


I 202: INFORMATION ORGANIZATION & RETRIEVAL FALL 2025

Class 25: Misinformation & Disinformation

HOW ARE THESE SIMILAR / DIFFERENT?

- Rumors
- Misinformation
- Disinformation



MISINFORMATION AND DISINFORMATION



Rumors

Unofficial and **unverified stories** about events that may turn out to be **true or false.**

Misinformation

Information that is false,
but **not necessarily**
intentionally false.

Disinformation

False or misleading information that is **purposefully** seeded and/or spread for a specific objective — e.g., financial, political, or reputational.

Disinformation

Often **built around a kernel of truth**, but layered with distortions or exaggerations.

Functions as a campaign — as set of information actions — rather than a single piece of content.

Mis vs DisInformation:

Motives Matter

MISINFORMATION

Information that is false, but not necessarily intentionally false.

- Uncertainty
- Out of fear
- To inform people
- To protect people
- A display of identity

DISINFORMATION

False and misleading information that is purposely seeded and/or spread for a specific objective.

- Seed and spread chaos
- Monetary gains
- Political objective
- Undermine trust among the public

**SATIRE OR PARODY**

No intention to cause harm but has potential to fool

**MISLEADING CONTENT**

Misleading use of information to frame an issue or individual

**IMPOSTER CONTENT**

When genuine sources are impersonated

**FABRICATED CONTENT**

New content is 100% false, designed to deceive and do harm

**FALSE CONNECTION**

When headlines, visuals or captions don't support the content

**FALSE CONTEXT**

When genuine content is shared with false contextual information

**MANIPULATED CONTENT**

When genuine information or imagery is manipulated to deceive

Types of Mis/Disinformation

1. SATIRE OR PARODY

No intention to harm, but has potential to fool

- Satire or parody is used to bring attention to an issue using intentionally false information. Yet sometimes, people believe the information
- This headline from the satirical news site **The Onion** about the U.S. Centers for Disease Control and Prevention (CDC) is clearly not true.

NEWS IN BRIEF

CDC Announces Children Will Be Last To Receive Covid Vaccine Because What Are Those Little Twerps Going To Do About It

9/23/20 10:09PM • SEE MORE: CORONAVIRUS



'Huh? Huh? What Are You Going To Do, Cry?'

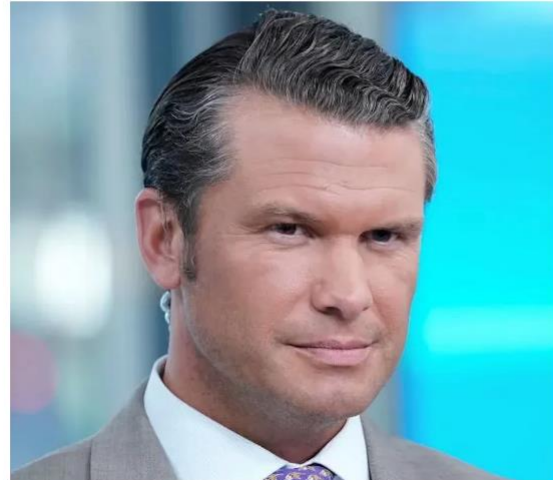
The Borowitz Report

Hegseth Rushed to Walter Reed Hospital After Poking Eye With Mascara Brush

ANDY BOROWITZ
NOV 17, 2025

1,274 333 141

Share



John Lamparski/Getty Images

Types of Mis/Disinformation

2. FALSE CONNECTION

When headlines, visuals and captions do not support the content.

- This image is taken out of context due to a false connection with a hurricane event.
- The message claims that a shark is swimming freely on the freeway in Houston.
- **More:** [“That shark photo is fake — and part of a bigger problem”](#) (USA TODAY, August 29, 2017)

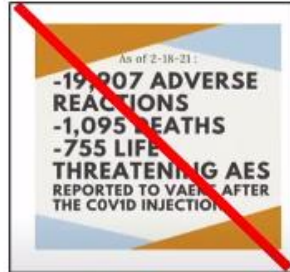


Types of Mis/Disinformation

3. MISLEADING CONTENT

Misleading use of information to frame an issue or individual.

- This image shows figures from the Vaccine Adverse Reporting System, where anyone can report symptoms after receiving the vaccine.
- However, the symptoms may not be related to the vaccine and have not been verified.
- These figures can mislead people to believing the vaccine produces high numbers of adverse reactions.



4. FALSE CONTEXT

When genuine content is shared with false contextual information.



"One of the things we want to make sure people understand is that they **should not be unnecessarily alarmed** if there are reports, once we start vaccinating, of someone or multiple **people dying** within a day or two of their vaccination who are residents of a long-term care facility. **That would be something we would expect, as a normal occurrence, because people die frequently in nursing homes.**"

- The headline in the post is phrased to suggest that getting vaccinated can lead to many deaths. However, the main body of the actual article states that there's no correlation between getting vaccinated and dying. The cases of vaccinated residents of nursing homes passing away has no connection to the vaccine.

Types of Mis/Disinformation

5. IMPOSTER CONTENT

Material involving impersonation of genuine sources, like using the branding of an established news agency.

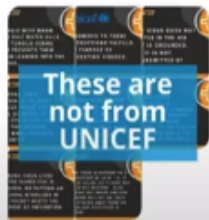
- These headlines warn of fake websites that scam people into providing personal information.
- The fake websites impersonate genuine sources of information.

