Generative Artificial Intelligence and Vaccine (and Public Health) Disinformation

Richard M. Carpiano, PhD, MPH

Professor of Public Policy

University of California, Riverside

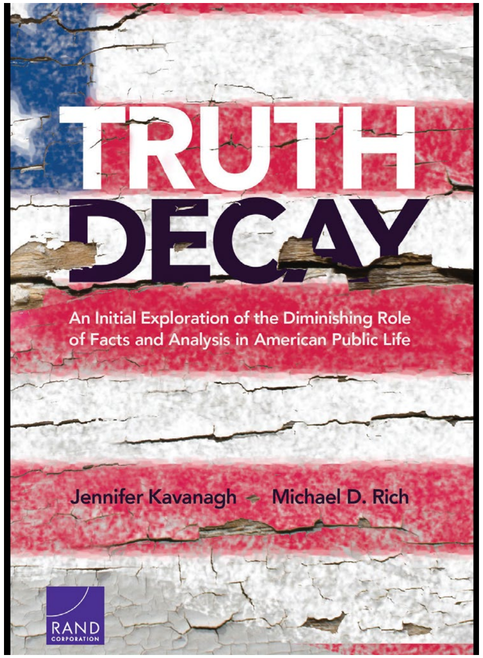
8th Vaccine Acceptance Conference Series

Mérieux Foundation

Annecy, France

October 30-November 1, 2023

We are in a period of “Truth Decay” (Kavanaugh & Rich, 2018)

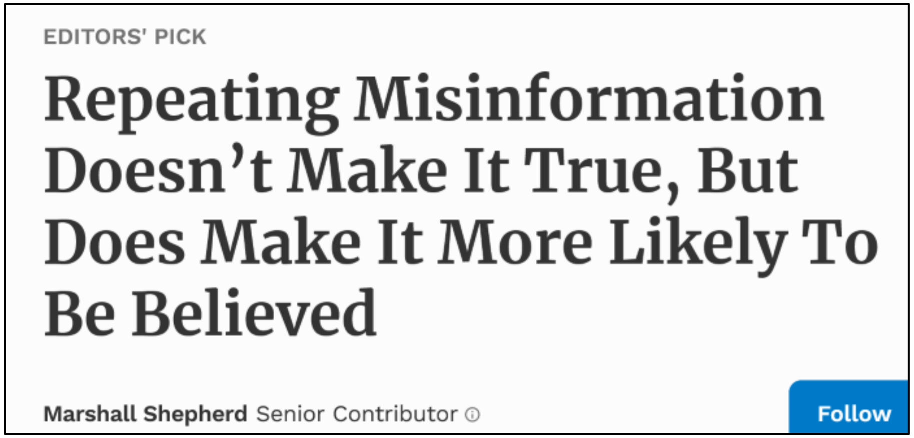

TRUTH DECAY
An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life
Jennifer Kavanagh
Michael D. Rich
RAND CORPORATION

- Defined as a set of 4 inter-related trends:
  1. increasing disagreement about facts and data
  2. a blurring of the line between opinion and fact
  3. increasing relative volume and resulting influence of opinion over fact
  4. Declining trust in formerly respected sources of factual information

- Driven by:
  1. cognitive processing and biases
  2. changes in the information system
  3. competing demands on the educational system
  4. social polarization

Information ecosystem facilitates spread of inaccurate information…

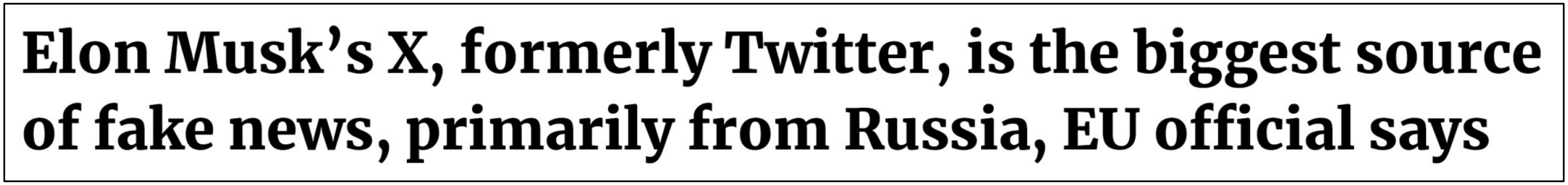
- “The real opposition is the media. And the way to deal with them is to flood the zone with shit.” –Steve Bannon, US political strategist


EDITORS' PICK

Repeating Misinformation Doesn’t Make It True, But Does Make It More Likely To Be Believed

Marshall Shepherd Senior Contributor ⓘ

Illusory Truth Effect; Fazio et al. (2015), J of Experiment Psychology: General


Elon Musk’s X, formerly Twitter, is the biggest source of fake news, primarily from Russia, EU official says

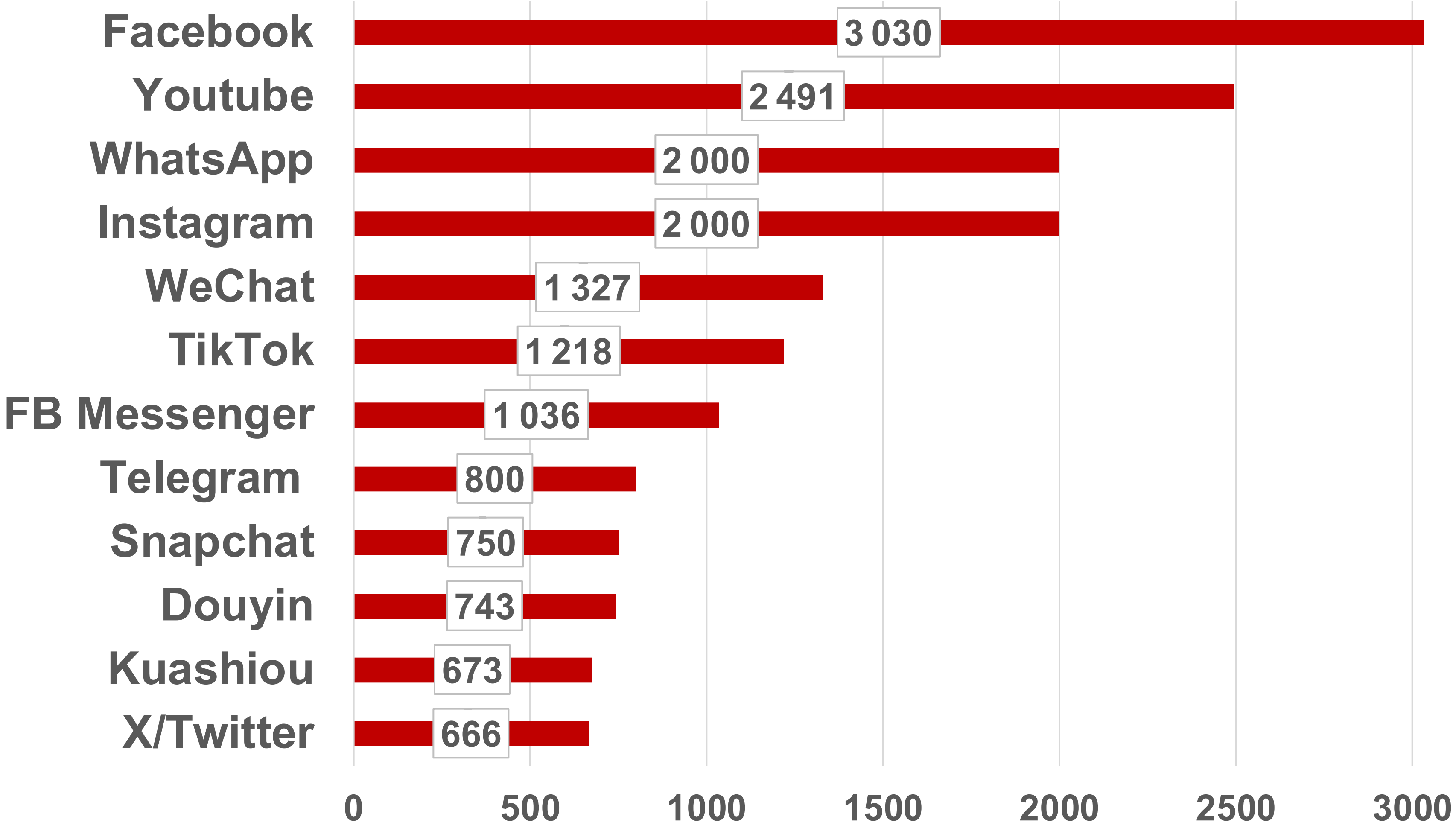
Sept. 7, 2023, Associated Press

Information ecosystem facilitates spread of inaccurate information…

- As of October 2023:
  - **5.3 billion internet users worldwide**
  - 65.7% of the global population (Statistica, 2023)
- Huge global target for **disinformation**
  - "false information which is deliberately intended to mislead—intentionally misstating the facts" (American Psychological Association, 2023)

