberg and Chueri 2025]

This difference persists even if the links are evaluated by balanced groups, or by groups of only conservatives [Mosleh et al. 2023]

So the issue is hot-button political, in addition to intersecting questions of freedom of expression

What do you think the platforms should do? [2min]



GIZMODO

FACEBOOK

Facebook Killed News Feed Fixes Over Fear of Conservative Backlash

Facebook said it did not "build and withhold any News Feed changes based on potential impact on any one political party." Internal documents say otherwise.

Summary

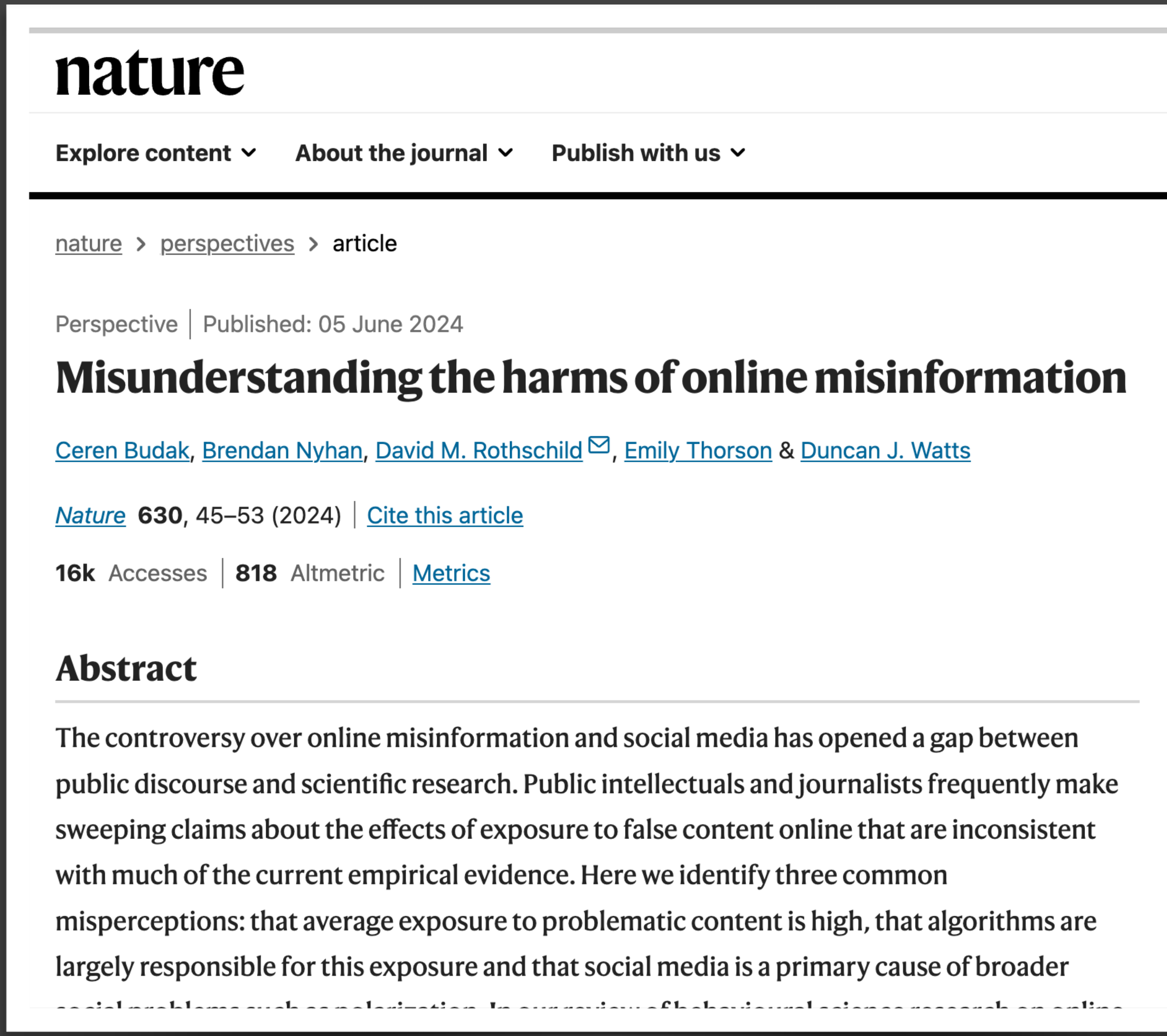
misinformation != disinformation

Disinformation is often created and amplified collectively by motivated actors and their audience

People share misinformation when they are not paying enough attention to accuracy cues

Misinformation is now as much a political issue as it is a sociotechnical one.

For more, check out [Budak et al. 2024]



nature

Perspective | Published: 05 June 2024

Misunderstanding the harms of online misinformation

Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson & Duncan J. Watts

Nature **630**, 45–53 (2024) | Cite this article

16k Accesses | **818** Altmetric | Metrics

Abstract

The controversy over online misinformation and social media has opened a gap between public discourse and scientific research. Public intellectuals and journalists frequently make sweeping claims about the effects of exposure to false content online that are inconsistent with much of the current empirical evidence. Here we identify three common misperceptions: that average exposure to problematic content is high, that algorithms are largely responsible for this exposure and that social media is a primary cause of broader social problems such as polarization. In our review of behavioural science research on online

References

Allcott, Hunt, and Matthew Gentzkow. "Social media and fake news in the 2016 election." Journal of economic perspectives 31.2 (2017): 211-236.

Allen, Jennifer, et al. "Evaluating the fake news problem at the scale of the information ecosystem." Science advances 6.14 (2020): eaay3539

Allen, Jennifer, et al. "Scaling up fact-checking using the wisdom of crowds." Science advances 7.36 (2021): eabf4393.