Feds warn of COVID-19 vaccine scammers using fake websites

By Natalie O'Neill

December 16, 2020 | 5:26pm | Updated

University of Iowa Health Care warns about vaccine-schedule scam



6. MANIPULATED CONTENT

Content that contains distortions of genuine information or imagery.

- This image shows chairs at the beach depicting an area that is set aside exclusively for vaccinated individuals.
- This is entirely fabricated, meant to spread a narrative that vaccinated individuals will enjoy perks and be afforded different rights.



Types of Mis/Disinformation

7. FABRICATED CONTENT

False content that is meant to deceive and do harm.

- This is a fake tweet that appears as though it was set from U.S. Senator Ted Cruz of Texas. It was circulated on social media during the February 2021 Texas winter storm that left people without power and high energy bills.



Spreading Mis/DisInformation –Trolls (Impersonators)

2016 U.S. ELECTION DISINFORMATION

- **RIGHT:** This image is a network graph of online Twitter conversations around Black Lives Matter (BLM) during the time of 2016 U.S. elections.
- The blue dots are authentic accounts belonging to pro- and anti-BLM individuals. Most interactions among members occurred within each of the communities.
- The orange dots and lines show the presence of Russian trolls. They successfully created fake personas and content and had their message retweeted millions of times. As shown, they infiltrated both communities with fake personas and content.

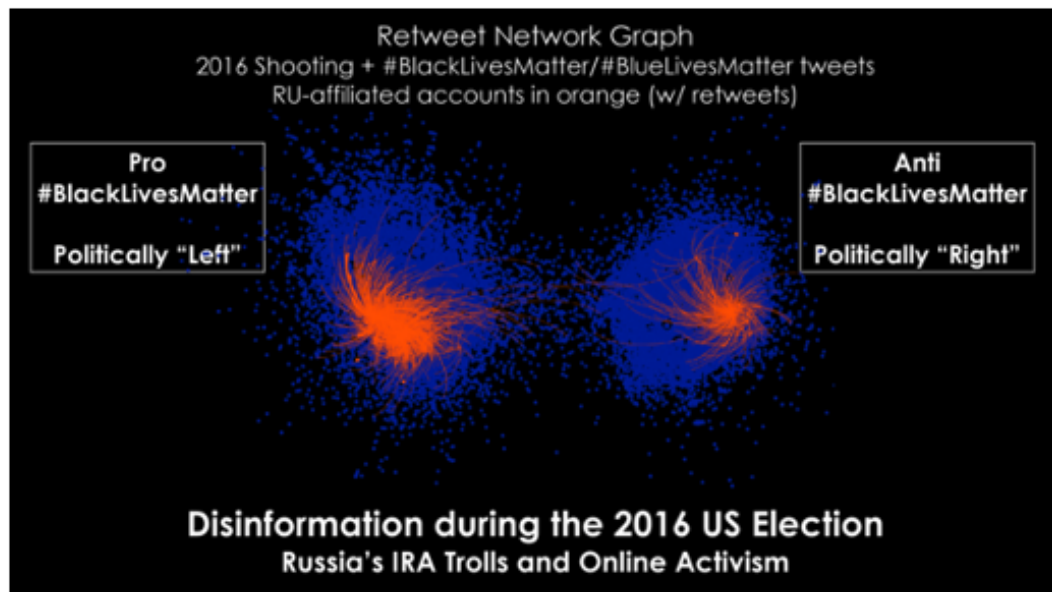


Image | Kate Starbird, UW HCDE, UW Center for an Informed Public

Starbird et al., Disinformation as Collaborative Work, CHI 2019

Slide adapted from Jevin West

STUDY OF USE OF VISUALIZATION BY “ANTI-MASKERS”

- Analyzed Facebook and Twitter posts with grounded coding

“This paper investigates how pandemic visualizations circulated on social media, and shows that people who mistrust the scientific establishment often deploy the same rhetoric of data-driven decision-making used by experts, but to advocate for radical policy changes.”

<https://vis.csail.mit.edu/covid-story/>

A vibrant, painterly landscape painting. In the foreground, a multi-tiered waterfall cascades over dark rocks, its water rendered with soft, flowing brushstrokes. To the left, a dense forest of tall, dark evergreen trees stands on a grassy slope. In the background, a large, dark, craggy mountain peak rises against a dramatic sky. The sky is filled with warm, orange, and yellow clouds, with a bright sun or moon setting or rising on the right side, creating a strong starburst effect with long, radiating rays of light. The overall style is reminiscent of classical landscape art, with visible brushwork and a rich, warm color palette.

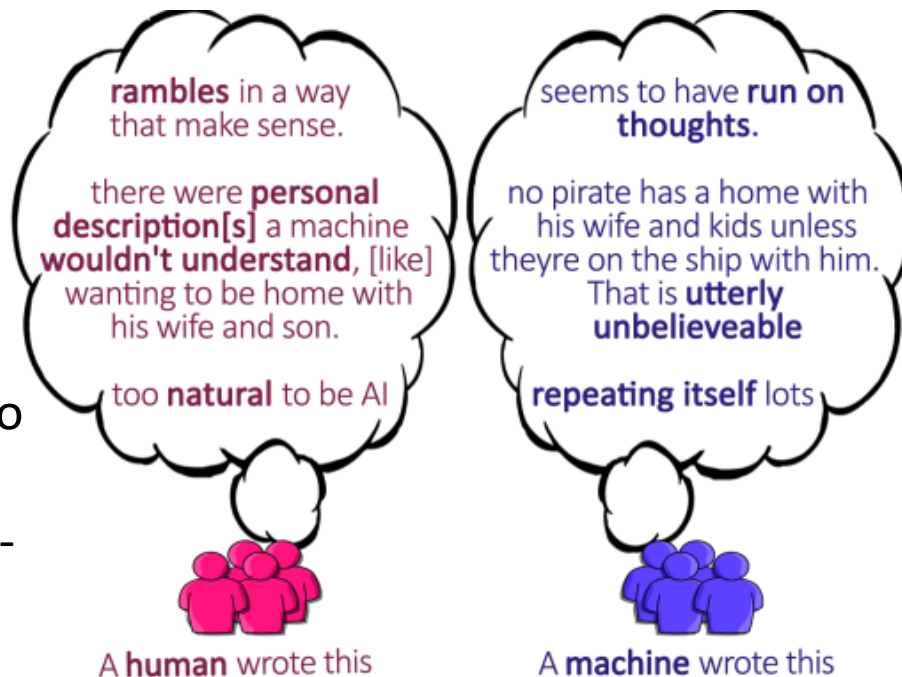
WHAT ABOUT AI?