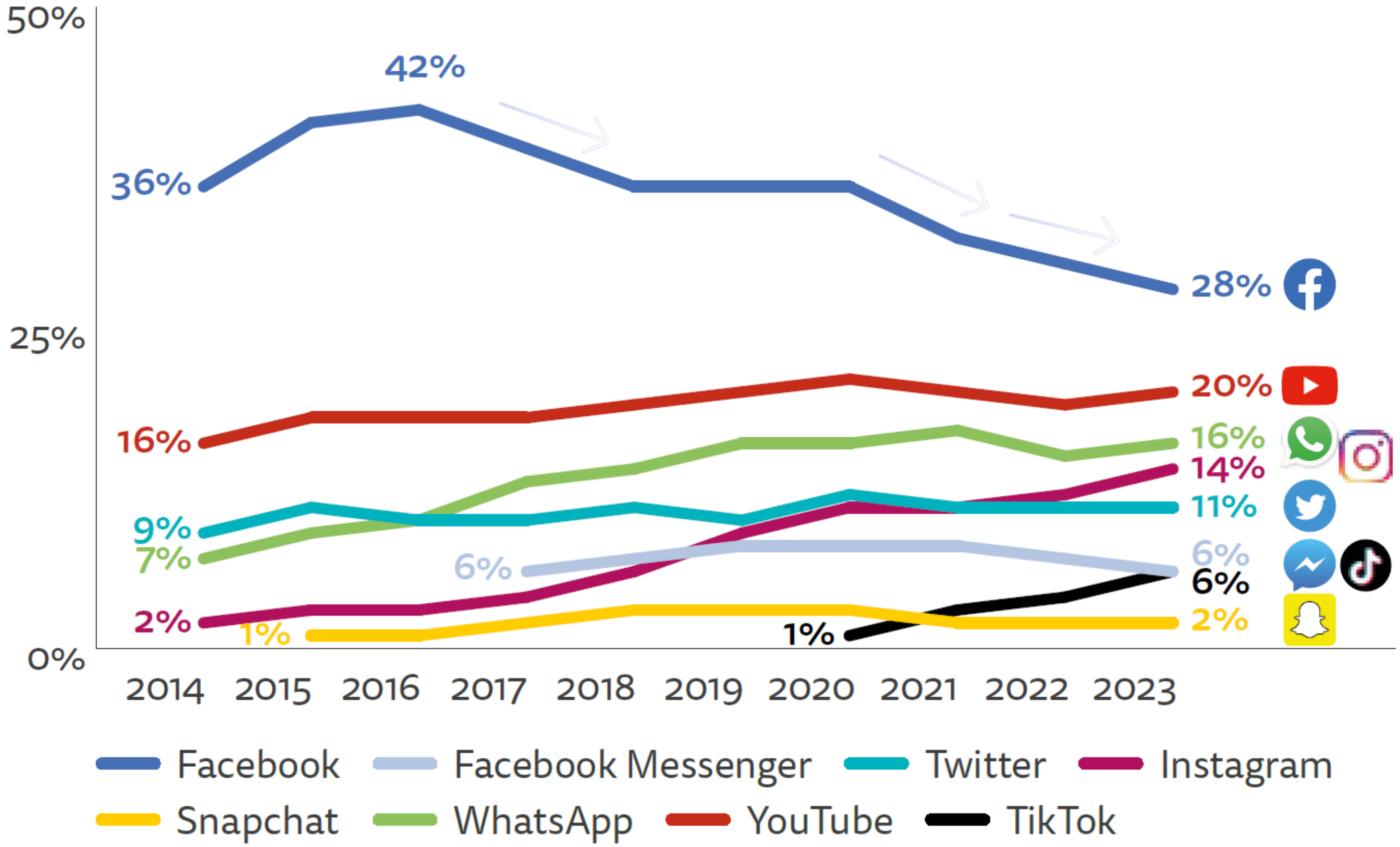
Source: https://calmatters.org/commentary/2022/11/california-disinformation-social-media-literacy-conspiracy/

Most popular social networks worldwide as of October 2023, by monthly active users (in millions)



Source: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

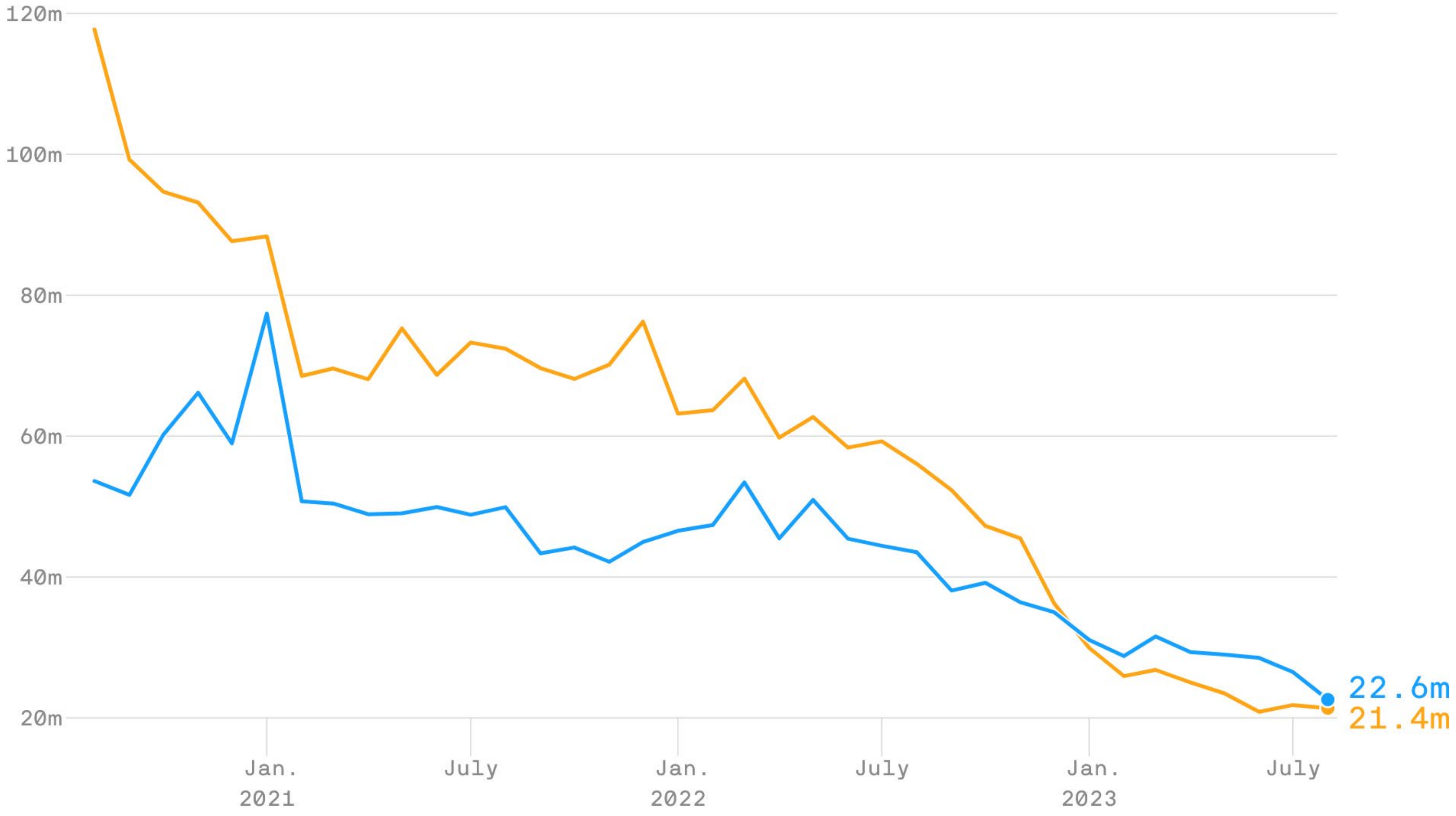
Survey: Proportion that used each social network for news in the last week, 2014-2023



Q12B. Which, if any, of the following have you used for news in the last week? *Base: Total sample in each country-year in UK, USA, Germany, France, Spain, Italy, Ireland, Denmark, Finland, Japan, Australia, Brazil, and Ireland ≈ 2000. Note: No data from Australia or Ireland in 2014.*

