ADDRESSING AI-DRIVEN MISINFORMATION:

Detection, Spread, and Mitigation in Science and Politics

Tutorial @ IC2S2 2025

21 July, 2025

Northwestern University

UNIVERSITY OF MARYLAND

KØBENHAVNS UNIVERSITET

Welcome!

Miriam Schirmer

Northwestern University

Julia Mendelsohn

University of Chicago

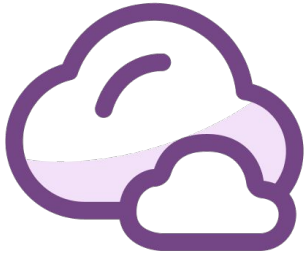
Dustin Wright

University of Copenhagen

Ágnes Horvát

Northwestern University

Do not edit
How to change the design



What's your background?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Do not edit
How to change the design

How confident are you in coding?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Agenda

9:00 - 9:55	Introduction
10:00 - 10:30	Hands-on: NLP for Fact Checking I
Coffee Break	
11:00 - 11:40	Hands-on: NLP for Fact Checking II
11:45 - 12:15	Mitigation Strategies
12:15 - 12:30	Q&A

We will have 5 minute breaks between each section.

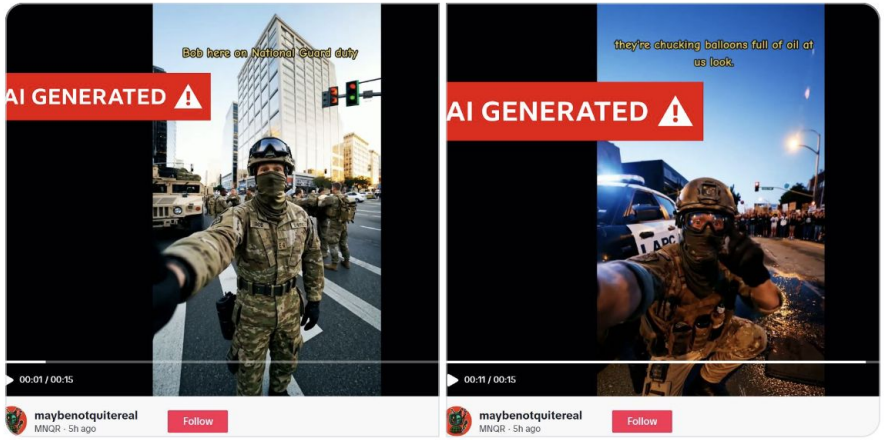
Link to tutorial website and materials

QR code for Google Drive folder with materials

Shayan Sardarizadeh
@shayan86.bsky.social

An AI-generated video of a National Guard member filming himself on duty in LA and provoking the protesters is being shared on TikTok.

The video is fake. The odd badges on his uniform, traffic lights and buildings, and "LAPC" on the police car all give it away.



June 9, 2025 at 1:20 PM

Fake Images and Conspiracy Theories Swirl Around L.A. Protests

Disinformation spreading on social media platforms has stoked an already tense situation.

Listen to this article · 7:03 min Learn more

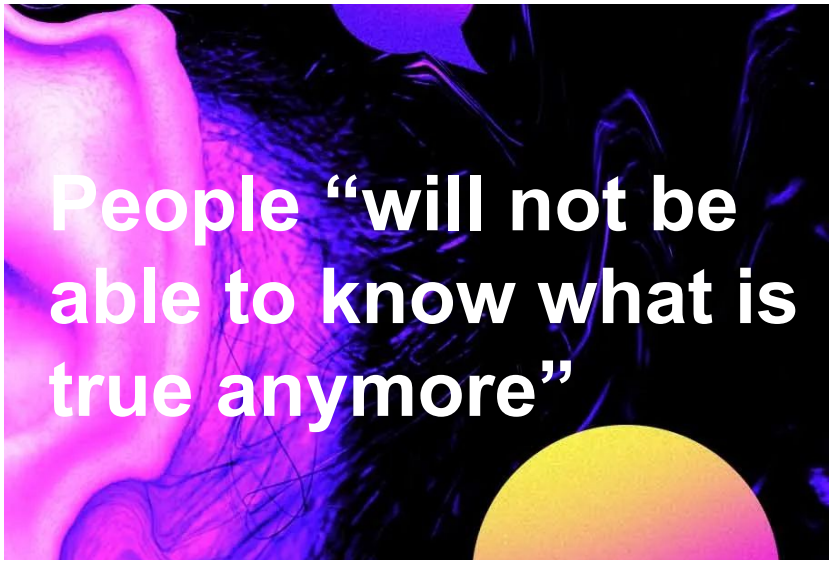
Share full article 55

In downtown Los Angeles on Sunday, protesters faced off with law enforcement officers. Disinformation about the events has circulated online. Mark Abramson for The New York Times

By Steven Lee Myers

June 10, 2025, 5:02 a.m. ET

https://www.axios.com/2023/02/21/chatbots-misinformation-nightmare-chatgpt-ai



https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html?smid=nytcore-ios-share&referringSource=highlightShare

COMMENTARY

Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown

https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/


Intro to
Misinformation in
Science and Politics
in Times of AI

Explosion of misinformation research since 2016

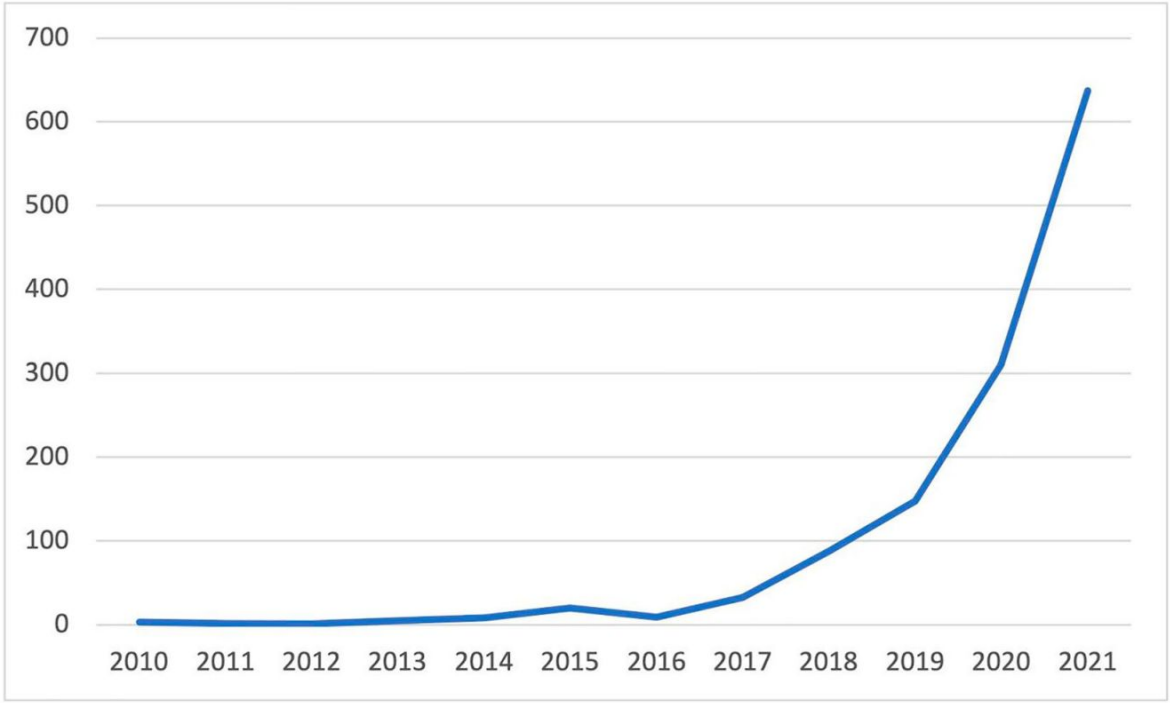


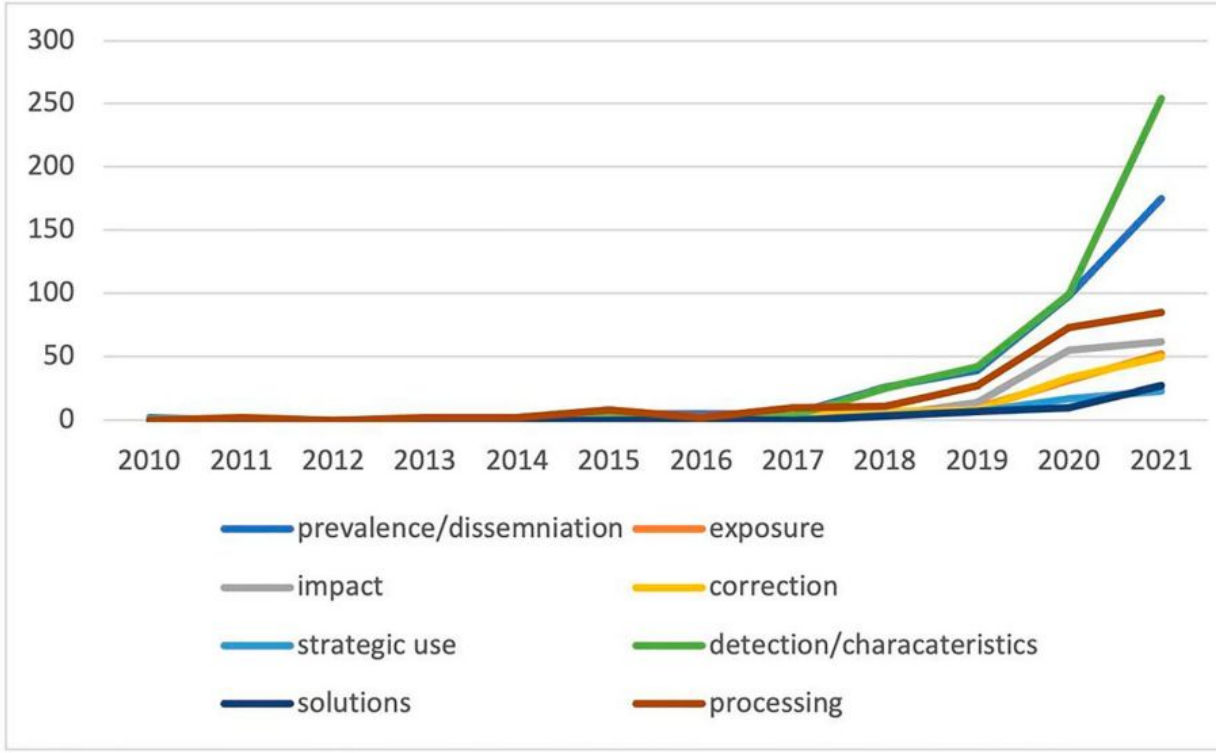
Figure 1. Number of publications across time.

Based on keywords: *misinformation, disinformation, fake news.*

And this figure only covers peer-reviewed journal articles, not including all the work published in CS conferences!

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association, 48(2), 139-166.

Explosion of misinformation research since 2016



Based on keywords: *misinformation, disinformation, fake news.*

And this figure only covers peer-reviewed journal articles, not including all the work published in CS conferences!

Figure 2. Evolution of empirical themes across time.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association, 48(2), 139-166.