<https://deepdreamgenerator.com/#gallery>

Automatically Generated Text

Once upon a time, there lived a pirate. He was the sort of pirate who would rather spend his time chasing away the sharks swimming around his ship than sail to foreign ports in search of booty. He was a good pirate, a noble pirate, an honest pirate. He was a pirate who would rather be at home with his wife and son than out on a ship in the middle of the ocean.

With no training, crowd workers were no better than chance at guessing if paragraphs generated by human or GPT-3 model



Once upon a time, there was a man in a place that was not a place at all.

He didn't know anything of a place or a time or who he was or what he was doing there. There was just him and the silence.

He sat there for a long time, not knowing what he was doing there. He thought, thought and thought, but he didn't know what to think. There was just him and the silence. He tried to speak, but no sound came from his mouth. He tried to move, but his body would not move. He sat there, but he didn't know for how long he was there.

* What do you think the source of this text is?

- ☒ **Definitely human-written**
- ☐ Possibly human-written
- ☐ Possibly machine-generated
- ☐ **Definitely machine-generated -- Correct Answer**

You cannot change your answer once you click submit.

Explanation

Note how the story is repetitive and doesn't seem to go anywhere.

Got it, next question

human-authored

Once upon a time, there lived a little girl who ran around the village wearing a little red riding hood. Don't ask me what a riding hood is because I don't even know. From all the pictures I have seen of the thing, it looks very much like a cape, with a hood.

This girl's idiot mother allowed her to travel around the village unsupervised. Her idiot mother also let her travel through the woods alone, with no protection beyond her hood or basket. Not a very smart parent, if you ask me. This girl can't have been older than ten or eleven.

machine-authored

Once upon a time, there was a man in a place that was not a place at all.

He didn't know anything of a place or a time or who he was or what he was doing there. There was just him and the silence.

He sat there for a long time, not knowing what he was doing there. He thought, thought and thought, but he didn't know what to think. There was just him and the silence. He tried to speak, but no sound came from his mouth. He tried to move, but his body would not move. He sat there, but he didn't know for how long he was there.

With training, accuracy improves slightly to 55%

Done, show me the next example

AUTOMATICALLY GENERATED TEXT

- Why might this be a misinfo problem? Or is it not?
- Is it hard for humans to write convincing misinformation without the help of programs?

● This article is more than **2 years old**

I thought I was immune to being fooled online. Then I saw the pope in a coat

Joel Golby

An encounter with an AI-generated image of his holiness has changed me: I now have sympathy for credulous baby boomers



📷 'I thought wearing a really big coat and looking like a Metal Gear Solid 2 boss battle might have been part of his ongoing cool guy shtick. Lord, forgive me.' Photograph: Reddit

AI-based Media Manipulation (Deep Fakes)

WHAT ARE DEEPFAKES?

A **deepfake** is a photo, video, or audio track created using artificial intelligence (AI) techniques to realistically simulate or alter people's faces, movements, and voices, among other simulations.

- Deepfake technology has enabled important advances in artistic expression, filmmaking, education, accessibility, and even historical or forensic re-creation.
- This technology allows media to be easily and convincingly altered or fabricated to make it seem as if someone did or said something that they never did. Hence, deepfakes can be used to mislead, misinform, and cause harm.

Deepfakes can be ...

- **POTENT**
Deepfakes can be difficult to distinguish due to increasingly advanced and evolving technology.
- **UNIVERSAL**
Deepfakes can be shared with a global audience across social media platforms.
- **IMMEDIATE**
Deepfakes can use open-source AI tools that basic skills can be used to easily create deepfakes.

To date most applications of deepfake technology have been for entertainment such as celebrity impersonations, and personal vendettas such as revenge porn and other exploitative uses.

AI-based Media Manipulation (Deep Fakes)

SPOTTING DEEPFAKES: WHICH FACE IS REAL?

It's becoming increasingly difficult to spot deep fakes due to continuous technological advancements.

With images there are ways to spot inconsistencies and help distinguish fake images from real ones, but few people go through the effort.

Can you guess which image on the right is real?



Image A



Image B

QUIZ | <https://www.whichfaceisreal.com/>

AI-based Media Manipulation (Deep Fakes)

HOW DEEPPAKES THREATEN DEMOCRACY: A CASE STUDY

In 2019, an attempted coup d'état took place in Gabon because people thought an authentic video of the President was a deepfake when it was actually real.