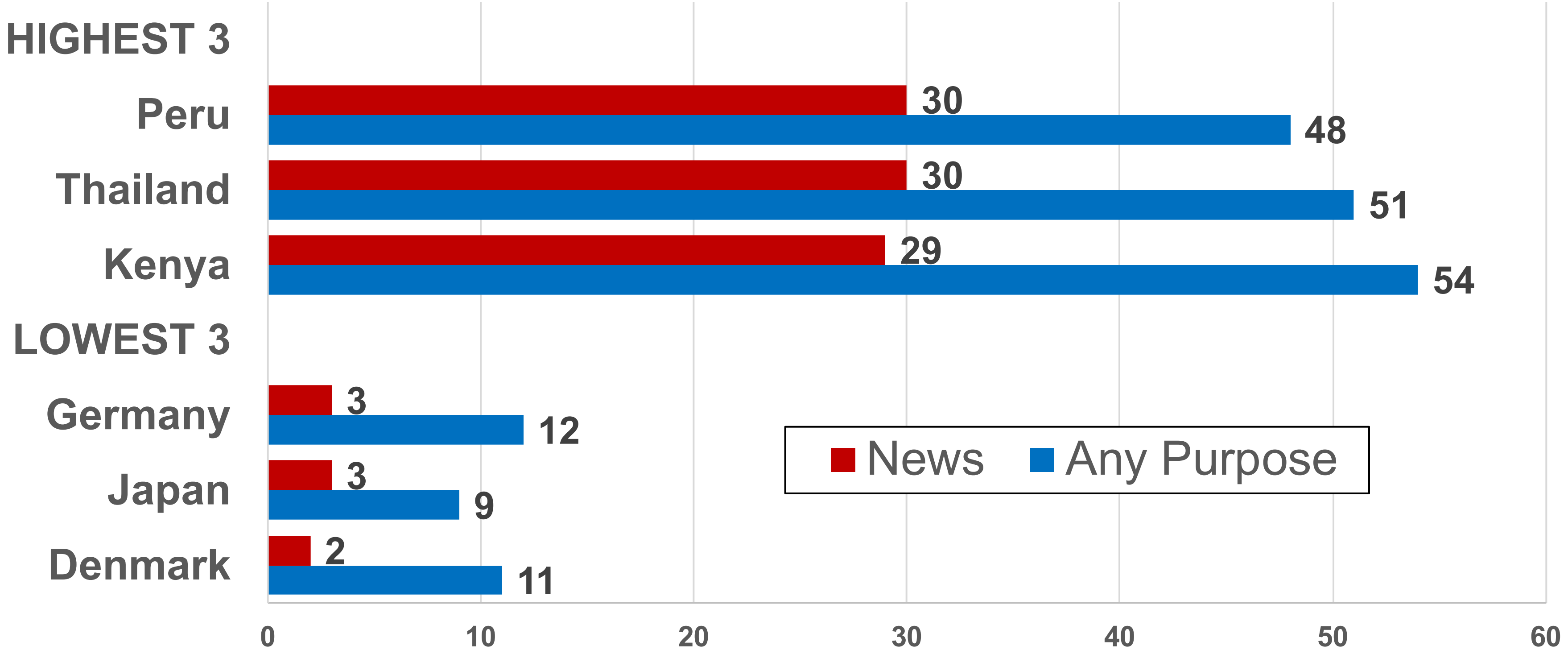
Survey Respondents ages 18+; Source: Reuters (2023) https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Facebook (orange) and X (blue) referrals to top global news sites declined, Aug. 2020-Aug. 2023



Source: https://www.axios.com/2023/10/03/social-media-traffic-news-sites-decrease

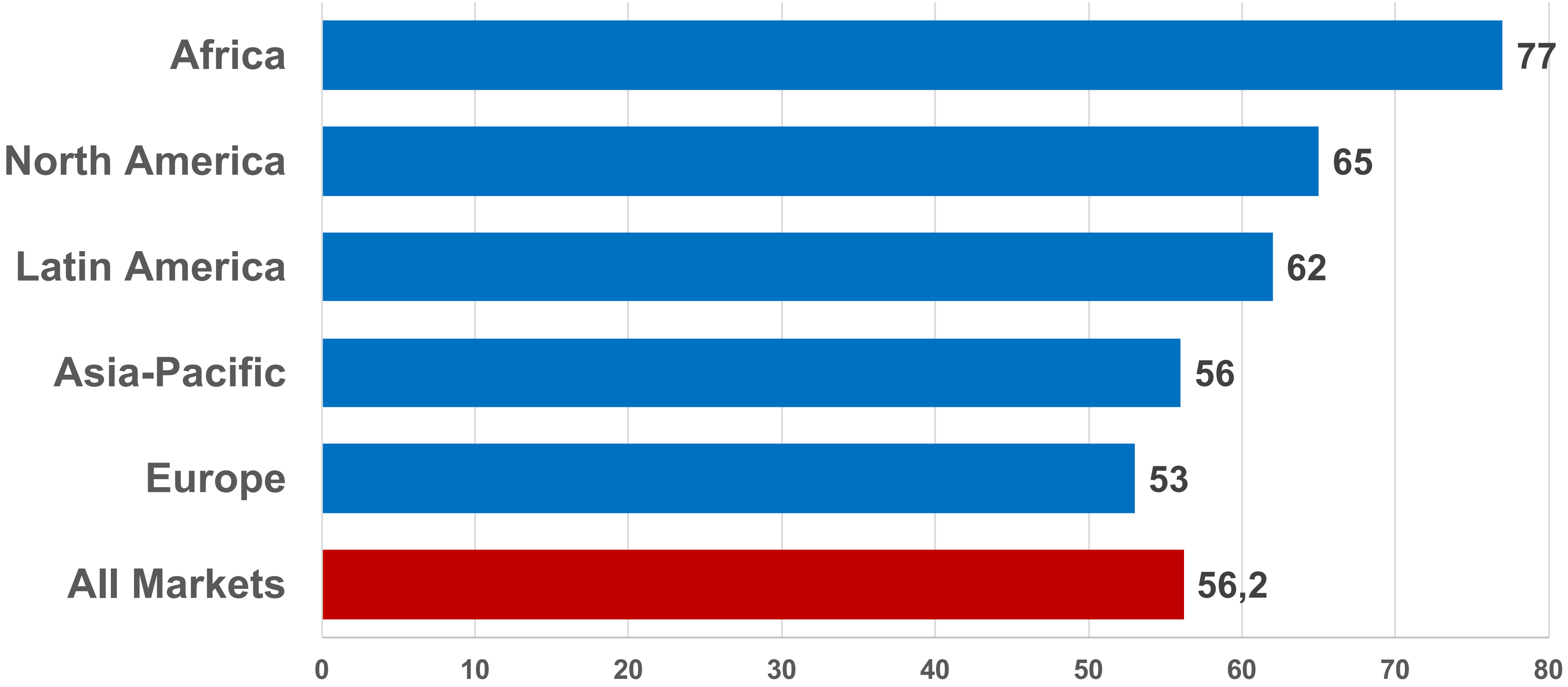
What about TikTok (video)? Percent who used TikTok for News/Any Purpose in last week, end of Jan.-early Feb. 2023



Q12B. Which, if any, of the following have you used for news in the last week? Base: Total sample in each market ≈ 2000. Note: TikTok has been banned in India and does not operate in Hong Kong

Source: Reuters (2023) https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

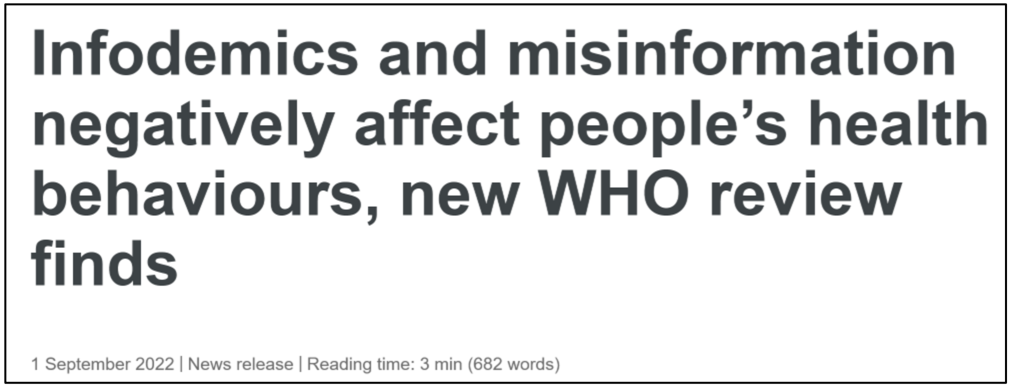
Proportion concerned about what is real and what is fake on the internet when it comes to news, end of Jan.-early Feb. 2023



Source: Reuters (2023) https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Potential for Infodemics (WHO, 2023)

- too much information including false or misleading information in digital and physical environments during a disease outbreak
- causes confusion and harmful, risk-taking behaviors
- leads to mistrust in health authorities
- undermines public health response



Infodemics and misinformation negatively affect people's health behaviours, new WHO review finds

1 September 2022 | News release | Reading time: 3 min (682 words)

Source: https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds

How might Generative AI make vaccine disinformation worse?