Some key concepts I

- **Information Disorder**: production and spread of false, misleading, or harmful information [Wardle & Derakshan, 2017].
  - Encompasses mis-, dis-, mal-information, but some use “misinformation” as the umbrella term.
- **Misinformation**: False information, but not intended to mislead or harm
  - Variation if misinformation must be unintentional, or if intention cannot be determined
- **Disinformation**: False information deliberately created or spread to harm
- **Malinformation**: Information based in reality used to cause harm

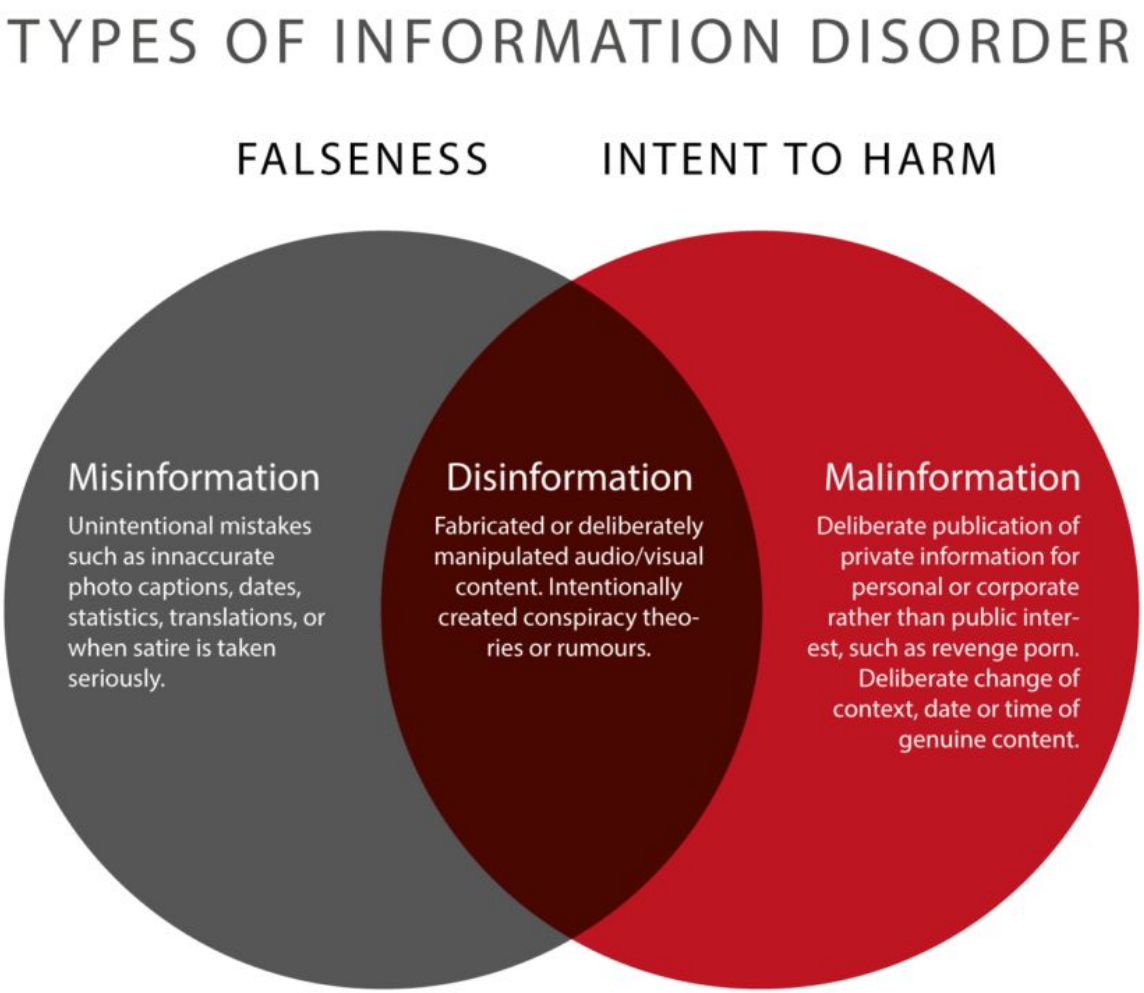
Some key concepts II

- **Fake News**: Disinformation that uses pseudojournalistic style
  - But the term has become a weaponized label to attack professional news media

- **Rumors:** unverified information statements that circulate to help people make sense of ambiguity and threat [DiFonzo & Bordia, 2007]

- **Propaganda**: the systematic dissemination of information, especially in a biased or misleading way, in order to promote a political cause or point of view [OED]

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

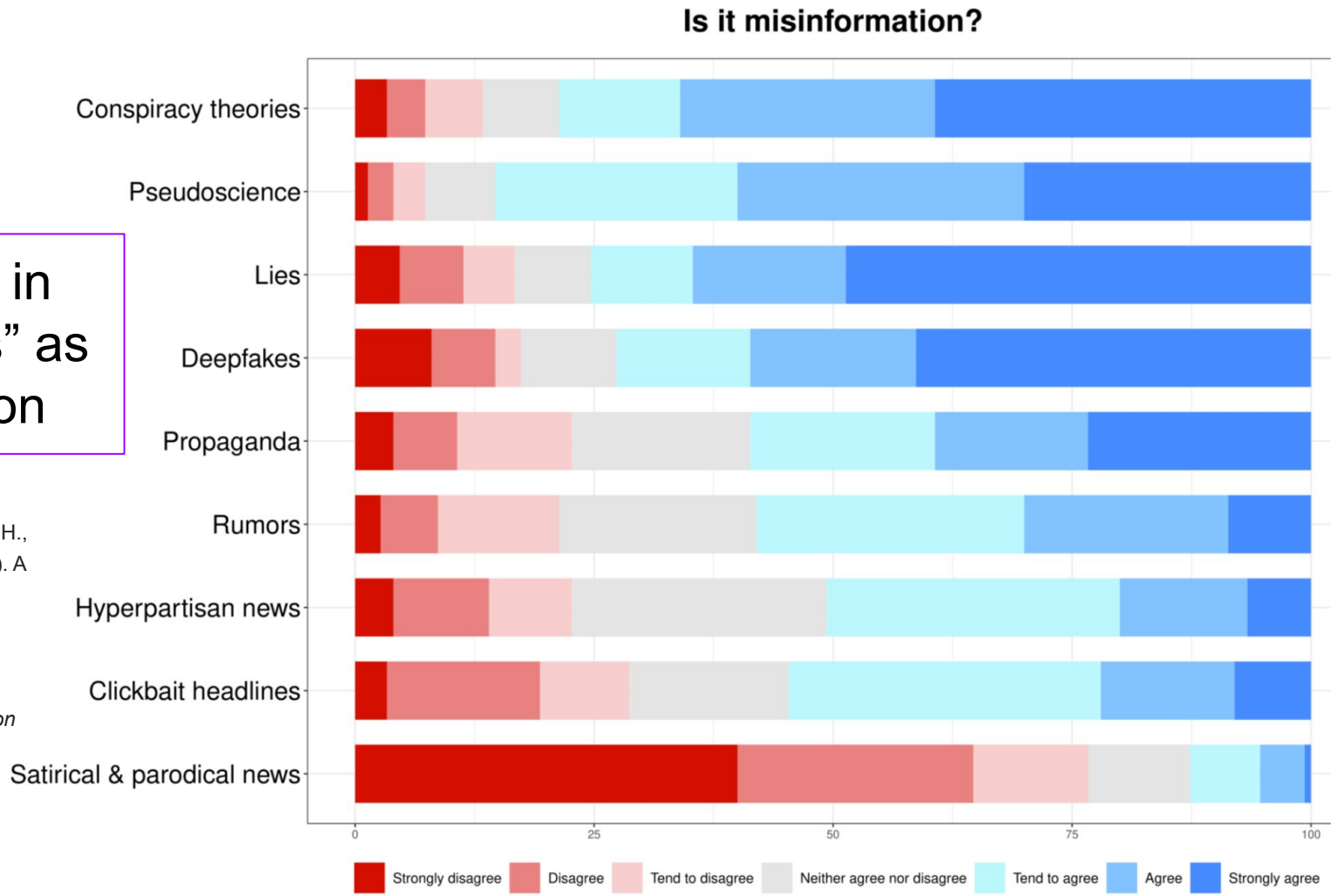


7 common forms of information disorder

- ***Satire or parody****: no intention to cause harm but has potential to fool
- **False connections**: when headlines, visuals, captions don't support the content
- **Misleading content**: misleading use of information to frame an issue or individual
- **False context**: genuine content shared with false context (e.g old images)
- **Imposter content**: genuine sources are impersonated
- **Manipulated content**: genuine information or imagery is manipulated to deceive
- **Fabricated content**: new content is 100% false, designed to deceive and do harm

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Experts vary in what “counts” as misinformation



Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, *4*(4), 1-34.

Why do people believe misinformation?

[Altay et al., 2023; Pantazi et al., 2021; Ecker et al., 2022]

- Identity-based biases
  - Partisanship, social identity, confirmation bias, motivated reasoning
- Generic cognitive biases
  - Repeated exposure (illusory truth), lack of cognitive reflection, inattention
- Attitudinal predispositions
  - Institutional trust, conspiracy thinking, etc.
- Message characteristics
  - Emotional appeals to fear and outrage, source cues
- Structural factors
  - Digital media literacy, education, access to reliable news

Misinformation has serious consequences

- Corrupt democratic processes [Broda & Strömbäck, 2024]
  - Democratic deliberation requires shared set of facts.
  - No shared facts threatens democratic processes' legitimacy (e.g. #Stopthesteal)
  - Concerns about disinformation affecting election outcomes
- Misperceptions lead to harassment, violence, and death
  - Terrorists fueled by belief in Great Replacement conspiracy theory have killed hundreds of people since 2018, primarily ethnic minorities.
  - Harassment of Haitians in wake of hoax about Springfield Ohio
- Economic effects of distrust
  - E.g. stock prices, consumer behavior [Kapantai et al., 2020]

Shayan Sardarizadeh
@shayan86.bsky.social

This video has racked up millions of views and been shared by Texas Senator Ted Cruz and conspiracy theorist Alex Jones.

While there's very real footage of LA protesters setting multiple vehicles on fire on Sunday, this particular video is from the George Floyd protests in 2020.

June 9, 2025 at 1:19 PM

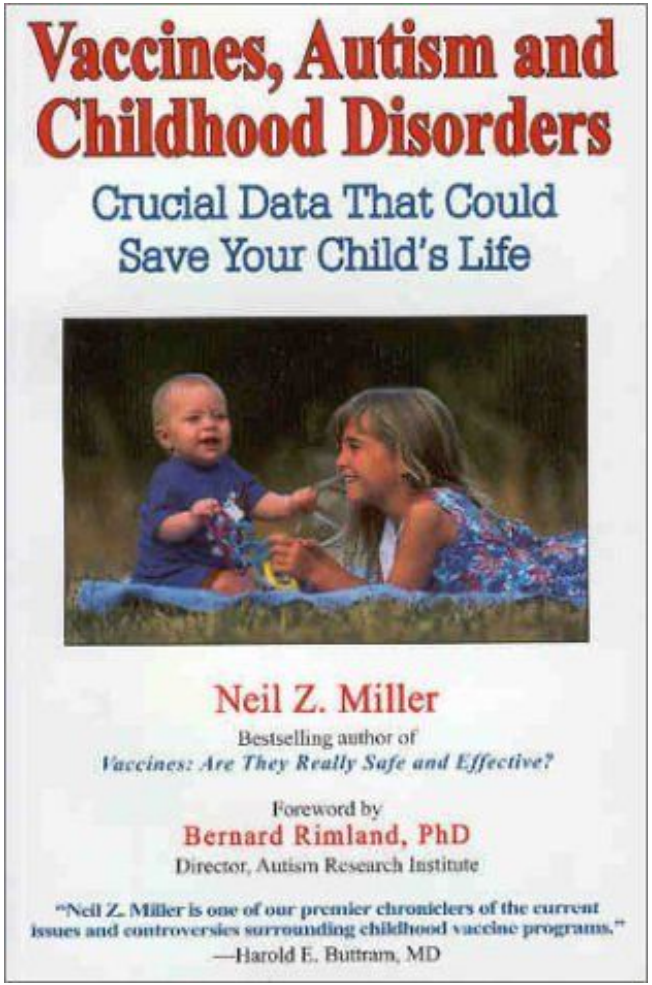
JD Vance
@JDVance

Follow

Months ago, I raised the issue of Haitian illegal immigrants draining social services and generally causing chaos all over Springfield, Ohio.

Reports now show that people have had their pets abducted and eaten by people who shouldn't be in this country. Where is our border czar?