- Several controversies around the health of the President of Gabon led to months of misinformation. **A New Years Eve address by the President had people questioning the authenticity of the video and many believed it was a deepfake.**



- **LEARN MORE** | [“The suspicious video that helped spark an attempted coup in Gabon”](#) (Washington Post Factchecker, February 2020).

HOW TO COMBAT MISINFORMATION?



STRATEGY: CONTENT LABELS

It matters how platforms label manipulated media. Here are 12 principles designers should follow

June 10, 2020



AUTHOR

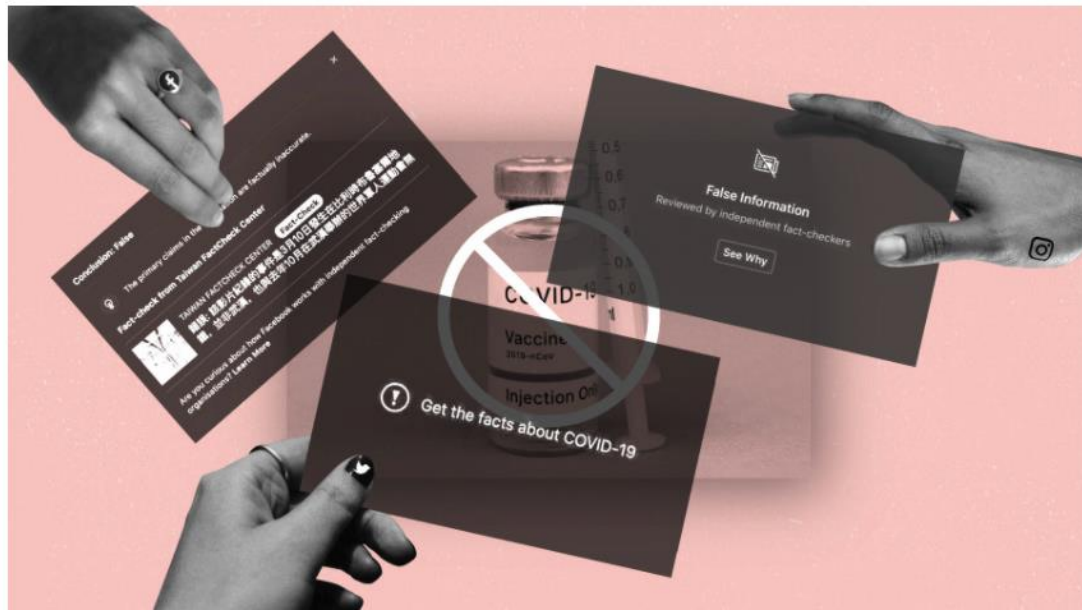
First Draft Guest

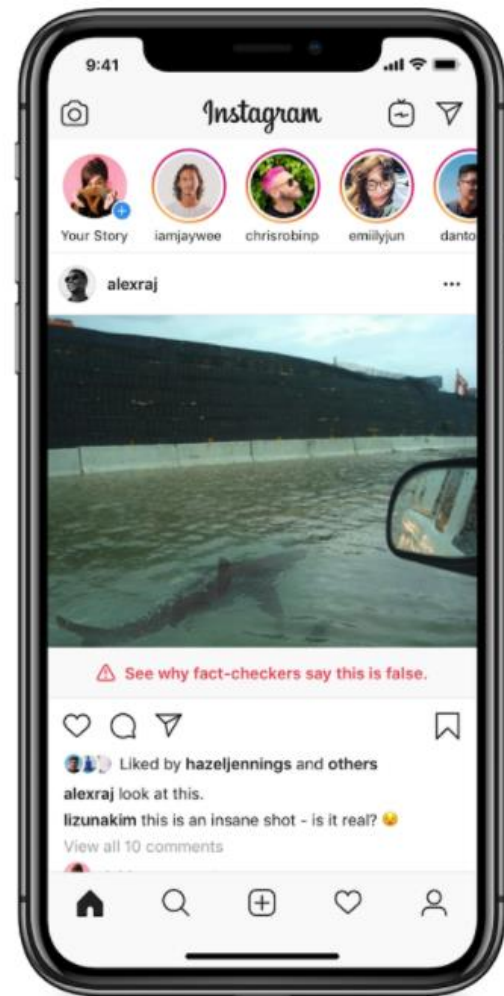
SHARE



TAGS

disinformation,
information disorder,
labelling, manipulated
media, manipulated

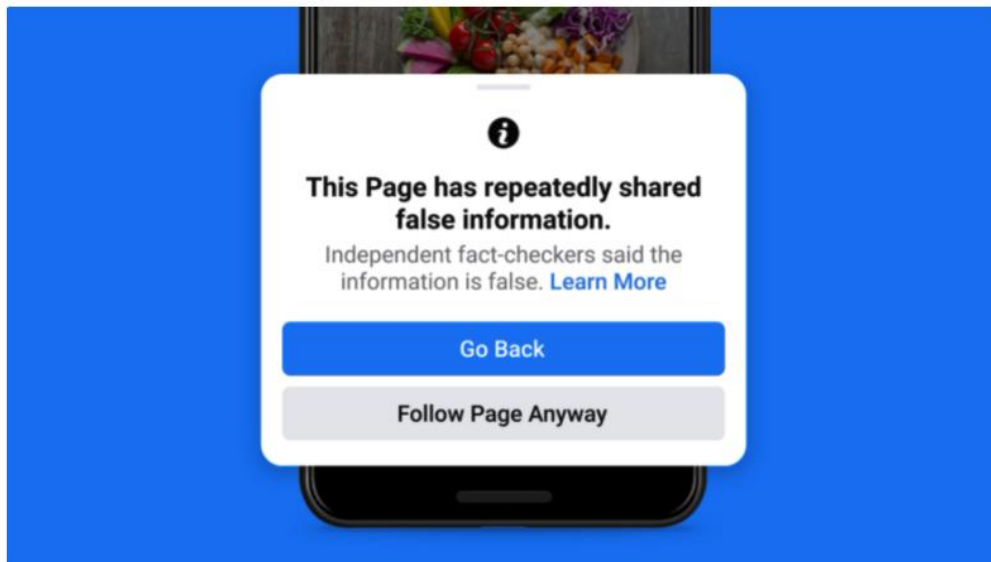




Facebook app

Taking Action Against People Who Repeatedly Share Misinformation

May 26, 2021



Counter-Measure: Fact Checking Web Sites

[SUBMIT A RUMOR](#)[LATEST](#)[TRENDING](#)[NEWS & POLITICS](#)[ENTERTAINMENT](#)[FACT CHECKS](#)[QUIZ](#)[Snopes Membership](#)

About Us

When misinformation obscures the truth and readers don't know what to trust, Snopes' fact-checking and original investigative reporting lights the way to evidence-based and contextualized analysis. We always link to and document our sources so readers are empowered to do independent research and make up their own minds.