- **Generative AI:** algorithms (such as ChatGPT and DALL-E) that can be used to create new content, including audio, code, images, text, simulations, and videos. (McKinsey & Co., 2023)
  - **“Social media manipulation generation 3.0”:** a technological leap blurring the line of “what is detectable as real vs. synthetic… the critical jump forward is in the plausibility of the messenger rather than the message.” (Marcelliano et al., 2023)

Source: https://crowdinsights.co/top-generative-ai-companies/

My Roadmap for Today

1. Review 4 Generative AI methods that have significant potential for disinformation efforts

2. Tactics for which Generative AI can be applied for producing vaccine—and broader public health-related–disinformation
    - factors that could mitigate such threats

3. Tactics for confronting AI-based disinformation
    - efforts to detect and counter them
    - some policy efforts and recommendations

Four Generative AI Technologies with Prime Potential for Disinformation Campaigns

1. Generative Text
2. Voice cloning
3. Deepfake videos
4. Deepfake Images

1. Generative Text

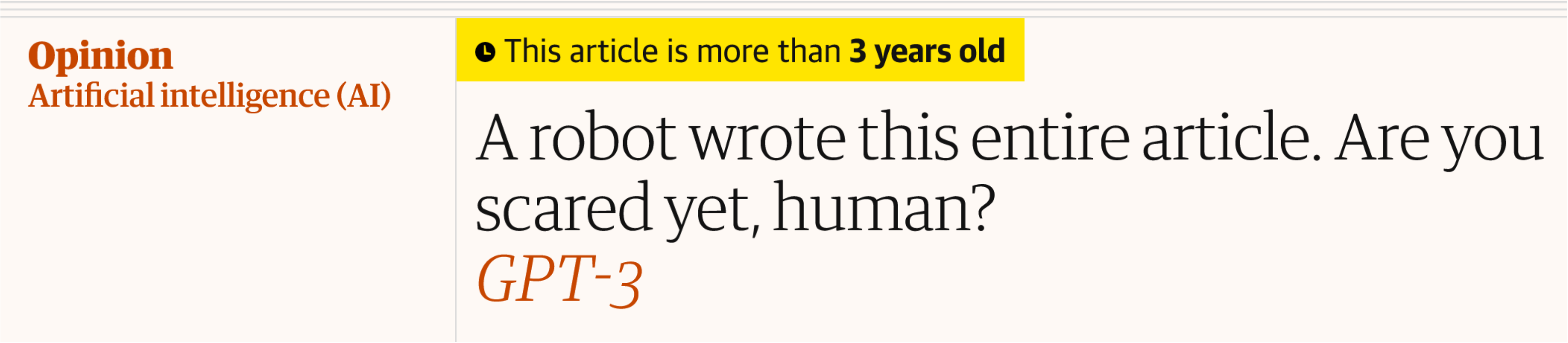
- Capacity to generate artificial, yet life-like text
  - For longer documents, *may* be less convincing,
  - For brief social media postings, perhaps more so

- can produce persuasive text, including articles that survey respondents rate as credible as real news articles (Goldstein et al., 2023)

1. Generative Text: Potential Risks

a. **Fake/misleading publications:**

- op-eds, studies/research reports (or press releases)
  - e.g., *Guardian* (2020) GPT3 op-ed
  - fed prompts, 8 versions cherry-picked for best parts and composed by humans into one essay



Opinion
Artificial intelligence (AI)

◷ This article is more than **3 years old**

A robot wrote this entire article. Are you scared yet, human?

GPT-3

Source:The Guardian (2023): https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

1. Generative Text: Potential Risk (cont'd)

b. **Astroturfing:** generate large numbers of inauthentic (fake) accounts (bots) to create the appearance of broad consensus on a vaccine-related falsehood or concern
   - Make antivax campaigns appear "grassroots" when it is not (e.g., different fake community groups—a phony coalition)

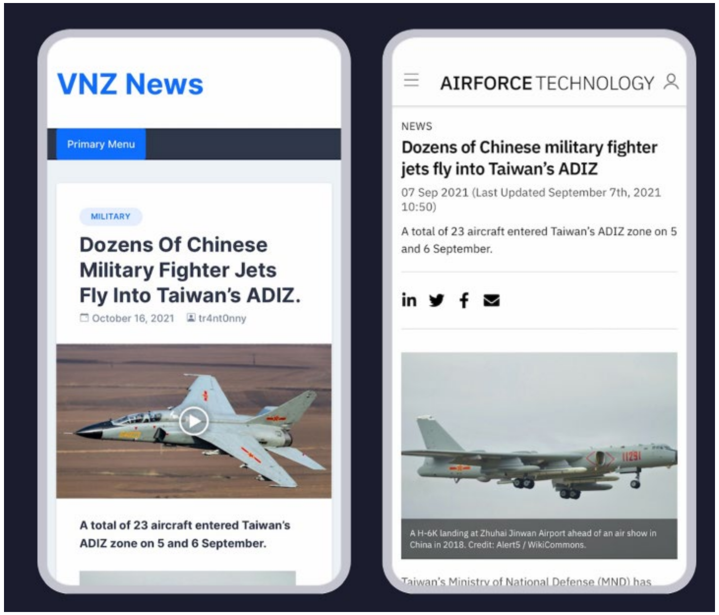
c. **Algorithm Gaming:** Mass-produce fake news stories to overwhelm truthful coverage and search engine results
   - e.g., Chinese government tactic to suppress info

1. Generative Text: Potential Risk (cont'd)

d. **More easily create fake sources**:

  d. fake experts (e.g., LinkedIn profiles) (Guled Ali, 2023)

  e. websites/blogs for posting pseudoscience and other claims, news, personal anecdotes (website farms)



VNZ News

Primary Menu

MILITARY

Dozens Of Chinese Military Fighter Jets Fly Into Taiwan's ADIZ.

October 16, 2021 tr4nt0nny

A total of 23 aircraft entered Taiwan's ADIZ zone on 5 and 6 September.

AIRFORCE TECHNOLOGY

NEWS

Dozens of Chinese military fighter jets fly into Taiwan's ADIZ

07 Sep 2021 (Last Updated September 7th, 2021 10:50)

A total of 23 aircraft entered Taiwan's ADIZ zone on 5 and 6 September.

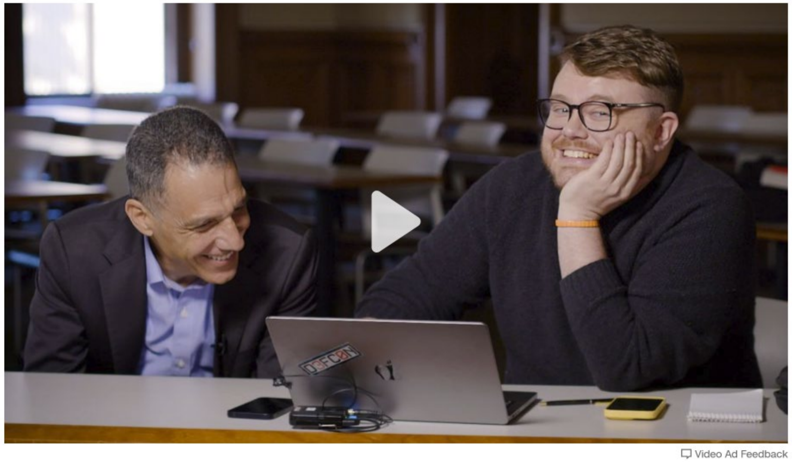
in 🐦 f ✉

A H-6K landing at Zhuhai Jinwan Airport ahead of an air show in China in 2018. Credit: Alert5 / WikiCommons.