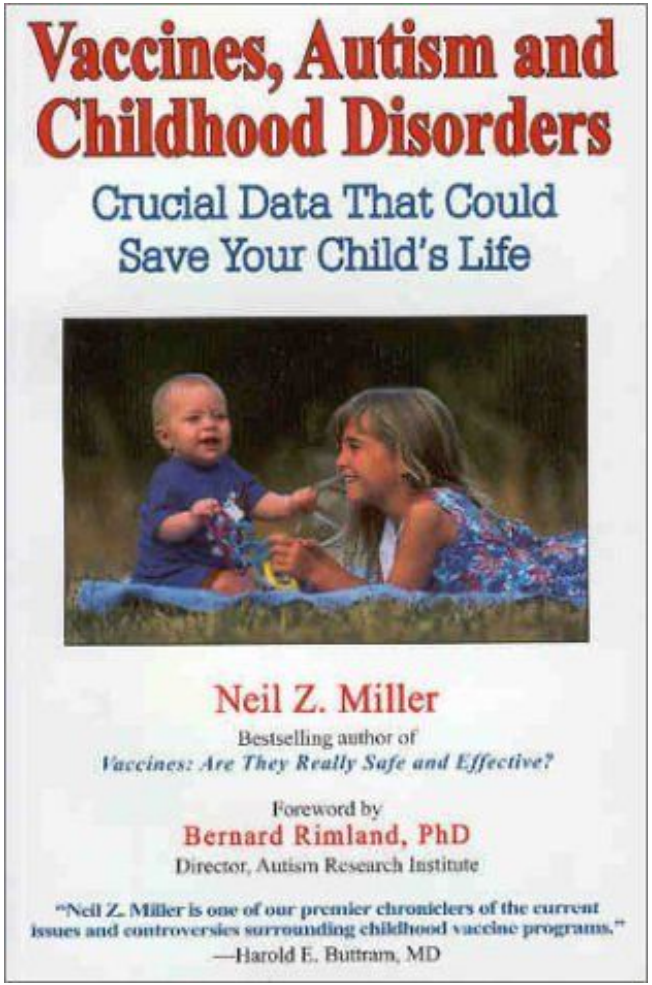
Vaccines and Autism



Vaccines and Autism

- Claim started with a 1998 paper by Andrew Wakefield in *The Lancet*
- Study based on 12 children, later **retracted** due to ethical violations and flawed data

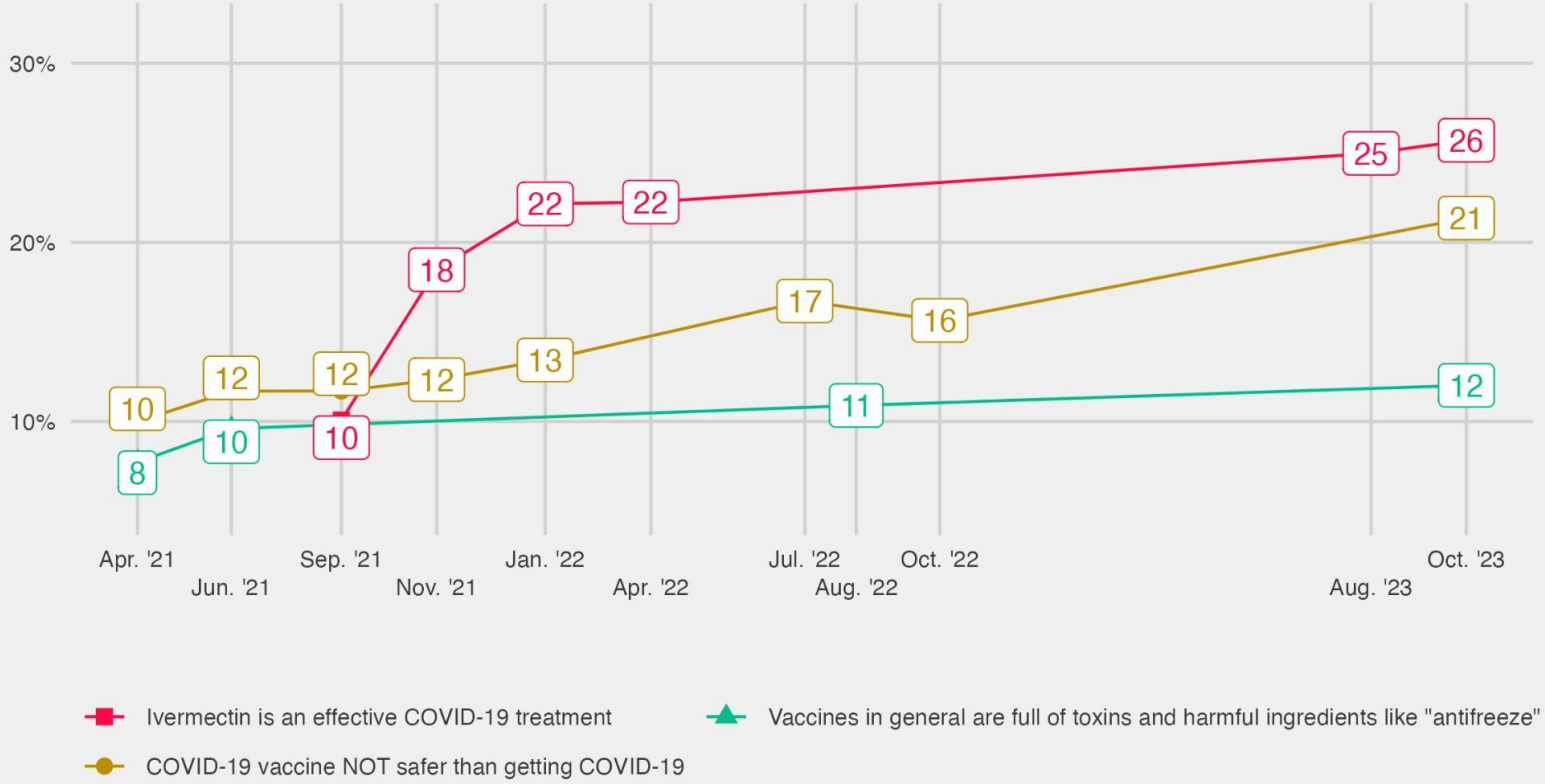
- Despite being debunked, the myth spread widely through media and celebrity endorsements and led to vaccine hesitancy and public distrust in science



Bergstrom, C. T., & West, J. D. (2021). *Calling bullshit: The art of skepticism in a data-driven world*. Random House Trade Paperbacks.

Increasing Belief in Vaccine Misinformation

(% of respondents holding science-inconsistent views)



Source: ASAPH Survey, April 2021 - October 2023
Note: Combined subcategories may not add to totals in topline and text due to rounding.
©2023 Annenberg Public Policy Center

Science Misinformation

Spread of **false or misleading information** about scientific topics, such as the effect of vaccinations (Bergstrom & West, 2021; Lewandowsky & van der Linden, 2017).

This includes **misinterpretations**, **flawed studies**, and **deliberate disinformation.**

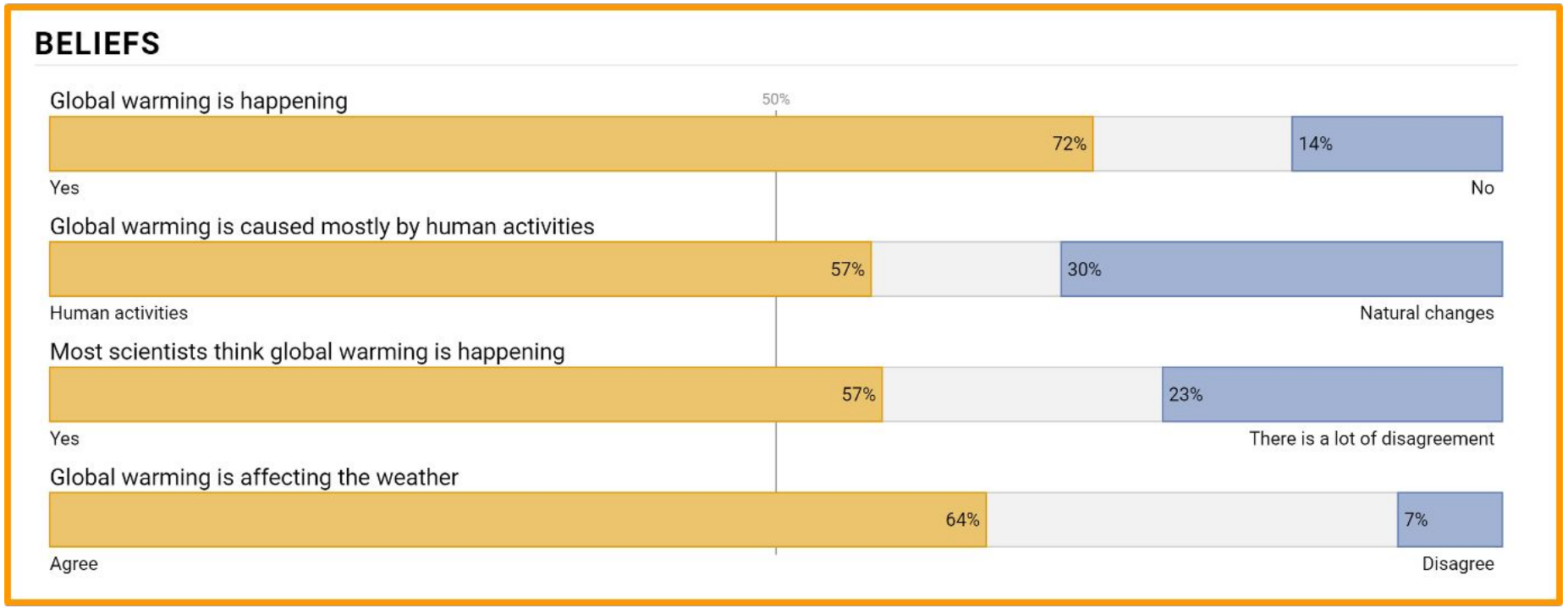
Can result from:

- Methodological flaws or exaggerated findings
- Misrepresentation by media (e.g., cherry-picked data)

How Science Misinformation Changes Beliefs


We the people
NEW DEAL
THERE IS NO
PLANET B

Beliefs about Global Warming



(Yale Climate Opinion Maps, 2021)

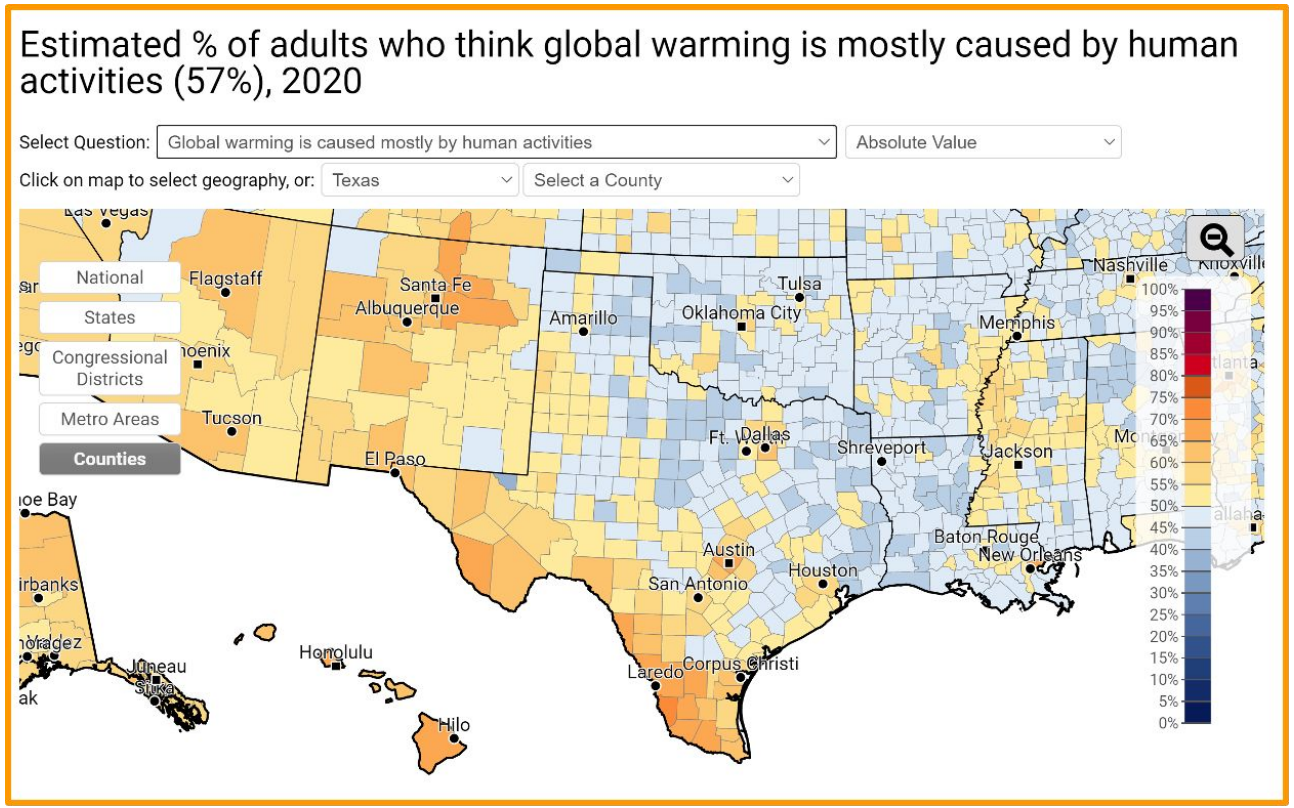
WHETHER THE PLANET IS WARMING DEPENDS ON ITS FRAMING: "GLOBAL WARMING" VS. "CLIMATE CHANGE"?