Snopes got its start in 1994, investigating urban legends, hoaxes, and folklore. As demand for reliable online fact checks grew, so did Snopes. Now it's the oldest and largest fact-checking site, widely regarded by journalists, folklorists, and readers as an invaluable research companion. Snopes.com is an independent publication owned and operated under "Snopes, Inc" which is a wholly-owned subsidiary of "Snopes Media Group, Inc" (SMG). SMG is owned by [Chris Richmond](#) (60%) and Drew Schoentrup (40%).

Snopes is a member of [IFCN](#) and in compliance with the International Fact Checking Network's highest [standards](#) of combating misinformation online. Should you need to file a complaint with IFCN, you may do so [here](#) but please [contact us](#) directly.

[Explore our FAQ](#) to learn more about us, or consider [becoming a Snopes member](#). If you have any questions, please [let us know](#).

About Us

[Transparency](#)[Fact Check Ratings](#)[Our Team](#)[Disclosures](#)[Media Mentions](#)[Privacy Policy](#)[Terms and Conditions](#)

“RUMOR CASCADES” STUDY

- Study done on Facebook in 2013
- “Rumors” categorized as true, false, or maybe
- How do they spread in social networks with “share” buttons?
- What happens when someone links to a fact-checking cite like Snopes.com?

A sample of 249,035 comments on either photos or shares of photos, posted during July and August 2013 and containing a valid link to a rumor covered by Snopes

Tagged 16,672 individual cascades, containing 62,497,651 shares.

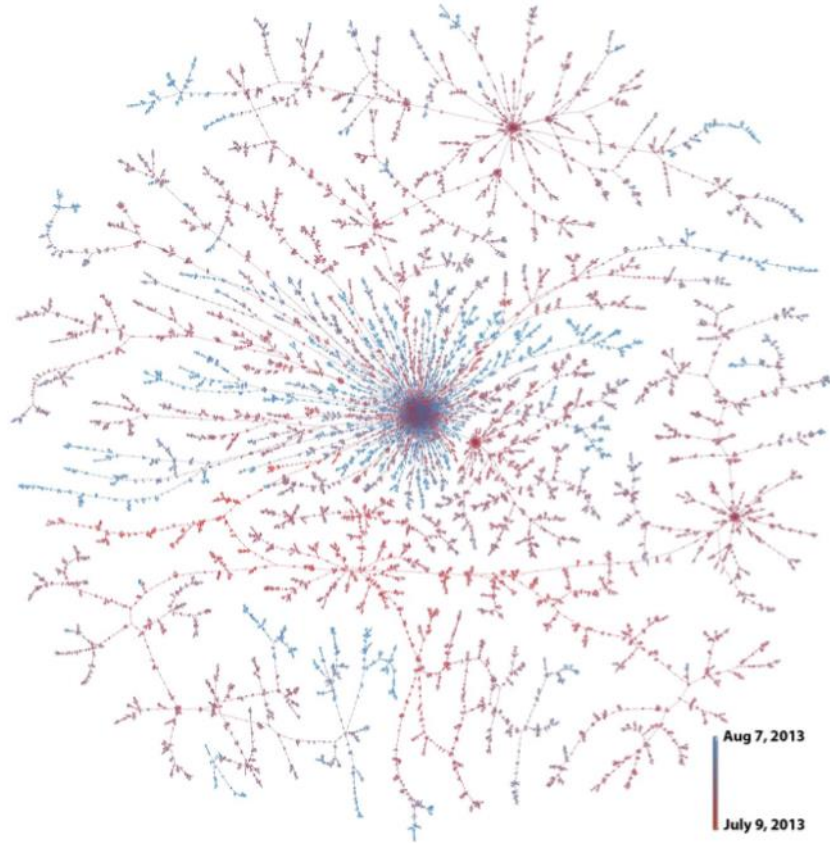


Figure 2: Cascade of reshares of a Cabela's sporting goods store receipt attributing addition sales tax to "Obamacare". Shares which did not prompt further shares are excluded. The coloring is from early (red) to late (blue).