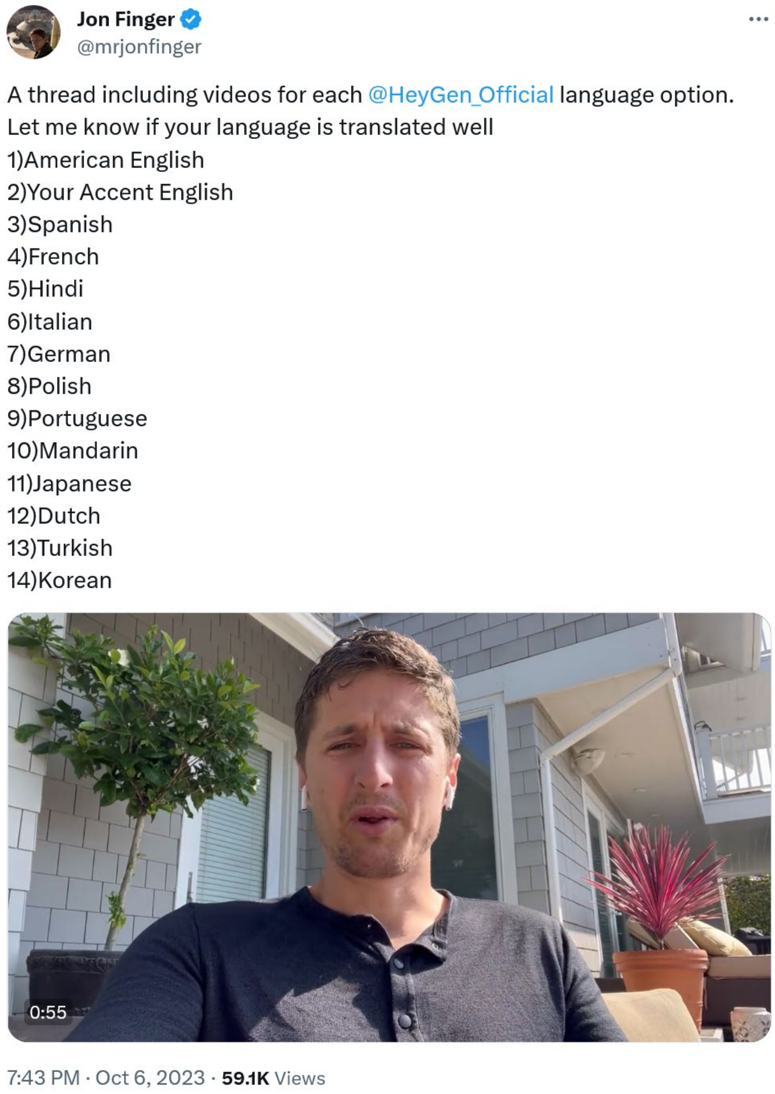
Taiwan's Ministry of National Defense (MND) has

2. Voice Cloning

- Simulates one’s voice
- Language translation (with convincing lip sync)
- Requires just short time of talking/audio to work



CNN reporter calls his parents using AI voice. Watch what happens next



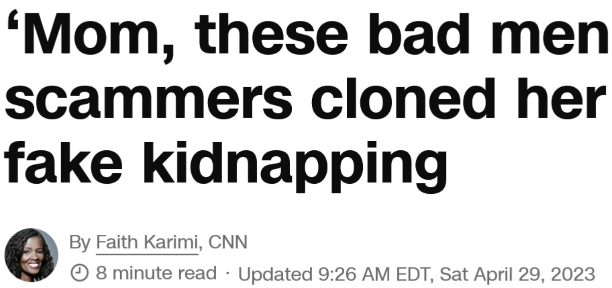
Jon Finger @mrjonfinger

A thread including videos for each @HeyGen_Official language option.
Let me know if your language is translated well
1)American English
2)Your Accent English
3)Spanish
4)French
5)Hindi
6)Italian
7)German
8)Polish
9)Portuguese
10)Mandarin
11)Japanese
12)Dutch
13)Turkish
14)Korean

7:43 PM · Oct 6, 2023 · **59.1K** Views

2. Voice Cloning: Potential Risks

- Easy for bad actors to:
  - expand reach from one population/subculture to many others
  - mimic trusted others (local clinicians, community leaders) in making false statements, convincing community members on claims that align with contrarians, grifters, and anti-vax groups (vaccines, public health, conspiracy theories)
  - prey on historically marginalized racial-ethnic groups



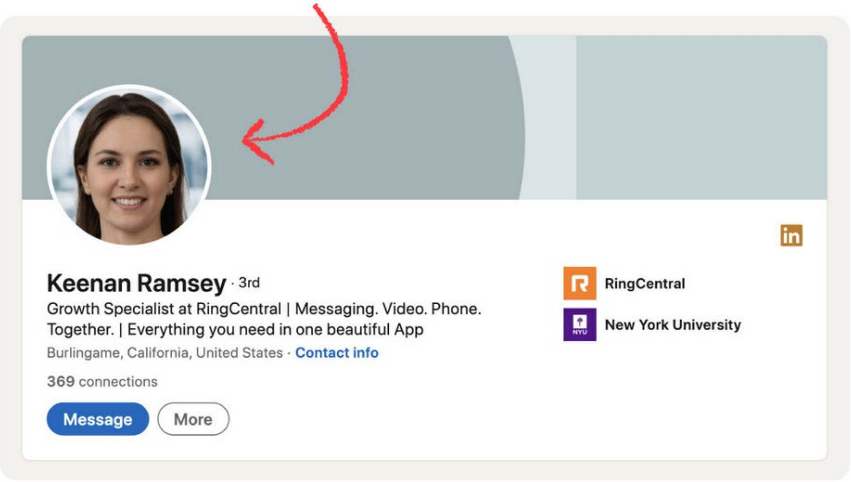
‘Mom, these bad men have me’: She believes scammers cloned her daughter’s voice in a fake kidnapping

By Faith Karimi, CNN

8 minute read · Updated 9:26 AM EDT, Sat April 29, 2023

3. Deepfake Images

- Using AI to create images of people who do not exist
- open-source text-to-image models can generate photorealistic images of anything (real or imagined) and do so at scale (e.g., Stable Diffusion)
- easier than using real stolen images
  - fakes are untraceable



Keenan Ramsey · 3rd

Growth Specialist at RingCentral | Messaging. Video. Phone. Together. | Everything you need in one beautiful App

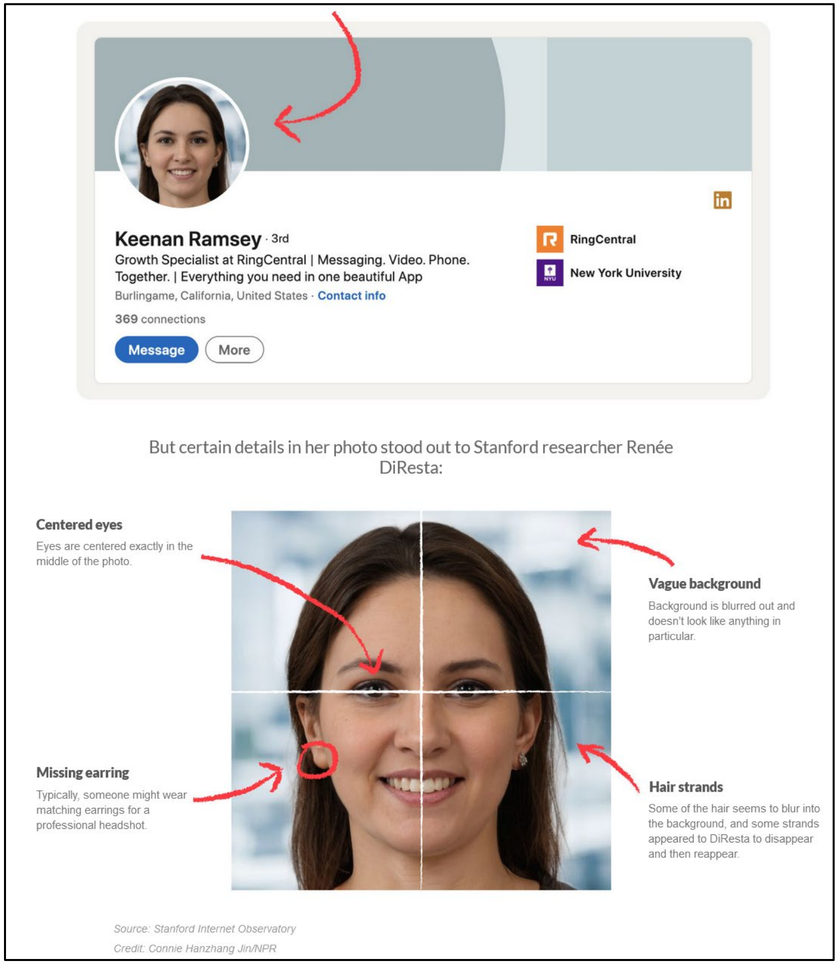
Burlingame, California, United States · Contact info

369 connections

Source: https://commons.wikimedia.org/w/index.php?curid=136083293

3. Deepfake Images: Potential Risk

- Could be used to create fake profiles of:
  - medical, public health, or scientific “experts”
  - vaccine-injured people
  - parents of purportedly vaccine-injured children



Keenan Ramsey · 3rd

Growth Specialist at RingCentral | Messaging. Video. Phone. Together. | Everything you need in one beautiful App

Burlingame, California, United States · Contact info

369 connections

RingCentral

New York University

But certain details in her photo stood out to Stanford researcher Renée DiResta:

Centered eyes
Eyes are centered exactly in the middle of the photo.

Vague background
Background is blurred out and doesn't look like anything in particular.

Missing earring
Typically, someone might wear matching earrings for a professional headshot.