Table 2. Distribution of Existence Beliefs by Political Self-identification and Question Wording (GW = "global warming"; CC = "climate change")

	Overall		Republicans		Democrats		Independents		Others	
Reported Existence Belief	GW	CC	GW	CC	GW	CC	GW	CC	GW	CC
1 = Definitely has not been happening	6.6%	3.9%	12.7%	5.1%	1.9%	1.3%	6.6%	5.4%	4.7%	5.7%
2 = Probably has not been happening	8.4%	6.8%	18.2%	14.3%	2.6%	1.0%	6.2%	5.4%	3.7%	4.8%
3 = Leaning has not been happening	8.2%	7.6%	14.1%	11.4%	2.9%	5.2%	9.2%	6.6%	6.5%	5.7%
4 = Unsure	9.0%	7.6%	11.0%	8.9%	5.7%	6.0%	8.5%	8.7%	16.8%	6.7%
5 = Leaning has been happening	13.9%	15.7%	14.1%	17.0%	12.1%	13.6%	15.1%	15.7%	17.8%	18.1%
6 = Probably has been happening	25.6%	27.8%	20.2%	28.9%	29.2%	27.0%	27.9%	28.5%	23.4%	25.7%
7 = Definitely has been happening	28.2%	30.5%	9.7%	14.3%	45.6%	45.8%	26.5%	29.8%	27.1%	33.3%
% High Belief (≥ 5)	67.7%	74.0%	44.0%	60.2%	86.9%	86.4%	69.5%	74.0%	68.3%	77.1%
Mean Belief	5.05	5.30	3.95	4.62	5.94	5.94	5.09	5.29	5.18	5.37
N	1162	1099	362	370	421	382	272	242	107	105

Republicans are more likely to accept "global warming" if it's framed as "climate change".

WHETHER THE PLANET IS WARMING DEPENDS ON WHERE YOU LIVE



People who live near the coastline tend to accept global warming (even in a Republican State, such as Texas).

But be aware of misperceptions about misinformation

But be aware of misperceptions about misinformation

From Altay et al. (2023) survey of misinformation experts:

- Agreement that **exposure to opposing viewpoints is higher online than offline**
- **Lack of consensus** for whether **misinformation** and **belief in conspiracy theories have increased** in the past ten years
- **Lack of consensus** on the impact of misinformation on the outcome of the **2016 U.S. election**
  - 54% of psychologists agreeing that misinformation played a decisive role
  - 73% of political scientists disagree

Budak et al. (2024) review of behavioral science research on online misinformation identify a

- pattern of **low exposure** to false and inflammatory content
- that is **concentrated among a narrow fringe**
- with **strong motivations to seek** out such information.

Budak, C., Nyhan, B., Rothschild, D.M. *et al.* Misunderstanding the harms of online misinformation. *Nature* 630, 45–53 (2024).

Three misperceptions about misinformation

Budak et al., 2024

Misperception 1:

Average exposure to misinformation is high and growing; most of the population is frequently exposed to misinformation

Three misperceptions about misinformation

Budak et al., 2024

Misperception 1:

Average exposure to misinformation is high and growing; most of the population is frequently exposed to misinformation

What the research says: Misinformation exposure is a small portion of people's information diets and concentrated among a small minority of users.

Three misperceptions about misinformation Budak et al., 2024

Misperception 1:

Average exposure to misinformation is high and growing; most of the population is frequently exposed to misinformation

What the research says: Misinformation exposure is a small portion of people's information diets and concentrated among a small minority of users.

Why this misperception is harmful: Focus on average exposure diverts attention away from the narrow fringe, where consumption and risk of real-world harm are highest.

Three misperceptions about misinformation Budak et al., 2024

Misperception 2:

Misinformation exposure is primarily driven by **social media platform algorithms**

Three misperceptions about misinformation

Budak et al., 2024

Misperception 2: Misinformation exposure is primarily driven by social media platforms' algorithms

What the research says:

(1) No evidence for large-scale algorithmic effects on public attitudes & behaviors
(2) There is *consumer demand* for misinformation: some deliberately seek it out

Three misperceptions about misinformation Budak et al., 2024

Misperception 2: Misinformation exposure is primarily driven by social media platforms' algorithms

What the research says:

(1) No evidence for large-scale algorithmic effects on public attitudes & behaviors
(2) There is *consumer demand* for misinformation: some deliberately seek it out

Why this misperception is harmful: Diverts attention *away* from consumer demand, role of media and political elites, and how platform design/features enable distribution of misinformation.

Three misperceptions about misinformation

Budak et al., 2024

Misperception 3:

Social media (and online misinformation) causes undesirable psychological, behavioral, and societal outcomes.

Three misperceptions about misinformation

Budak et al., 2024

Misperception 3: Social media (and online misinformation) causes undesirable psychological, behavioral, and societal outcomes.

What the research says: Correlation != causation. Little support from existing studies designed to measure causal effects. Making causal claims requires unrealistic counterfactuals, or ability to systematically vary platform features.

Three misperceptions about misinformation

Budak et al., 2024

Misperception 3: Social media (and online misinformation) causes undesirable psychological, behavioral, and societal outcomes

What the research says: Correlation != causation. There's little support from existing studies designed to measure causal effects. Making causal claims requires unrealistic counterfactuals, or ability to systematically vary platform features.

Why this misperception is harmful:

(1) Neglects indirect harms. For example, misperceptions about misinformation on social media could reduce public trust (which is also why we shouldn't exaggerate its presence)
(2) Assumption undermines need to do more research on causal effects of platform features and other interventions.

DeepSeek's Answers Include Chinese Propaganda, Researchers Say

Since the Chinese company's chatbot surged in popularity, researchers have documented how its answers reflect China's view of the world. Some of its responses amplify propaganda Beijing uses to discredit critics.

Listen to this article · 5:46 min Learn more

Share full article

144

How the Indian Media Amplified Falsehoods in the Drumbeat of War

During the conflict between India and Pakistan, even some long-trusted outlets reported unverified information and fabricated stories.

Listen to this article · 5:55 min Learn more

Share full article

A recent live telecast in Mumbai of Prime Minister Narendra Modi describing military operations against Pakistan. Rajanish Kakade/Associated Press

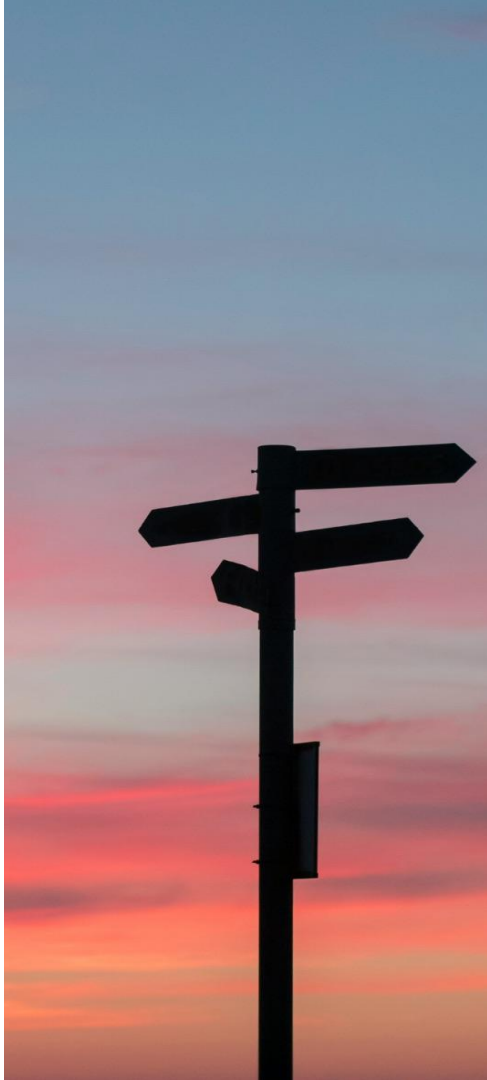
By Anupreeta Das and Pragati K.B.
Reporting from New Delhi

May 17, 2025

AI and Misinformation

AI is changing the landscape of misinformation, however, the extent and direction is unknown:

~ 1,300 studies on how AI is impacting **science misinformation** (Schirmer et al., in prep)

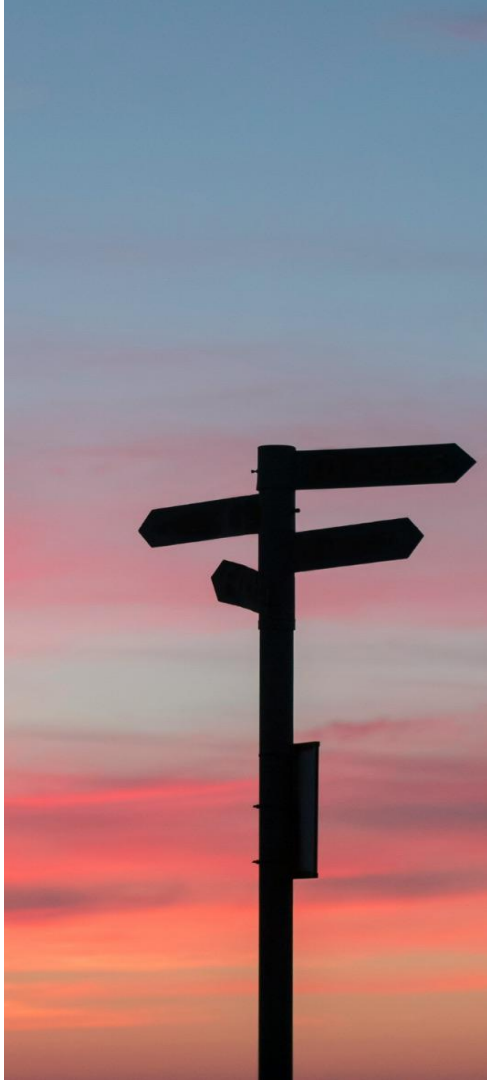


AI and Misinformation

AI is changing the landscape of misinformation, however, the extent and direction is unknown:

~ 1,300 studies on how AI is impacting science misinformation (Schirmer et al., in prep)

- **AI-generated deepfakes** used to impersonate public figures and spread false narratives (Shoaib et al., 2023; Vaccari & Chadwick, 2020)
- **Recommendation algorithms and personalization** (Pathak et al., 2023; Xu et al., 2023)
- **Mass production** (Barman et al., 2024; De Angelis et al., 2023; Tomassi et al., 2024)
- **Synthetic consensus:** Bots and AI agents simulate public opinion to manipulate perception (Calvo and Garcia, 2024)
- **Multilingual manipulation:** AI spreads misinformation in low-moderation languages (Lenti et al., 2023)

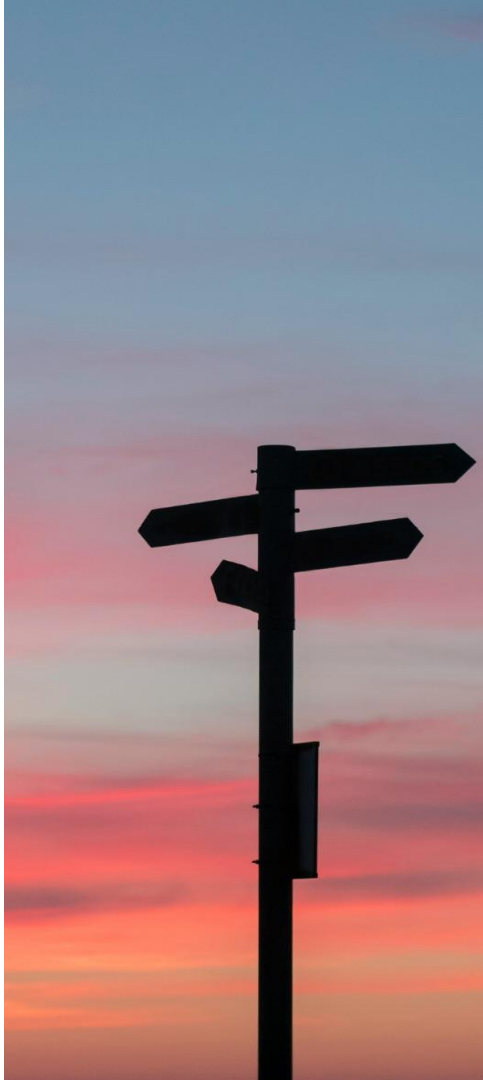


AI and Misinformation

But:

AI also creates opportunities to address misinformation (Chen & Shu, 2024; Saeidnia et al., 2025), e.g., by

- **Enhanced detection** of misinformation on social media platforms (Wang et al., 2025; Yang et al., 2019)
- **Disaster response**: Managing misinformation during disasters, enhancing public trust and community resilience (Imran et al., 2020; Vicari & Komendatova, 2023)
- AI tools for **community based grassroot initiative**s in marginalized communities (Ozawa et al., 2024)
- Improved digital access through **translation tools** (Zaki & Ahmed, 2024)



So what should researchers focus on?

[Altay et al., 2023; Budak et al., 2024, Saeidnia et al., 2025]

Misinformation exposure, effects, and interventions are situated among the narrow fringe.

How we can reduce audience demand for misinformation and its amplification by media and political elites?

- One area: Online media environments in **non-Western and/or authoritarian contexts, or disadvantaged communities**
  - Average exposure & effects of online misinformation might be higher where there's less exposure to other news sources, lower media freedom, low trust in media

So what should researchers focus on?

[Altay et al., 2023; Budak et al., 2024, Saeidnia et al., 2025]

Misinformation exposure, effects, and interventions among the narrow fringe.