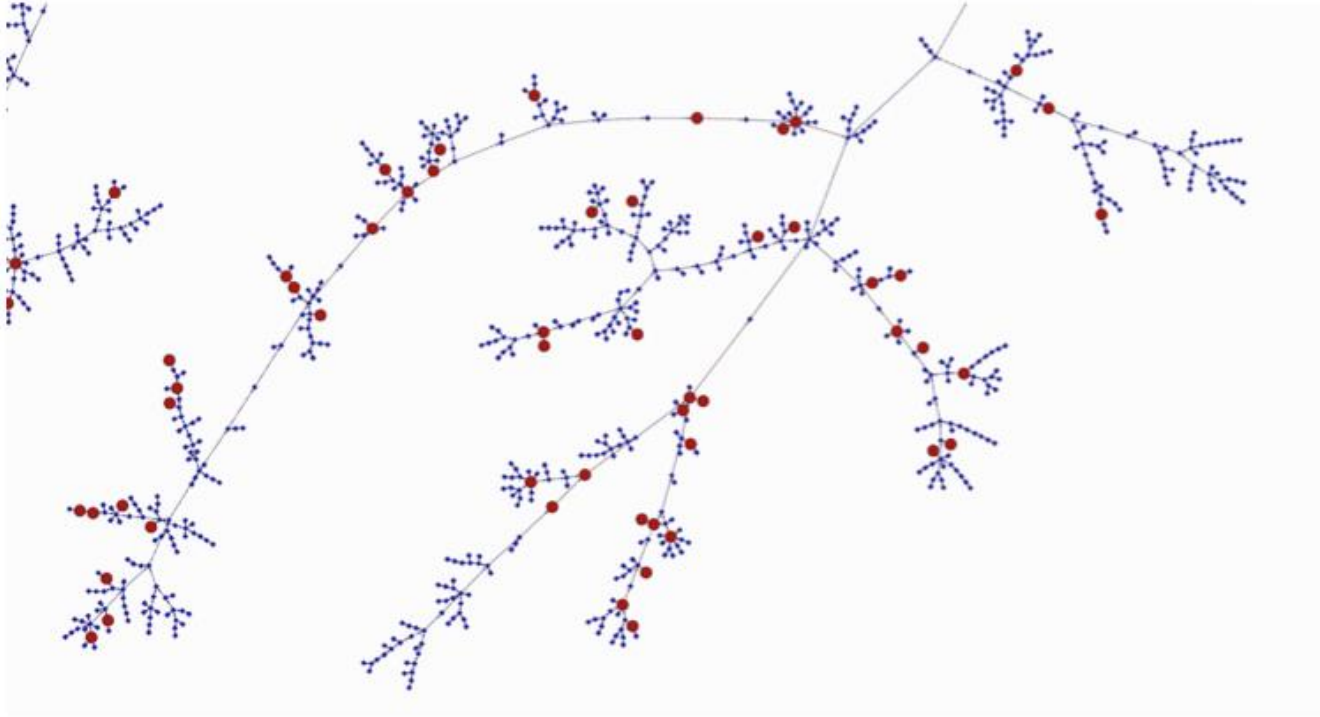


Figure 3: A branch of the cascade shown in Figure 2. Shares that received a Snopes link are displayed in red.

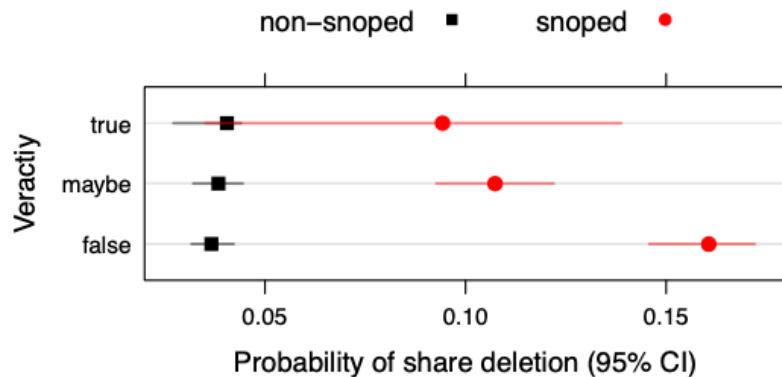


Figure 10: Probability that a reshare will be deleted as a function of the veracity of the rumor and whether it was snoped.

Evidence that rapid fact-checking can help mitigate false rumors (a bit)

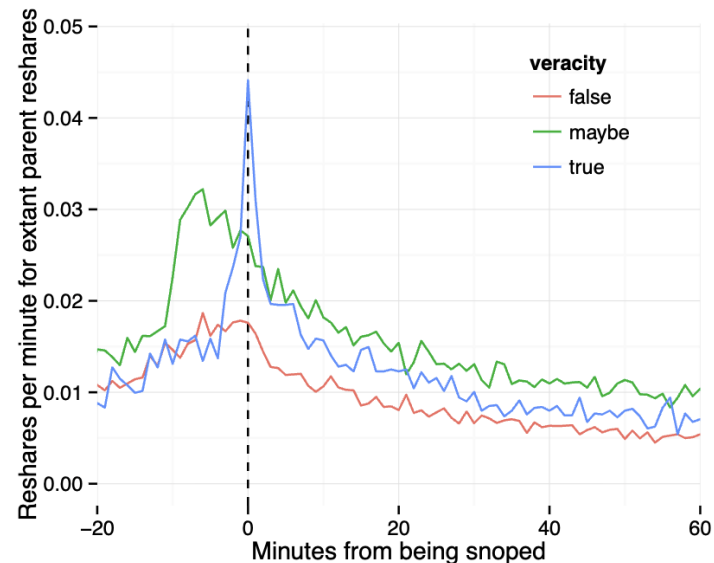


Figure 13: Reproductive rate of reshares that are eventually snoped by time from being snoped. The act of snoping coincides with a boost for true stories, but appears to be detrimental in the short run to the propagation of the false cascade.