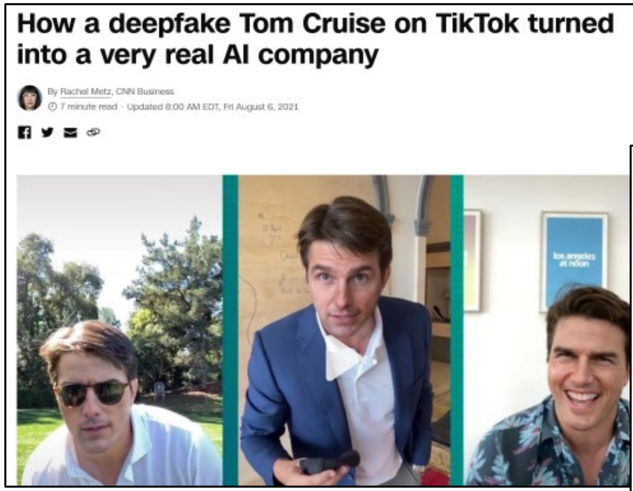
Hair strands
Some of the hair seems to blur into the background, and some strands appeared to DiResta to disappear and then reappear.

Source: Stanford Internet Observatory
Credit: Connie Hanzhang Jin/NPR

Source: Bond, S. (2022). NPR. https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles

4. Deepfake Videos

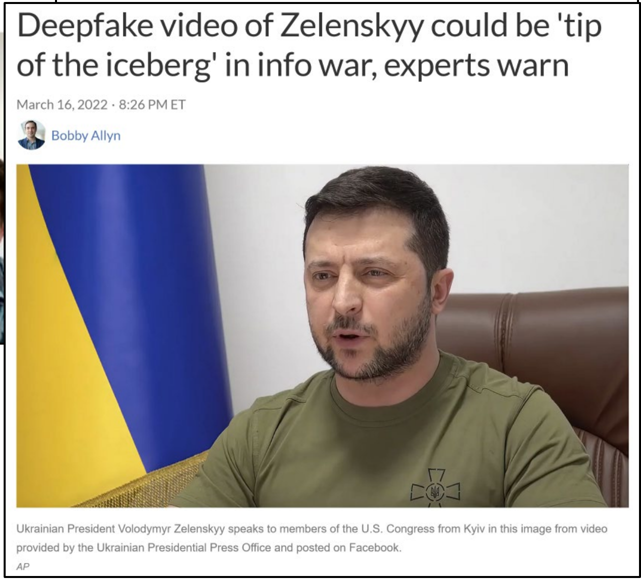
- an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said (Merriam-Webster, 2023)



How a deepfake Tom Cruise on TikTok turned into a very real AI company

By Rachel Metz, CNN Business

7 minute read · Updated 8:00 AM EDT, Fri August 6, 2021



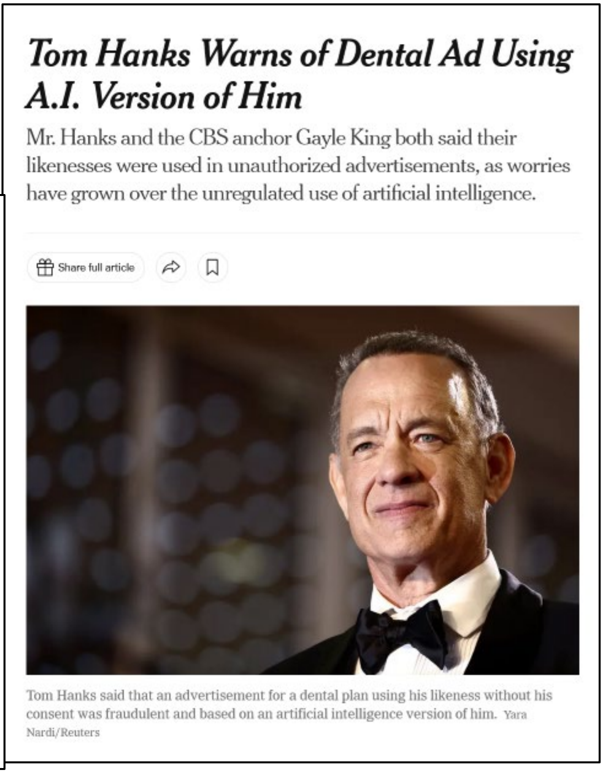
Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn

March 16, 2022 · 8:26 PM ET

Bobby Allyn

Ukrainian President Volodymyr Zelenskyy speaks to members of the U.S. Congress from Kyiv in this image from video provided by the Ukrainian Presidential Press Office and posted on Facebook.

AP



Tom Hanks Warns of Dental Ad Using A.I. Version of Him

Mr. Hanks and the CBS anchor Gayle King both said their likenesses were used in unauthorized advertisements, as worries have grown over the unregulated use of artificial intelligence.

Tom Hanks said that an advertisement for a dental plan using his likeness without his consent was fraudulent and based on an artificial intelligence version of him. Yara Nardi/Reuters

4. Deepfake Videos: Potential Risks

- Videos portraying elected leaders, scientists, public health officials, trusted community figures committing acts/making statements that are discrediting:
  - related to vaccines, disease threats, and public health policies or
  - in general to undermine them and public’s trust in them (not unlike women victimized by deepfake pornography)

- Play to conspiracy theories and audience’s confirmation biases about a disease threat, government, science, medicine, “big pharma”

Good News: Factors that hamper Deep Fakes

1. **“Shallow fakes” just as effective**
   - videos shared with misleading contexts or small edits
   - more widespread and easier to produce

2. **Cost for amateurs** (time, equipment, other resources)
   - but getting more affordable and accessible

Good News: Factors that hamper Deep Fakes (cont'd)

3. **Time:** requires planning to create content for a specific strategy; operations need to be planned well in advance
   - Limits rapid deployment
   - Could be good for getting ahead of disinformation (establishing the narrative, pre-bunking)

4. **Extensive model training data required**
   - Why you see them for prominent people (politicians, actors) vs. lesser-known people

Also… End users

- Studies find that viewers (so far) have some decent capacity on average to catch deepfakes when they see it.

- Conjecture: generational—youth are probably already most familiar with it
  - potential for inoculation theory approaches?

5 Approaches to Counter Deep & Related AI Fakes

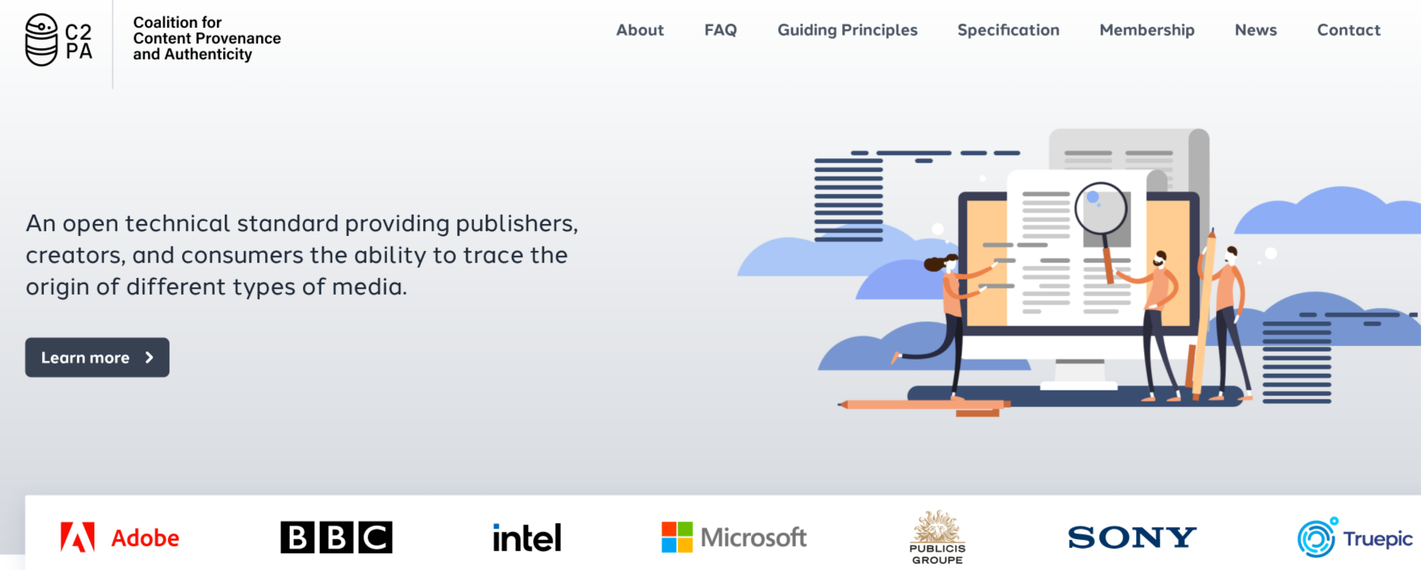
1. Detection
2. Provenance
3. Open-Source Intelligence Techniques and Journalistic Approaches
4. Media Literacy
5. Regulation

1. Detection

- Develop systems to automatically detect deepfakes
  - US Government investment (Defense Advanced Research Projects Agency; DARPA)
  - Facebook "Deepfake Challenge Competition"
- Industry can:
  - provide access to their repositories of images to serve as training data for keeping detection programs up-to-date (Google has done this)
  - provide access to "Radioactive" training data that would make content more easily detectable later by detection programs
  - limit access to most high tech/effective detectors
- Government could limit public access to government-funded detectors for finding deepfakes that undermine national security.