How we can reduce audience demand for misinformation and its amplification by media and political elites?

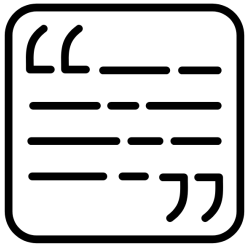
- One area: Online media environments in **non-Western and/or authoritarian contexts, or disadvantaged communities**
    - Average exposure & effects of online misinformation might be higher where there's less exposure to other news sources, lower media freedom, low trust in media
    -
- **Specifically for AI**, mapping the AI landscape has to be expanded for certain domains (e.g., science misinformation)
- Understanding platform-specific dynamics (e.g., TikTok vs. Reddit vs. YouTube) and how AI-generated content spreads differently, specifically regarding causal effects of platform features and direct impact on individuals


Take
a
Break


Hands-On:
NLP for Fact Checking

Detection of Misinformation

Fact checking is time consuming



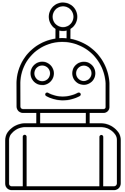
Claim to check

Find Evidence

Determine Veracity

Write a Report

How can AI help fact checkers?



Detection of misinformation

In this practical section, we will cover a set of NLP methods intended to help with the four major steps of fact checking.

We will cover a range of methods, from older classification approaches to the use of LLMs and their evaluation.

Future Work in the Era of LLMs

- Transcription of audio from debates, interviews, etc. (Augenstein et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence 2024.)
- Organizing and normalizing claims into fine-grained frames (Sundriyal et al. 2023. From chaos to clarity: Claim normalization to empower fact-checking. EMNLP Findings 2023)
- Summarizing evidence to speed up claim verification (Kevin Roitero et al. 2025. Efficiency and Effectiveness of LLM-Based Summarization of Evidence in Crowdsourced Fact-Checking. SIGIR 2025)
- Finding previously fact-checked claims (Shaar et al. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. ACL 2020.)
- For more info, see: Augenstein et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking


Take
a
Break

Mitigating Misinformation

Do not edit
How to change the design

Do you know of any effective strategies to counter misinformation?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido


Human-Centered Solutions to
Address Misinformation

Human-Centered Approaches

- Support **prebunking** and **inoculation** strategies (Lewandowsky & van der Linden, 2021)

    a. **Debunking**: Correcting misinformation *after* it has been encountered, usually by providing factual information.
    b. **Prebunking**: “Making people aware of potential misinformation *before* it is presented,” often by explaining common manipulation techniques.
    c. **Inoculation**: A psychological strategy that works like a vaccine: exposing people to a weakened form of misinformation or its tactics in advance, along with refutations, to build cognitive resistance to future exposure.

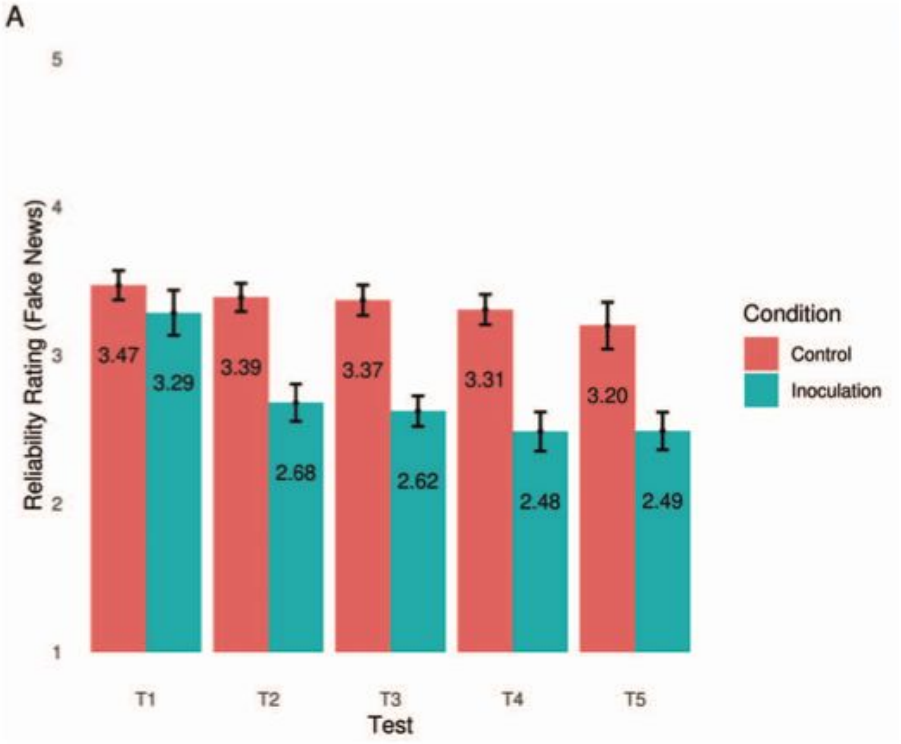
The Effectiveness of Inoculation Strategies (Maertens et al., 2021)

Participants rated **reliability of news headlines** before and after intervention.

Intervention: Randomly assigned to play ***Bad News*** (game that teaches inoculation) or **Tetris** (control).

Follow-up ratings collected **immediately, after 1 week, 5 weeks, and 13 weeks**.

Result:
Participants who played *Bad News* consistently rated fake news as **less reliable** than control group, even **13 weeks later**.



Human-Centered Approaches

- Support pre-bunking and inoculation strategies.
- Build **partnerships between scientists, journalists, and within communities**.

Diaspora Communities Fighting Misinformation (Ozawa et al., 2024)

- Conducted **interviews with 12 leaders** from diaspora civil society organizations in the U.S. (**Latinx and Asian communities**)
- **Community Organization Efforts**:
  - Some organizations already combat misinformation using:
    - **Volunteer teams**
    - **Fact-checking websites**
    - **Social listening tools** (i.e., software designed to monitor online conversations and mentions of a topic across the web)
  - However, they **require support** through:
    - **Targeted media literacy programs**
    - Culturally and **linguistically adapted interventions**

Crash Course Media Literacy Preview

Learning playlist

CrashCourse
16.4M subscribers

Subscribe

7.8K Share Save

Human-Centered Approaches

- Support pre-bunking and inoculation strategies.
- Build partnerships between scientists, journalists, and technologists.
- Educate users on AI-generated misinformation and scientific reasoning.
- Encourage **transparent communication** of scientific uncertainty.

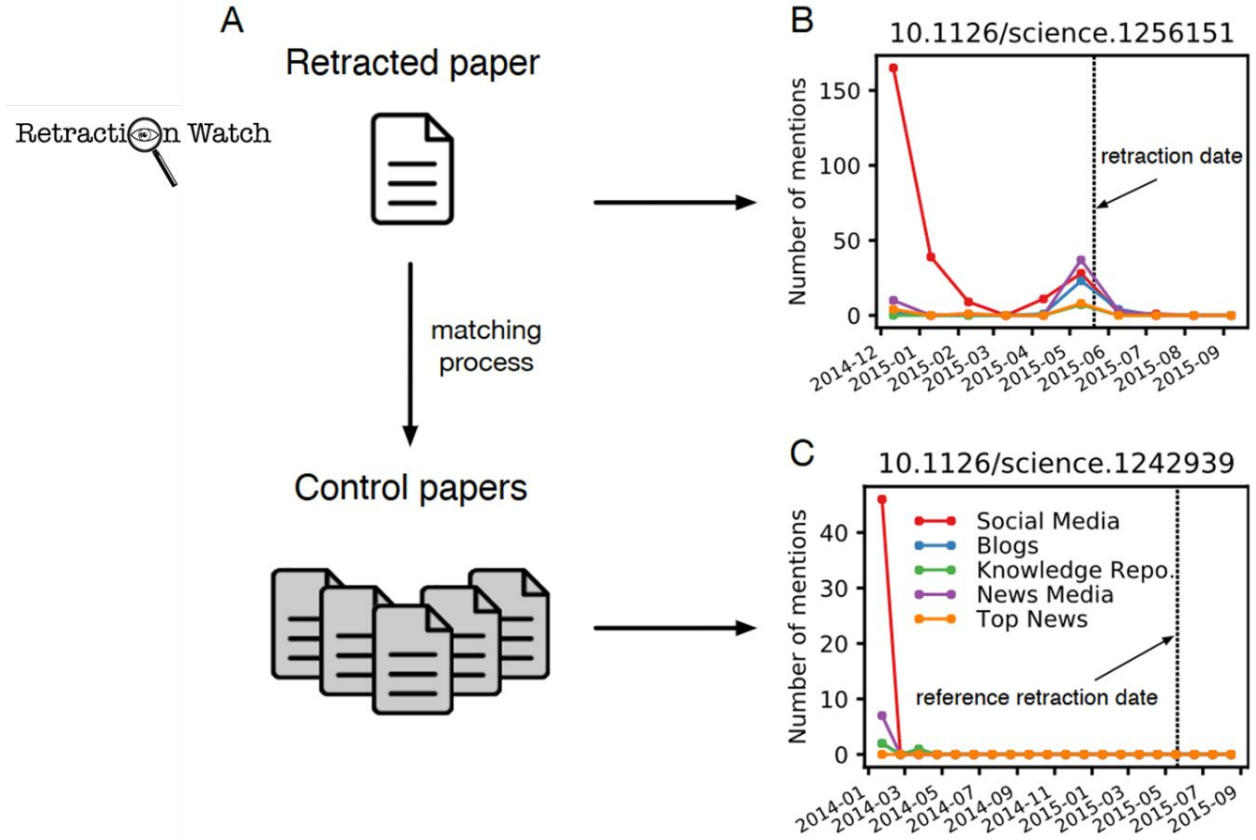
Retraction Watch

The most highly cited retracted papers

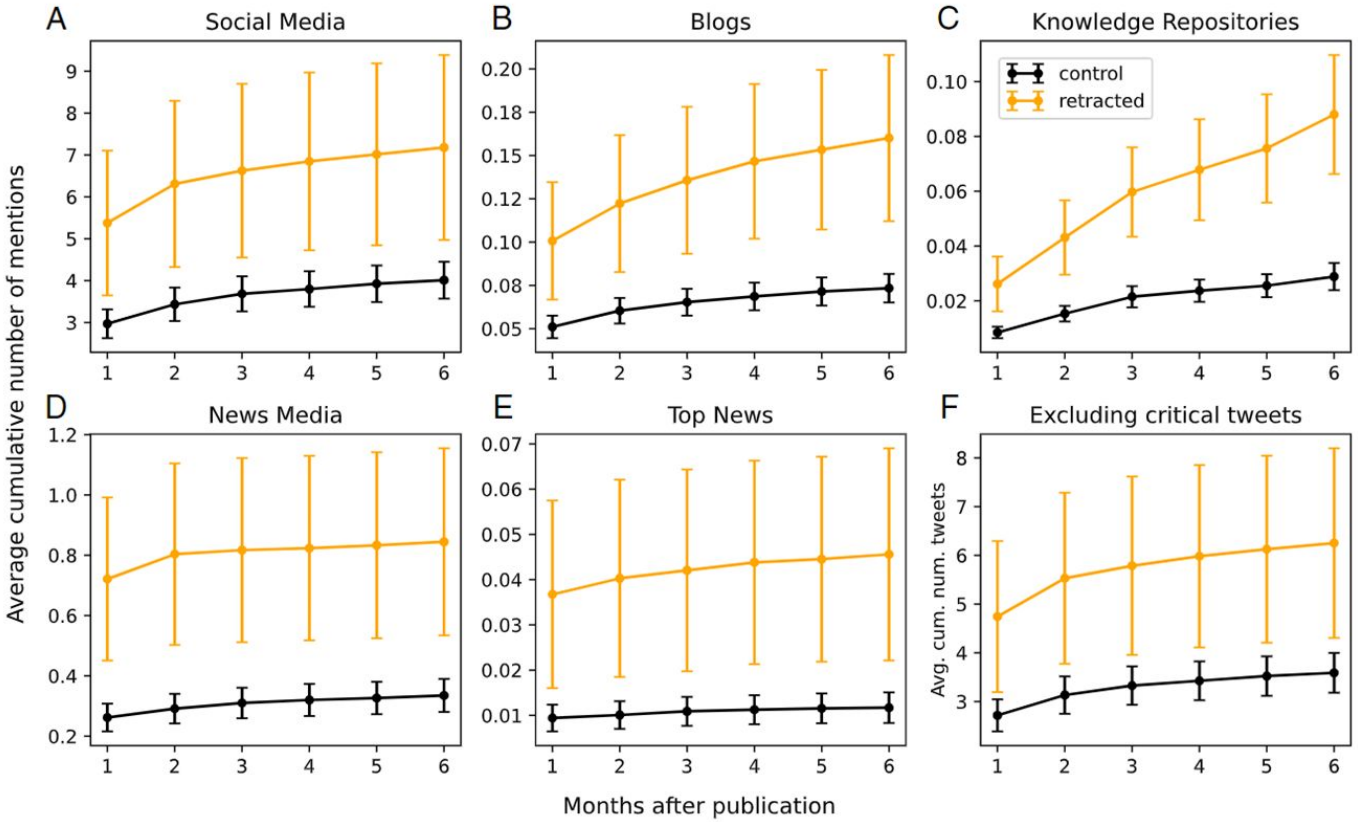
https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/