STRATEGY: DE/PREBUNKING



The Debunking Handbook 2020 ...

Authors

Stephan Lewandowsky
University of Bristol and
University of Western Australia
cogsciwa.com

John Cook
George Mason University
climatechangecommunication.org

Ullrich Ecker
University of Western Australia
emc-lab.org

Dolores Albarracín
University of Illinois at Urbana Champaign
psychology.illinois.edu/people/dalbarra

Michelle A. Amazeen
Boston University
bu.edu/com/profile/michelle-amazeen/

Panayiota Kendeou
Department of Educational Psychology,
University of Minnesota
cehd.umn.edu/edpsych/people/kend0040/

Doug Lombardi
University of Maryland
sciencelearning.net

Eryn J. Newman
Research School of Psychology,
The Australian National University
erynjnewman.com

Gordon Pennycook
Hill Levene Schools of Business, University of Regina
gordonpennycook.net

Ethan Porter
School of Media and Public Affairs; Institute for Data,
Democracy and Politics; Department of Political
Science (courtesy), George Washington University
ethanporter.com

David G. Rand
Sloan School and Department of Brain and
Cognitive Sciences, MIT
daverand.org

David N. Rapp
School of Education and Social Policy & Department
of Psychology, Northwestern University
rapplab.sesp.northwestern.edu

Jason Reifler
University of Exeter
jasonreifler.com

Jon Roozenbeek
University of Cambridge
chu.cam.ac.uk/people/view/jon-roozenbeek

Philipp Schmid
Department of Psychology, University of Erfurt
philippschmid.org

Colleen M. Seifert
University of Michigan
lsa.umich.edu/psych

Gale M. Sinatra
Rossier School of Education,
University of Southern California
motivatedchangelab.com/

Briony Swire-Thompson
Network Science Institute, Northeastern University
Institute of Quantitative Social Science,
Harvard University,
brionyswire.com


Sander van der Linden
Department of Psychology, University of Cambridge
psychol.cam.ac.uk/people/sander-van-der-linden

Emily K. Vraga
Hubbard School of Journalism and Mass
Communication, University of Minnesota
emilyk.vraga.org

Thomas J. Wood
Department of Political Science, Ohio State University
polisci.osu.edu/people/wood.1080

Maria S. Zaragoza
Department of Psychology, Kent State University
kent.edu/psychology/profile/maria-s-zaragoza

Reviewers: Lisa Fazio, Anastasia Kozyreva, Philipp
Lorenz-Spreen, Jay Van Bavel
Graphic Design: Wendy Cook



STRATEGIES FOR COMBATING MISINFORMATION

- Standard approach: debunking
 - Can help, but there are barriers
 - Can backfire, but there is increasing evidence that this is not as much of an issue as once thought
- Newer, more successful approach: **inoculation**
 - Expose people to weakened versions of the misinformation
 - Slowly build up cognitive resistance
 - Prophylactic vs Therapeutic (for pre-existing views)

PREBUNKING

- **Prevention (Prebunking)**
- The most effective strategy is to prevent misinformation from taking root in the first place. This is achieved through a technique called **inoculation** or **prebunking**.
- This involves:
 - ***Forewarning*** people that they may be exposed to misinformation.
 - ***Explaining the manipulative tactics*** or misleading argumentation strategies used by those who spread false information (e.g., *ad hominem attacks, false analogies, emotional appeals*).
- By exposing people to a weakened dose of the techniques, they cultivate "**cognitive antibodies**" that make them resilient to subsequent manipulation attempts.

PRE/DEBUNKING METHOD

- Replace a myth with a clear, simple factual alternative
- Avoid reinforcing the myth through repetition or complexity,
- Use visuals to strengthen understanding
- Inoculate people against misleading tactics
- Delivered this with respectful communication.

FACT

Lead with the fact if it's clear, pithy, and sticky—make it simple, concrete, and plausible. It must “fit” with the story.

WARN ABOUT THE MYTH

Warn beforehand that a myth is coming... mention it once only.

EXPLAIN FALLACY

Explain how the myth misleads.

FACT

Finish by reinforcing the fact—multiple times if possible. Make sure it provides an alternative causal explanation.

FACT

Scientists observe human fingerprints all over our climate

The warming effect from greenhouse gases like carbon dioxide has been confirmed by many lines of evidence. Aircraft and satellites measure less heat escaping to space at the exact wavelengths that carbon dioxide absorbs energy. The upper atmosphere cools while the lower atmosphere warms—a distinct pattern of greenhouse warming.

Lead with the fact if it's clear, pithy, and sticky—make it simple, concrete, and plausible.

Provide a factual alternative that fills a causal "gap", explaining what happened if the misinformation is corrected.

Do not rely on a simple retraction ("this claim is not true").

MYTH

A common climate myth is that climate has always changed naturally in the past, therefore modern climate change must be natural also.

Warn that a myth is coming.

Repeat the misinformation, only once, directly prior to the correction.

FALLACY

This argument commits the single cause fallacy, falsely assuming that because natural factors have caused climate change in the past, then they must always be the cause of climate change.

Explain how the myth misleads.

This logic is the same as seeing a murdered body and concluding that people have died of natural causes in the past, so the murder victim must have also died of natural causes.

Point out logical or argumentative fallacies underlying the misinformation.