2. Content Provenance

- Ways to show source, history/alteration of media content
- Private sector: Coalition for Content Provenance and Authenticity (C2PA) collaboration to develop technical standards and methods
  - E.g., for digitally capturing provenance of photo images

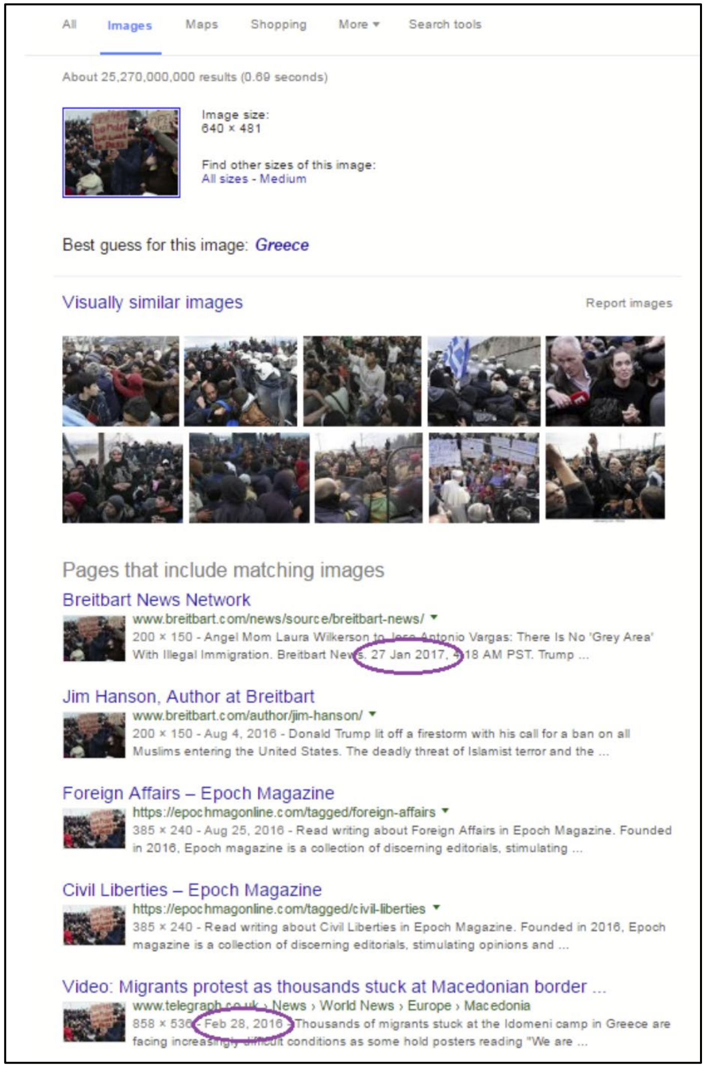


An open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media.

3. Open-Source Intelligence Techniques and Journalistic Approaches

- Develop and share tools that can be used to identify deepfakes and other disinformation content
    - e.g., reverse image search

- Important for journalists in small/midsize news organizations, who need to rely on such open-source tools to verify content authenticity



About 25,270,000,000 results (0.69 seconds)

Image size:
640 × 481

Find other sizes of this image:
All sizes - Medium

Best guess for this image: *Greece*

Visually similar images

Pages that include matching images

Breitbart News Network
www.breitbart.com/news/source/breitbart-news/ ▾
200 × 150 - Angel Mom Laura Wilkerson to Jose Antonio Vargas: There Is No 'Grey Area' With Illegal Immigration. Breitbart News. 27 Jan 2017, 4:18 AM PST. Trump ...

Jim Hanson, Author at Breitbart
www.breitbart.com/author/jim-hanson/ ▾
200 × 150 - Aug 4, 2016 - Donald Trump lit off a firestorm with his call for a ban on all Muslims entering the United States. The deadly threat of Islamist terror and the ...

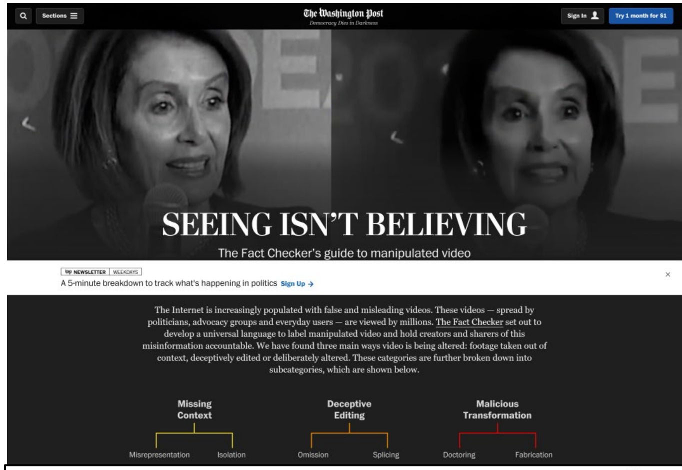
Foreign Affairs – Epoch Magazine
https://epochmagonline.com/tagged/foreign-affairs ▾
385 × 240 - Aug 25, 2016 - Read writing about Foreign Affairs in Epoch Magazine. Founded in 2016, Epoch magazine is a collection of discerning editorials, stimulating ...

Civil Liberties – Epoch Magazine
https://epochmagonline.com/tagged/civil-liberties ▾
385 × 240 - Read writing about Civil Liberties in Epoch Magazine. Founded in 2016, Epoch magazine is a collection of discerning editorials, stimulating opinions and ...

Video: Migrants protest as thousands stuck at Macedonian border ...
www.telegraph.co.uk › News › World News › Europe › Macedonia
858 × 536 - Feb 28, 2016 - Thousands of migrants stuck at the Idomeni camp in Greece are facing increasingly difficult conditions as some hold posters reading "We are ...

4. Media Literacy

- Empower public via efforts to train them to detect deepfake content (by itself & as part of mis-/disinfo education)
  - Create resources:
    - e.g., *Washington Post* (2019) guide to manipulated videos
  - Build awareness by creating and publicizing deepfake content
    - e.g., fake Nixon speech by MIT researchers
- Issues: Scaling up? Systematic dissemination and evaluation?

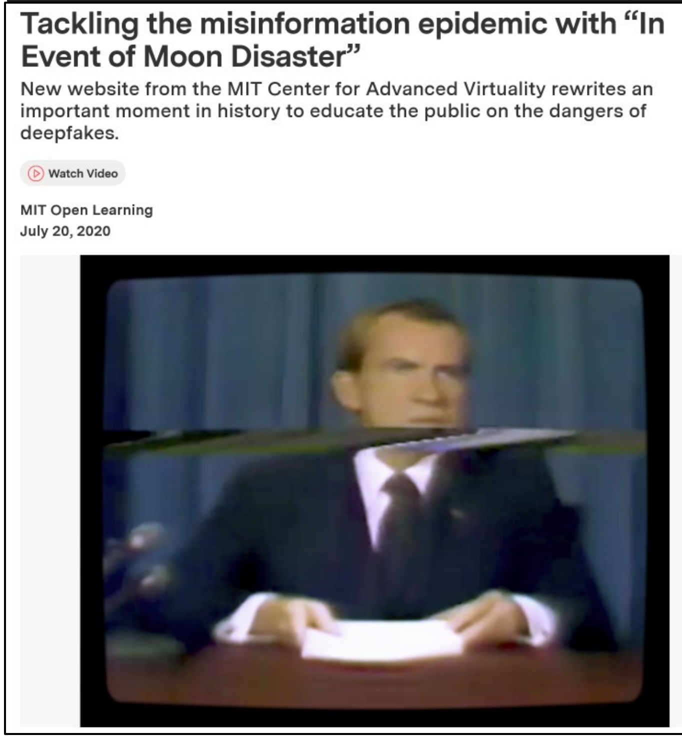


SEEING ISN'T BELIEVING

The Fact Checker's guide to manipulated video

A 5-minute breakdown to track what's happening in politics Sign Up →

The Internet is increasingly populated with false and misleading videos. These videos — spread by politicians, advocacy groups and everyday users — are viewed by millions. The Fact Checker set out to develop a universal language to label manipulated video and hold creators and sharers of this misinformation accountable. We have found three main ways video is being altered: footage taken out of context, deceptively edited or deliberately altered. These categories are further broken down into subcategories, which are shown below.