Allen, Jennifer, Duncan J. Watts, and David G. Rand. "Quantifying the impact of misinformation and vaccine-skeptical content on Facebook." Science 384.6699 (2024): eadk3451.

Aslett, Kevin, et al. "News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions." Science advances 8.18 (2022): eabl3844.

Atreja, Shubham, Hemphill, Libby, and Resnick, Paul. ''Remove, Reduce, Inform: What Actions do People Want Social Media Platforms to Take on Potentially Misleading Content?'' 2023. arXiv

Bail, Christopher A., et al. "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017." Proceedings of the national academy of sciences 117.1 (2020): 243-250.

Baribi-Bartov, Sahar, Briony Swire-Thompson, and Nir Grinberg. "Supersharers of fake news on Twitter." Science 384.6699 (2024): 979-982.

Benkler, Yochai, et al. "Mail-in voter fraud: Anatomy of a disinformation campaign." Berkman Center Research Publication 2020-6 (2020).

Bittman, Lawrence. ''The KGB and Soviet Disinformation: An Insider's View''. 1985. Pergamon-Brassey's, Washington, DC.

References

Bliss, Nadya, et al. "An agenda for disinformation research." arXiv preprint arXiv:2012.08572 (2020).

Budak, Ceren, et al. "Misunderstanding the harms of online misinformation." Nature 630.8015 (2024): 45-53.

Chuai, Yuwei, et al. "Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?." Proceedings of the ACM on Human-Computer Interaction 8.CSCW2 (2024): 1-52.

Farrell, Henry, and Bruce Schneier. "Common-knowledge attacks on democracy." Berkman Klein Center Research Publication 2018-7 (2018).

Grinberg, Nir, et al. "Fake news on Twitter during the 2016 US presidential election." Science 363.6425 (2019): 374-378.

Guess, Andrew M., Brendan Nyhan, and Jason Reifler. "Exposure to untrustworthy websites in the 2016 US election." Nature human behaviour 4.5 (2020): 472-480.

Guess, Andrew, and Alexander Coppock. "Does counter-attitudinal information cause backlash? Results from three large survey experiments." British Journal of Political Science 50.4 (2020): 1497-1515.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." Science advances 5.1 (2019): eaau4586.

Kahan, Dan M. "Misconceptions, misinformation, and the logic of identity-protective cognition." (2017).

Lin, Hause, et al. "Reducing misinformation sharing at scale using digital accuracy prompt ads." PsyArXiv (2024).

References

Martel, Cameron, and David G. Rand. "Fact-checker warning labels are effective even for those who distrust fact-checkers." Nature Human Behaviour 8.10 (2024): 1957-1967.

Mosleh, Mohsen, et al. "Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment." proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

Mosleh, Mohsen, et al. "Trade-offs between reducing misinformation and politically-balanced enforcement on social media." (2023).

Nyhan, Brendan, and Jason Reifler. "When corrections fail: The persistence of political misperceptions." Political Behavior 32.2 (2010): 303-330.

Pennycook, Gordon, and David G. Rand. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation." Nature communications 13.1 (2022): 2333.

Pennycook, Gordon, and David G. Rand. "The psychology of fake news." Trends in cognitive sciences 25.5 (2021): 388-402.

Pennycook, Gordon, et al. "Shifting attention to accuracy can reduce misinformation online." Nature 592.7855 (2021): 590-595.

Pennycook, Gordon, et al. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." Management science 66.11 (2020): 4944-4957.

References

Prochaska, Stephen, et al. "Mobilizing Manufactured Reality: How Participatory Disinformation Shaped Deep Stories to Catalyze Action during the 2020 US Presidential Election." Proceedings of the ACM on Human-Computer Interaction 7.CSCW1 (2023): 1-39.

Renault, Thomas, David Restrepo Amariles, and Aurore Troussel. "Collaboratively adding context to social media posts reduces the sharing of false news." arXiv preprint arXiv:2404.02803 (2024).

Rodrigo, Chris Mills. "Critics Fear Facebook Fact-Checkers Losing Misinformation Fight." The Hill (2020). https://thehill.com/policy/technology/478896-critics-fear-facebook-fact-checkers-losing-misinformation-fight/

Sheng, Jeff T. "Ethnographic uncovering: Hidden communities." Contexts 19.2 (2020): 46-53.

Shirky, Clay. Here comes everybody: The power of organizing without organizations. Penguin, 2008.

Starbird, Kate, Ahmer Arif, and Tom Wilson. "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations." Proceedings of the ACM on Human-Computer Interaction 3.CSCW (2019): 1-26.

Starbird, Kate. ''Beyond 'Bots and Trolls' — Understanding Disinformation as Collaborative Work''. 2019. https://youtu.be/SvEItxWb6Ek

Starbird, Kate. 2021. https://twitter.com/katestarbird/status/1390408145428643842]

References

Stewart, Leo G., Ahmer Arif, and Kate Starbird. "Examining trolls and polarization with a retweet network." Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web. Vol. 70. 2018.

Törnberg, Petter, and Juliana Chueri. "When Do Parties Lie? Misinformation and Radical-Right Populism Across 26 Countries." The International Journal of Press/Politics (2025): 19401612241311886.

Van Bavel, Jay J., and Andrea Pereira. "The partisan brain: An identity-based model of political belief." Trends in cognitive sciences 22.3 (2018): 213-224.

Wood, Thomas, and Ethan Porter. "The elusive backfire effect: Mass attitudes' steadfast factual adherence." Political Behavior 41 (2019): 135-163.

Social Computing

CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.