Article	Year of retraction	Citing Articles before retraction	Citing Articles after retraction	Total cites (journals indexed by Web of Science)
1. Pluripotency of mesenchymal stem cells derived from adult. *Nature*. June 20, 2002. *Y Jiang, BN Jahagirdar, RL Reinhardt, RE Schwartz, CD Keene, XR Ortiz-Gonzalez, M Reyes, T Lenvik, T Lund, M Blackstad, J Du, S Aldrich, A Lisberg, WC Low, DA Largaespada, CM Verfaillie*	2024	4491	21	4512

Retractions in the media (Peng et al., 2022)



Retractions in the media (Peng et al., 2022)

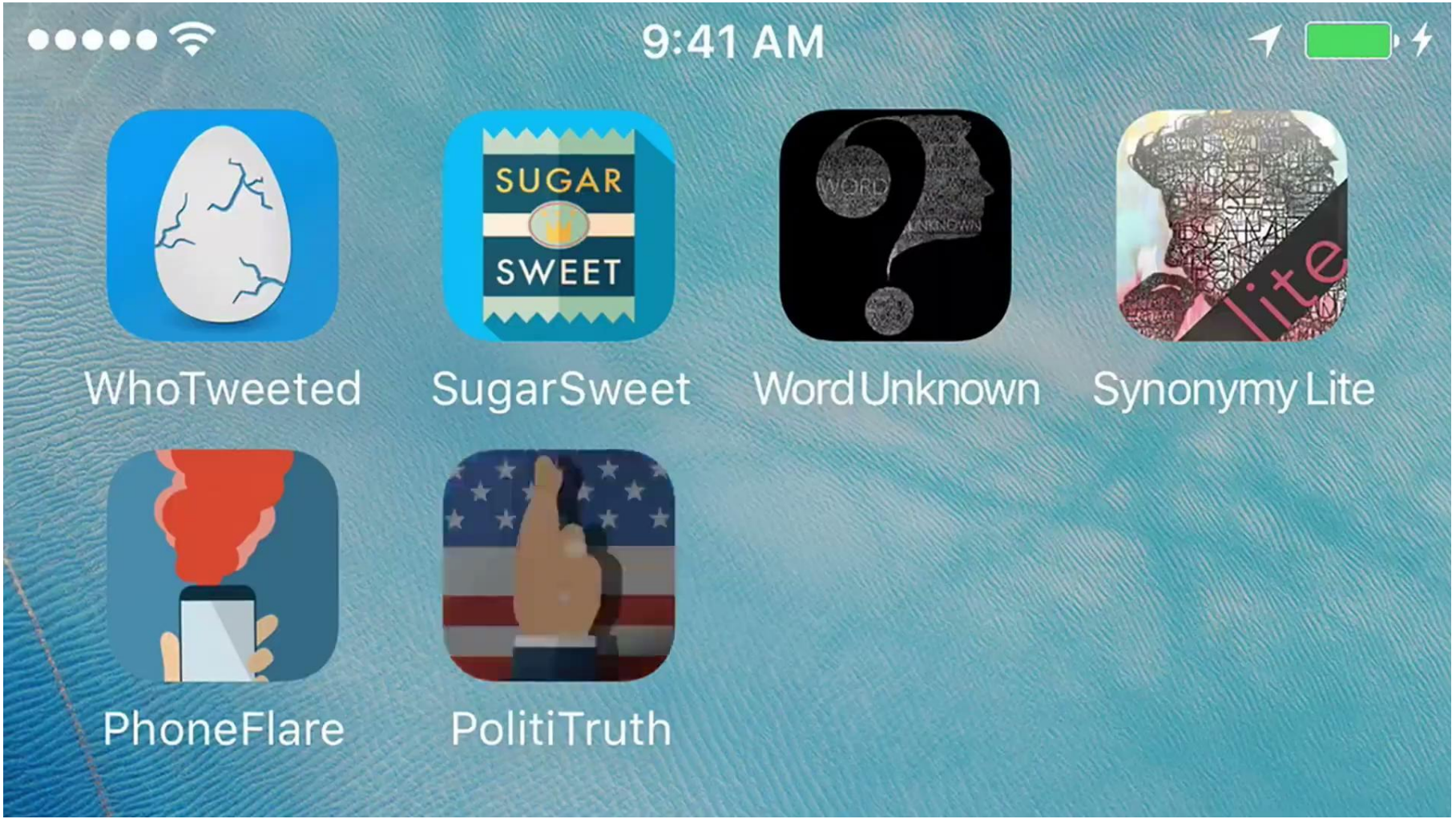


- Untrustworthy research is mentioned more than expected even or curated digital platforms.
- Correcting the public record is difficult if not impossible.


Technical Solutions to
Address Misinformation

Technical Solutions for Addressing Misinformation

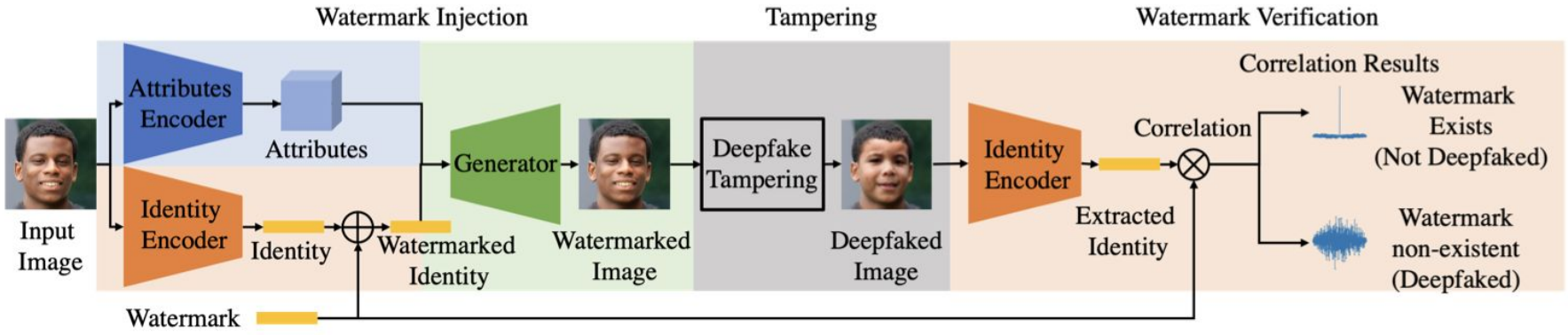
- AI-powered **fact-checking tools**, e.g., ClaimBuster (Hassan et al., 2017) and PoliTruth (Christopher Cinq-Mars Jarvis, 2017)


9:41 AM
WhoTweeted
SUGAR
SWEET
SugarSweet
WORD
UNKNOWN
WordUnknown
lite
Synonymy Lite
PhoneFlare
PolitiTruth

Technical Solutions for Addressing Misinformation

- AI-powered fact-checking tools, e.g., ClaimBuster (Hassan et al., 2017)
- Detection of synthetic content via **watermarking or AI detectors**.

Watermark Injection to Identify Deepfakes



Zhao et al., 2023

Technical Solutions for Addressing Misinformation

- AI-powered fact-checking tools, e.g., ClaimBuster (Hassan et al., 2017)
- Detection of synthetic content via watermarking or AI detectors.
- Educational tools, e.g., getbadnews.com

ROVOKE

INTRIGUE

BAD NEWS

MANIPULATE

From fake ne ɔs! How bad are you? Get as many followers as you can.

INVENT

START

EXAGG SHOUT

WHISPER

FRIGHTEN

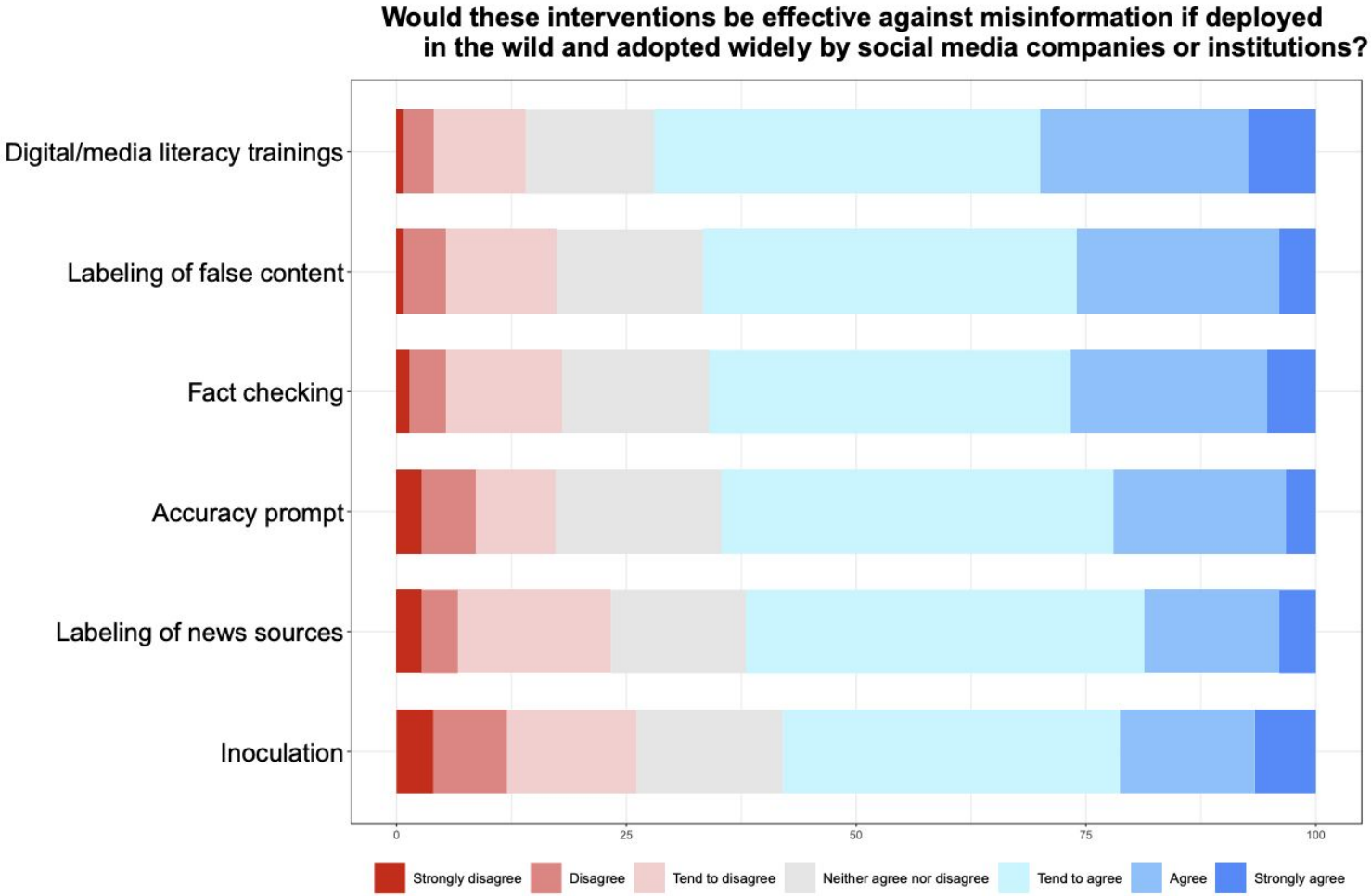
Technical Solutions for Addressing Misinformation

- AI-powered fact-checking tools, e.g., ClaimBuster (Hassan et al., 2017)
- Detection of synthetic content via watermarking or AI detectors.
- Educational tools, e.g., getbadnews.com
- Conversations with LLMs.