Tackling the misinformation epidemic with "In Event of Moon Disaster"

New website from the MIT Center for Advanced Virtuality rewrites an important moment in history to educate the public on the dangers of deepfakes.

MIT Open Learning
July 20, 2020

5. Regulation: US State Efforts

- NY: Deepfake likeness bill
- Also, California, Texas, Connecticut
- Good step, but problems:
  - How to handle bad actors outside these states?
  - First Amendment scrutiny in court?

5. Regulation: US Federal Efforts

1. White House (July 2023):
   - secured voluntary commitments with 7 top AI companies "to help move toward safe, secure, and transparent development of AI technology"

2. White House (Oct. 30, 2023—yesterday):
   - issues Executive Order on **"Safe, Secure, and Trustworthy Artificial Intelligence"**
     - **Among its many actions:** "Protect Americans from AI-enabled fraud and deception by establishing standards and best practices for detecting AI-generated content and authenticating official content."

5. Regulation: Private Sector Efforts

- Industry (March 2023): Open call for pause on A.I., citing "Profound Risks to Society"
    - AI R&D should be refocused on making… systems more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, & loyal."
    - "In parallel, AI developers must work with policymakers to dramatically accelerate development of robust AI governance systems."

5. Regulation: Private Sector Efforts (cont’d)

- Google
  - Aug. 2023: (subsidiary) DeepMind launched watermarking tool for AI-generated images
    - first Big Tech firm to publicly launch one, after July WH pledge to develop them
    - But limited roll-out (“experimental”) and proprietary
  - Sept. 2023: will require election ads to “prominently disclose” AI content

5. Regulation: Globally

- UK White Paper on AI (March 2023): supports existing regulators to develop a sector-focused, principles-based approach. Not legally binding.
- EU AI Act (June 2023, still being finalized): Generative AI would have to comply with transparency requirements
  - E.g., disclosing content was AI-generated
- China (Aug. 15, 2023): “Interim Measures for the Management of Generative Artificial Intelligence Services to regulate the use of generative artificial intelligence (AI)”

In Conclusion

1. **Generative AI offers easier means to existing duplicitous ends**
   - not a new problem but a worsening of an existing, vexxing problem: amplifying vaccination and public health mis-/disinformation
2. **Many efforts underway to respond to negative impacts**
   - But still early days, uncoordinated, rely heavily on industry compliance/benevolence
   - Not focused on health disinformation, but direct spillover effects and potential
3. **Requires greater (and new?) partnerships**
   - with government & industry for infodemic surveillance and response—for public protection and domestic/global health security

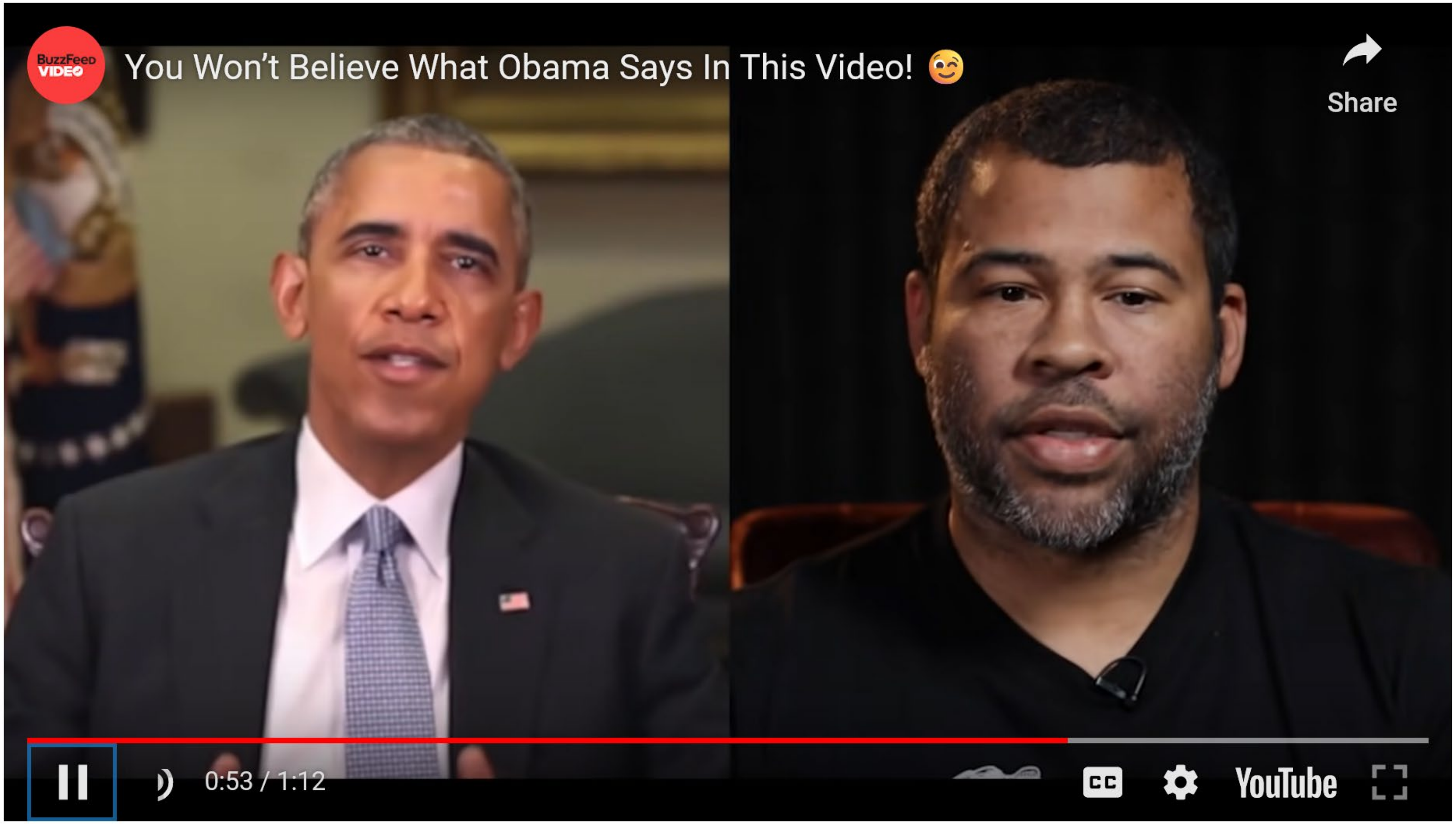
Thank You!

richard.carpiano@ucr.edu

X (Twitter): @RMCarpiano
Bluesky: @rmcarpiano.bsky.social

Acknowledgments:
Reed Berkowitz, Curioser LLC

- https://youtu.be/cQ54GDm1eL0



References

1. Chan, K. (2023, September). Musk’s X is the biggest purveyor of disinformation, EU official says. Associated Press. https://apnews.com/article/disinformation-musk-x-twitter-european-union-9f7823726f812bb357ee4225b884354f
2. Heikkilä, M. (2023, August). Google DeepMind has launched a watermarking tool for AI-generated images. MIT Technology Review. https://www.technologyreview.com/2023/08/29/1078620/google-deepmind-has-launched-a-watermarking-tool-for-ai-generated-images/
3. Helmus, T.C. (2022, July). Artificial intelligence, deepfakes, and disinformation: A primer. RAND. https://www.rand.org/pubs/perspectives/PEA1043-1.html
4. Heywood, D. (2023). The UK’s approach to regulating AI. TaylorWessing. https://www.taylorwessing.com/en/interface/2023/ai---are-we-getting-the-balance-between-regulation-and-innovation-right/the-uks-approach-to-regulating-ai
5. Kavanaugh, J., & Rich, M. D. (2018). Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life. RAND. https://www.rand.org/pubs/research_reports/RR2314.html
6. Kirby, P. (2023, October). New York Governor signs deepfake likeness bill into law. Government Technology. https://www.govtech.com/policy/new-york-governor-signs-deepfake-likeness-bill-into-law
7. Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Navarre Chao, L., & Smithrand, J. (2023, September). The rise of generative AI and the coming era of social media manipulation 3.0. RAND https://www.rand.org/pubs/perspectives/PEA2679-1.html
8. McKinsey & Company. (2023, January). What is generative AI? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai
9. Narayan, J., Hu, K., Coulter, M., & Mukherjee, S. (2023, April). Elon Musk and others urge AI pause, citing 'risks to society’ Associated Press. https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/

References (continued)

10. Rogers, A. & Kinder, T. (2023, September). The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0. Financial Times. https://www.ft.com/content/d1e5a0a9-9438-4fb4-905c-3dd737246600
11. The Washington Post. (2019). *The Washington Post*'s guide to manipulated media. https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/
12. The White House. (2023, July). Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/