FACT

Just as a detective finds clues in a crime scene, scientists have found many clues in climate measurements confirming humans are causing global warming. Human-caused global warming is a measured fact.

Finish by reinforcing the fact.

Repeat the fact multiple times if possible.

ANIMATED PREBUNKING VIDEOS

False Dichotomies



→ Watch Video

A false dichotomy (or false dilemma) is a logical fallacy in which a limited number of choices or sides are presented as mutually exclusive, when in reality more options are available. It's also known as the "either-or fallacy".

Debunking + AI

RESEARCH

RESEARCH ARTICLE SUMMARY

ARTIFICIAL INTELLIGENCE

Durably reducing conspiracy beliefs through dialogues with AI

Thomas H. Costello*, Gordon Pennycook, David G. Rand

INTRODUCTION: Widespread belief in unsubstantiated conspiracy theories is a major source of public concern and a focus of scholarly research. Despite often being quite implausible, many such conspiracies are widely believed. Prominent psychological theories propose that many people want to adopt conspiracy theories (to satisfy underlying psychic “needs” or motivations), and thus, believers cannot be convinced to abandon these unfounded and implausible beliefs using facts and counter-evidence. Here, we question this conventional wisdom and ask whether it may be possible to talk people out of the conspiratorial “rabbit hole” with sufficiently compelling evidence.

RATIONALE: We hypothesized that interventions based on factual, corrective information may seem ineffective simply because they lack sufficient depth and personalization. To test this hypothesis, we leveraged advancements in large language models (LLMs), a form of artificial intelligence (AI) that has access to vast amounts of information and the ability to generate bespoke arguments. LLMs can thereby directly refute particular evidence each

individual cites as supporting their conspiratorial beliefs.

To do so, we developed a pipeline for conducting behavioral science research using real-time, personalized interactions between research subjects and AI. Across two experiments, 2190 Americans articulated—in their own words—a conspiracy theory in which they believe, along with the evidence they think supports this theory. They then engaged in a three-round conversation with the LLM GPT-4 Turbo, which we prompted to respond to this specific evidence while trying to reduce participants’ belief in the conspiracy theory (or, as a control condition, to converse with the AI about an unrelated topic).

RESULTS: The treatment reduced participants’ belief in their chosen conspiracy theory by 20% on average. This effect persisted undiminished for at least 2 months; was consistently observed across a wide range of conspiracy theories, from classic conspiracies involving the assassination of John F. Kennedy, aliens, and the illuminati, to those pertaining to topical events such as COVID-19 and the 2020 US presidential elec-

tion; and occurred even for participants whose conspiracy beliefs were deeply entrenched and important to their identities. Notably, the AI did not reduce belief in true conspiracies. Furthermore, when a professional fact-checker evaluated a sample of 128 claims made by the AI, 99.2% were true, 0.8% were misleading, and none were false. The debunking also spilled over to reduce beliefs in unrelated conspiracies, indicating a general decrease in conspiratorial worldview, and increased intentions to rebut other conspiracy believers.

CONCLUSION: Many people who strongly believe in seemingly fact-resistant conspiratorial beliefs can change their minds when presented with compelling evidence. From a theoretical perspective, this paints a surprisingly optimistic picture of human reasoning: Conspiratorial rabbit holes may indeed have an exit. Psychological needs and motivations do not inherently blind conspiracists to evidence—it simply takes the right evidence to reach them. Practically, by demonstrating the persuasive power of LLMs, our findings emphasize both the potential positive impacts of generative AI when deployed responsibly and the pressing importance of minimizing opportunities for this technology to be used irresponsibly. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: tcostello@american.edu
Cite this article as T. H. Costello *et al.*, *Science* 385, eadq1814 (2024). DOI: [10.1126/science.adq1814](https://doi.org/10.1126/science.adq1814)

S READ THE FULL ARTICLE AT
<https://doi.org/10.1126/science.adq1814>

USING AI TO DEBUNK

- Hypothesized that interventions based on factual, corrective information may seem ineffective simply because they lack sufficient depth and personalization.
- To test this, used an LLM chatbot to generate bespoke arguments.
- Study design:
 - *2190 Americans articulated—in their own words—a conspiracy theory in which they believe, along with the evidence they think supports this theory.*
 - *Then engaged in a three-round conversation with the LLM GPT-4 Turbo,*
 - *Prompted to respond to this specific evidence while trying to reduce participants' belief in the conspiracy*
 - *Control condition: discuss an unrelated topic*
- Results:
 - *Reduced participants' belief in their chosen conspiracy theory by 20% on average.*

USING AI TO DEBUNK

Exercise: Check out some conversations:

<https://tinyurl.com/5n6kruc7>



STRATEGY: LEGISLATION AND REGULATION

California Passes Election ‘Deepfake’ Laws, Forcing Social Media Companies to Take Action

The state joins dozens of others in regulating the A.I. fakery in ways that could impact this year’s presidential race.

Sept. 17, 2024

Congress passes bill to fight deepfake nudes, revenge porn



A poster is displayed next to Sen. Ted Cruz (R-Texas) as he unveils the Take It Down Act to protect victims against nonconsensual intimate image abuse in June 2024 on Capitol Hill. (Andrew Harnik/Getty Images)



By [Will Oremus](#)

The U.S. House of Representatives on Monday voted overwhelmingly to pass a bill aimed at cracking down on the posting of sexual images and videos of people online without their consent, including AI-generated “deepfake” nudes of real people.

HOW ELSE TO RESPOND TO Mis/DISINFORMATION?