Dialogues with AI Reduce Conspiracy Beliefs (Costello et al., 2025)

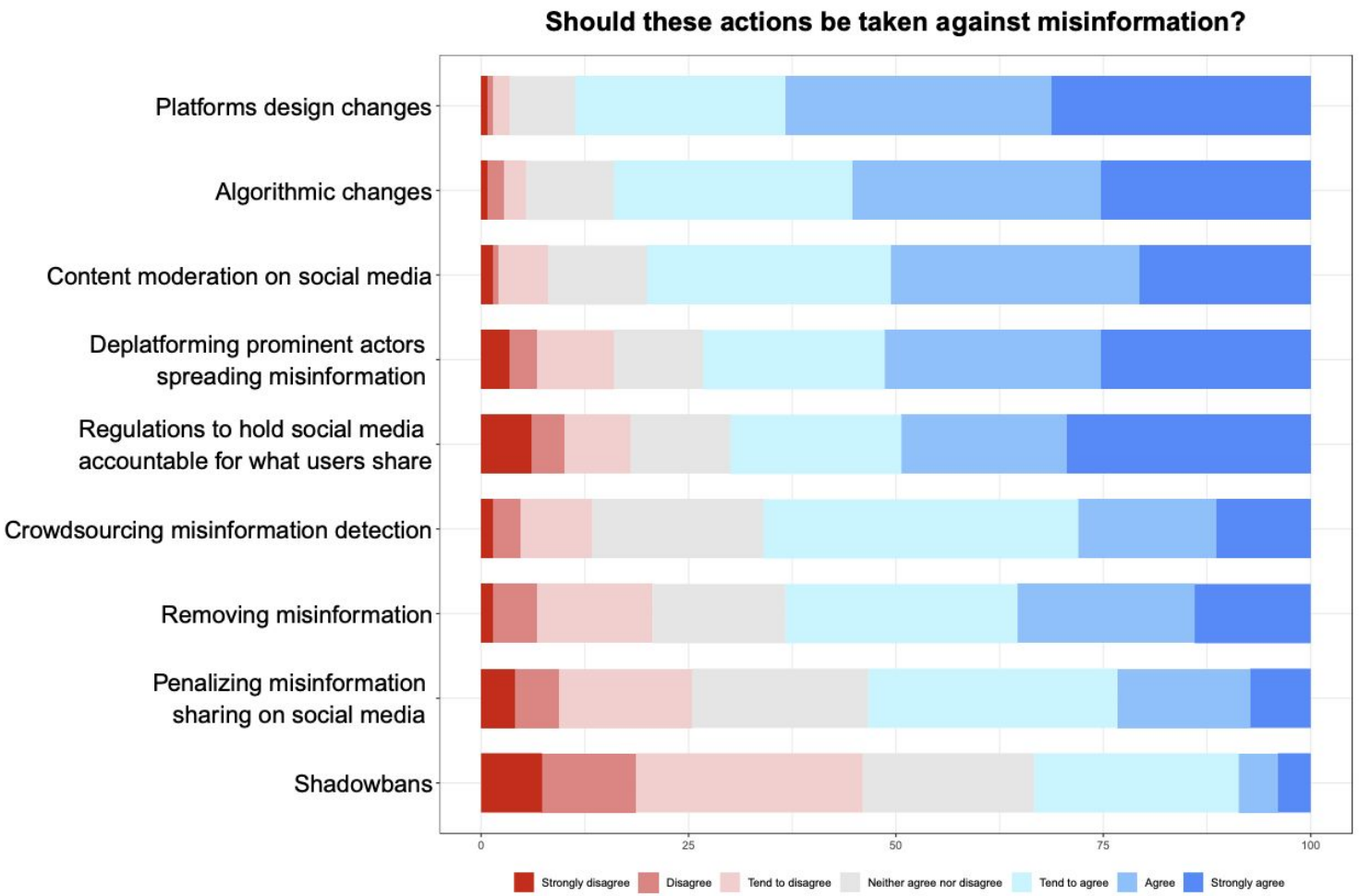
- Researchers engaged **2,190 conspiracy believers** in personalized evidence-based dialogues with **GPT-4 Turbo**.
- When a professional fact-checker evaluated a sample of 128 **claims made by the AI, 99.2% were true**, 0.8% were misleading, and none were false.
- The intervention **reduced conspiracy belief by ~20%**.
- The **effect remained 2 months later**, and generalized across a wide range of conspiracy theories 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 election; and occurred even for participants whose conspiracy **beliefs were deeply entrenched and important to their identities**.
- The **debunking** also spilled over to **reduce beliefs in unrelated conspiracies**, indicating a general decrease in conspiratorial worldview.

What do the experts say?



Altay et al., 2023

Figure 4. Effectiveness of interventions against misinformation. *Stacked bar plot (in percentage) representing participants' responses on the effectiveness of interventions against misinformation.*

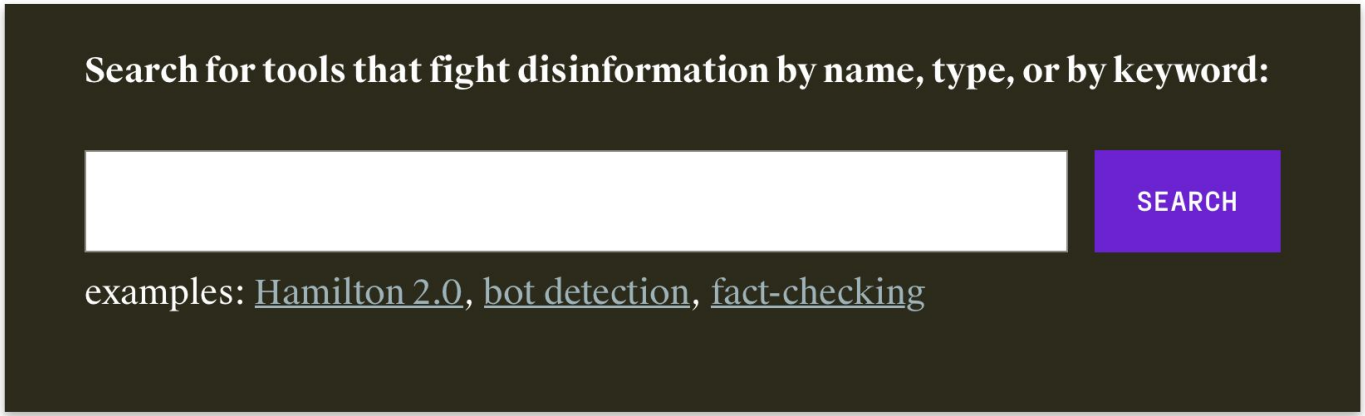


Altay et al., 2023

Figure 5. Effectiveness of system-level actions against misinformation. *Stacked bar plot (in percentage) representing participants' responses on actions that should be taken against misinformation.*

RAND School of Public Policy

Tools That Fight Disinformation Online



https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html

What can we do as researchers?

Develop better detection or spread simulation tools

Collaborate with NGOs, policy makers, and platforms

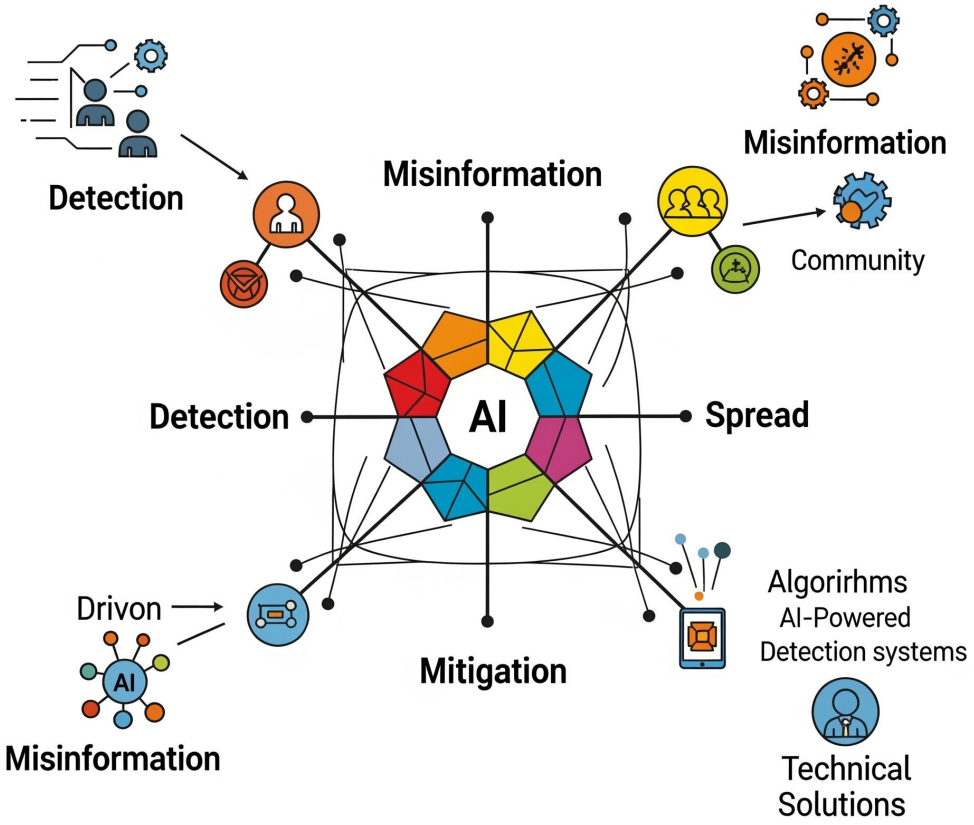
Learn about science communication

Engage with local communities, e.g., through participatory studies

Study human cognition to understand human engagement with misinformation

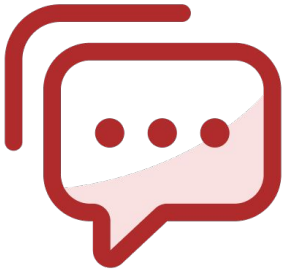
Key Takeaways

- AI impacts misinformation in various, complex ways.
- It is crucial in both spreading and combating false content.
- We covered detection, spread, and mitigation strategies.
- Human and technical solutions are essential for addressing this.



Q&A

Do not edit
How to change the design



Audience Q&A

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Contact Information

Miriam Schirmer

Northwestern University

miriam.schirmer@northwestern.edu

Julia Mendelsohn

University of Chicago

juliame@umd.edu

Dustin Wright

University of Copenhagen

dw@di.ku.dk

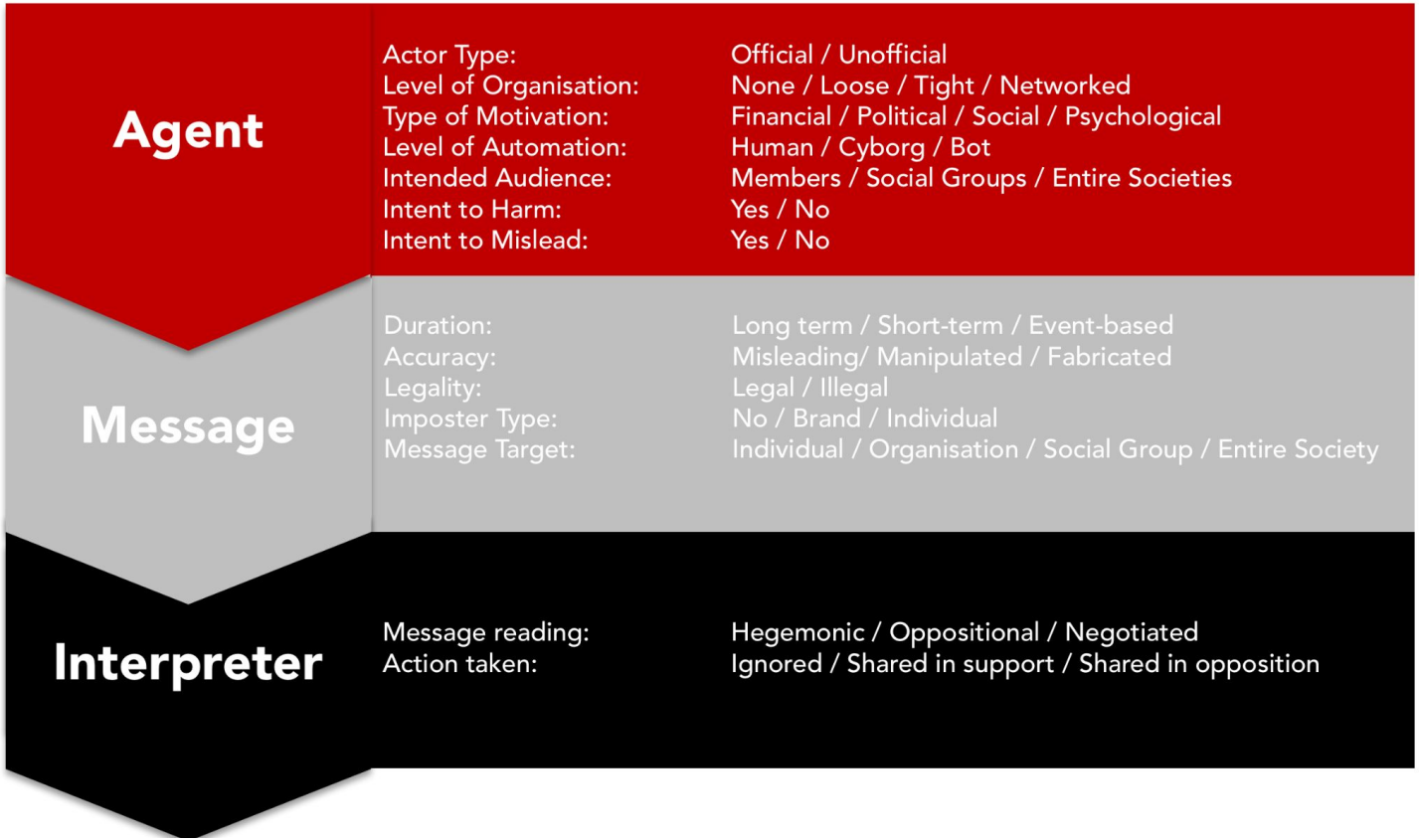
Ágnes Horvát

Northwestern University

a-horvat@northwestern.edu

APPENDIX

Making sense of information disorder



Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

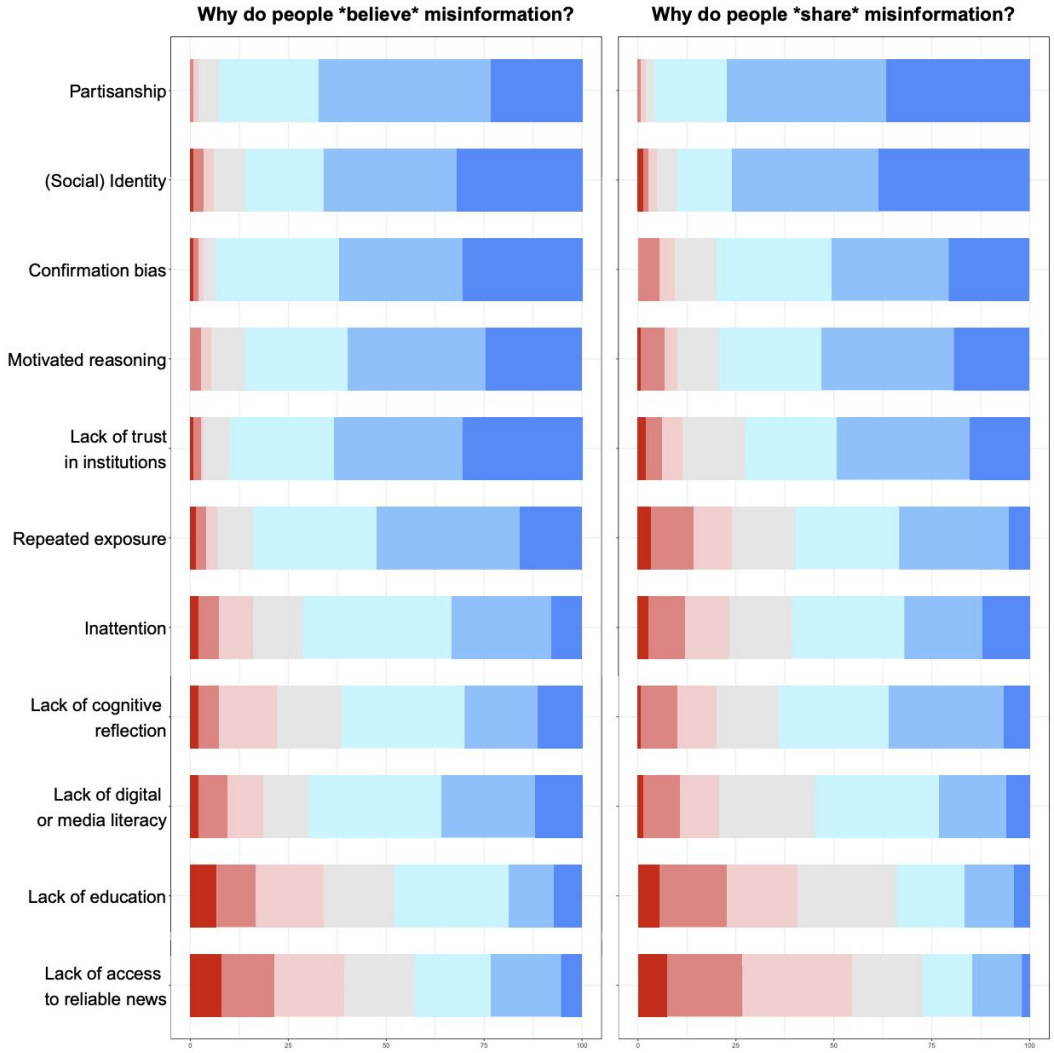
Kapantai et al. - Taxonomy of disinformation

Dimensions/measurement	Motive			Facticity				Verifiability	
	Profit	Ideological	Psychological	Unclear	Mostly true	Mostly false	False	Yes	Not
Clickbait	✓		✓		✓			✓	
Conspiracy Theories		✓	✓			✓			✓
Fabrication				✓			✓		✓
Misleading connection			✓		✓				✓
Hoax			✓				✓		✓
Biased or one-sided		✓				✓		✓	
Imposter			✓			✓		✓	
Pseudoscience	✓		✓		✓				✓
Rumors				✓		✓			✓
Fake Reviews	✓						✓		✓

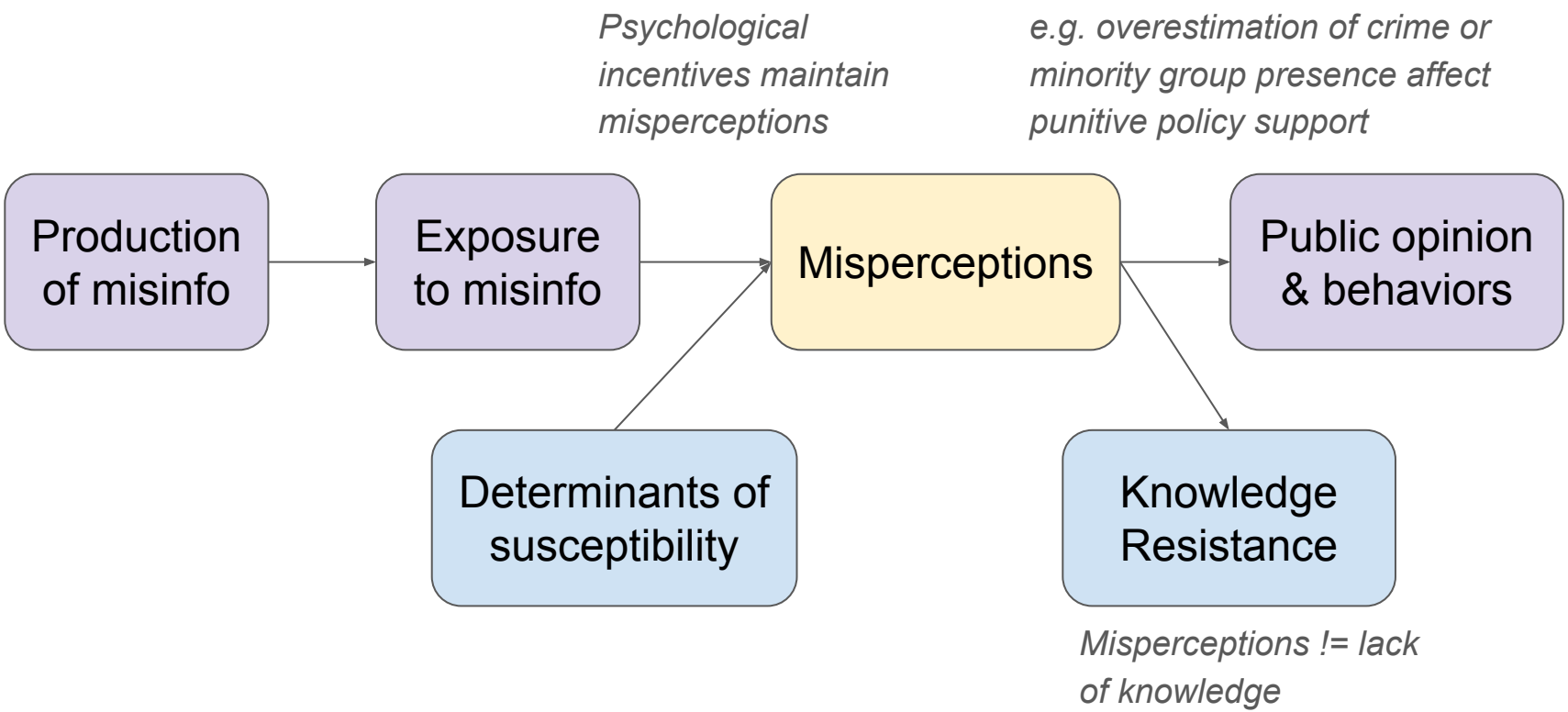
Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New media & society*, *23*(5), 1301-1326.

Experts have highest agreement on role of identity-based biases

Discrepancy: most interventions target the least-agreed upon determinants



Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, *4*(4), 1-34.



Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association, 48(2), 139-166.

(Figure ours)

References

References

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. Harvard Kennedy School (HKS) Misinformation Review.

Barman, D., Guo, Z., & Conlan, O. (2024). The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. Machine Learning with Applications, 100545.

Bergstrom, C. T., & West, J. D. (2021). Calling bullshit: The art of skepticism in a data-driven world. Random House Trade Paperbacks.

Broda, E., & Strömbäck J. (2024). Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review, Annals of the International Communication Association, 48(2), 139–166.

Budak, C., Nyhan, B., Rothschild, D.M. et al (2024). Misunderstanding the harms of online misinformation. Nature 630, 45–53.

Calvo, P., & Saura García, C. (2024). Generative AI and Democracy: the synthetification of public opinion and its impacts. Available at SSRN 4911710.

Chen, C., & Shu, K. (2024). Combating misinformation in the age of llms: Opportunities and challenges. AI Magazine, 45(3), 354-368.

Costello, T. H., Pennycook, G., & Rand, D. (2025). Just the facts: How dialogues with AI reduce conspiracy beliefs. OSF Preprint.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Frontiers in Public Health, 11, 1166120.

References

Dessler, A. E., & Parson, E. A. (2019). The science and politics of global climate change: A guide to the debate. Cambridge University Press.

Drolsbach, C., & Pröllochs, N. (2025). Characterizing AI-Generated Misinformation on Social Media. *arXiv preprint arXiv:2505.10266*.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology, 1(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., ... & Tremayne, M. (2017). Claimbuster: The first-ever end-to-end fact-checking system. Proceedings of the VLDB Endowment, 10(12), 1945-1948.

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. Information Processing & Management, 57(5), 102261.

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2020). A systematic literature review on disinformation: Toward a unified taxonomical framework. New Media & Society, 23(5), 1301-1326.

Lenti, J., Mejova, Y., Kalimeri, K., Panisson, A., Paolotti, D., Tizzani, M., & Starnini, M. (2023). Global misinformation spillovers in the vaccination debate before and during the COVID-19 pandemic: multilingual Twitter study. JMIR infodemiology, 3, e44714.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European review of social psychology, 32(2), 348-384.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. Journal of Experimental Psychology: Applied, 27(1), 1.

Ozawa, J. V., Woolley, S., & Lukito, J. (2024). Taking the power back: How diaspora community organizations are fighting misinformation spread on encrypted messaging apps. Harvard Kennedy School Misinformation Review.

References

Pantazi, M., Hale, S., & Klein, O. (2021). Social and Cognitive Aspects of the Vulnerability to Political Misinformation. Political Psychology, 42(S1), 267–304. https://doi.org/10.1111/pops.12797

Pathak, R., Spezzano, F., & Pera, M. S. (2023). Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. ACM Transactions on the Web, 17(4), 1-26.

Pielke Jr, R. A. (2004). When scientists politicize science: making sense of controversy over The Skeptical Environmentalist. Environmental Science & Policy, 7(5), 405-417.

Saeidnia, H. R., Hosseini, E., Lund, B., Tehrani, M. A., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches. Knowledge and Information Systems, 67, 3139–3158.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023, November). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In 2023 International Conference on Computer and Applications (ICCA)(pp. 1-7). IEEE.

Tomassi, A., Falegnami, A., & Romano, E. (2024). Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society. Plos one, 19(5), e0303183.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society, 6(1), 2056305120903408.

References

Vicari, R., & Komendatova, N. (2023). Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. Humanities and Social Sciences Communications, 10(1), 1-14.

Wang, J., Wang, X., & Yu, A. (2025). Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy. Scientific Reports, 15(1), 1118.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe.

Xu, D., Fan, S., & Kankanhalli, M. (2023, October). Combating misinformation in the era of generative AI models. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9291-9298).

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. Human Behavior and Emerging Technologies, 1(1), 48-61.

Zaki, M. Z., & Ahmed, U. (2024). Bridging linguistic divides: The impact of AI-powered translation systems on communication equity and inclusion. Journal of Translation and Language Studies, 5(2), 20–30.

Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 4602-4611).

All pictures are downloaded from unsplash.com.