BULLETIN OF THE ATOMIC SCIENTISTS
2021, VOL. 77, NO. 3, 119–123
https://doi.org/10.1080/00963402.2021.1912093

Routledge
Taylor & Francis Group

Meme warfare: AI countermeasures to disinformation should focus on popular, not perfect, fakes

Michael Yankoski, Walter Scheirer and Tim Weninger

ABSTRACT
From QAnon conspiracy theories to Russian government sponsored election interference, social media disinformation campaigns are a part of online life, and identifying these threats amid the posts that billions of social media users upload each day is a challenge. To help sort through massive amounts of data, social media platforms are developing AI systems to automatically remove harmful content primarily through text-based analysis. But these techniques won't identify all the disinformation on social media. After all, much of what people post are photos, videos, audio recordings, and memes. Developing the entirely new AI systems necessary to detect such multimedia disinformation will be difficult.

KEYWORDS
Artificial intelligence; social media; disinformation; memes; deepfakes; shallowfakes

Despite several alarming headlines in the press last year like "Deepfakes are coming for American Democracy" (Hwang and Watts 2020), the sophisticated, artificial intelligence- (AI-) generated images and videos called "deepfakes" didn't end up influencing the presidential election (Simonite 2020), and they haven't yet shaped major world events. Rather, as with the case of the viral video that appeared to show Joe Biden wearing a wire on the presidential debate stage (it was a shirt crease), cruder manipulations and simple misrepresentations can be far more effective and influential.

Researchers are only just beginning to understand the threat posed by multimedia (that is, visual and audio) disinformation. From QAnon conspiracy theories to Russian government-sponsored election interference, social media disinformation campaigns are a daunting front in online life, and identifying these threats amid the posts that billions of social media users upload each day is a challenge. To help sort through massive amounts of data, social media platforms are developing AI systems to automatically remove harmful content, primarily through text-based analysis. But these techniques won't identify all the disinformation on social media. After all, much of what people post are photos, videos, audio recordings, and memes.

Academic researchers, including us, have worked to develop AI systems with the technological sophistication to detect faked media items such as photos and videos (Yankoski, Weninger, and Scheirer 2020). In our analyses of disinformation in multimedia content, what we have found is that sophisticated faked content – often called deepfakes – just isn't the most pressing problem. Through our ingestion platforms and media analysis engines, what we are seeing is not the proliferation of fake images, videos, and audio that are so real as to convince someone of something untrue, but rather the proliferation of narratives that emotionally reaffirm a belief that an audience already has (Theisen et al. 2020). Deepfake manipulations on the Internet are niche. The real challenge in using AI to detect multimedia disinformation lies in understanding much more crudely produced content: It is the meme that social media companies and policymakers need to worry about.

Memes and the power of shallow

In 1976, the evolutionary biologist Richard Dawkins needed a term to explain how cultural artifacts evolve over time as they spread across society, replicating through imitative acts (Dawkins 2016). He coined the term "meme," a portmanteau of the ancient Greek word for imitation "mimeme" and the English word "gene." Since then, memes have become an essential form of visual communication. Anything a person can conceive of and express visually is potential meme material.

Memes generally are a kind of shallowfake. In contrast to increasingly realistic, AI-generated deepfakes, these are manipulations that range in production value from plausible to obviously fake. In other words, they're easy to make. Shallowfakes can even be original images or videos that someone has simply relabeled as depicting something else or has subtlety edited to change the perception of the content, for example, by slowing

CONTACT Michael Yankoski myankosk@nd.edu

down the video frame rate (Denham 2020). What is important is that they replicate and spread as rapidly as possible.

Shallowfakes are much better for meme making than deepfakes, which strive for realism. For shallowfakes like memes, it is often the case that the less they correspond to reality, the more effective they can be in spreading online and even swaying human behavior.

Take the saga of a Reddit group called r/wallstreetbets. In late January 2021, the "Redditors," as the platform's users are known, sparked an astronomical rise in the stock price of GameStop. A video game store popular in the early 2000s (Eavis 2021), GameStop had lost traction as shoppers and gamers moved online. For a few wild days in midwinter, however, the price of GameStop stock, which had hovered in the single-digit dollar range for much of 2020, rocketed to triple digits and closed at $347 on January 27 (Yahoo Finance 2021). Hedge funds that had placed bets against GameStop hemorrhaged money and the entire incident prompted congressional inquiries (Flatley and Wasson 2021), a Securities and Exchange Commission investigation (Newmyer and Zapotosky 2021), and numerous lawsuits.

A big part of the r/wallstreetbets story are the memes that Redditors used to generate an enormous response and to coordinate action by playing off of people's emotions and beliefs. For example, one meme depicted a financial commentator who had a negative analysis of GameStop as a beggar. "On the way to the grocery store, found Andrew Left at his new job," the post's caption read. While the Redditors, in this case, weren't pushing disinformation, their tactics perfectly illustrate the power of shallowfakes and memes to drive behavior. Another, perhaps more alarming, example of memes serving as powerful tools for disinformation has been anti-vaccination messaging during the pandemic (Buts 2021). Social media users have posted memes questioning the value of a COVID-19 vaccine, saying, for example, that the flu vaccine hasn't eliminated the flu. Others have linked the vaccines to autism. Such meme-based disinformation campaigns could reinforce existing divides in vaccine support (Funk and Tyson 2021) (Figures 1 and 2).

AI for detecting disinformation

Researchers have put significant resources into the creation of sophisticated AI systems to rapidly detect threats as they emerge in online social media networks. There are text-based hate speech detection systems (PeaceTech Lab, n.d.; Woolley and Howard 2018; Technologies & International Development Lab, n.d.). One of us (W.S.) is developing sophisticated image manipulation detection algorithms to detect doctored images (Theisen et al. 2020; Yankoski, Weninger, and Scheirer 2020). Deepfake video detection systems have the capacity to both detect irregularities, such as noise inconsistency in the video itself (Guarnera, Giudice, and Battiato 2020;

Figure 1. An r/wallstreetbets meme depicts a financial analyst as a beggar.

Figure 2. A Reddit user posted an example of an anti-vaccination meme.

Verdoliva 2020), as well as differences between the affective cues contained in the audio versus the video components of a media item (Mittal et al. 2020).

The problem is that these technologies are often isolated from one another, and thus relatively incapable of detecting meme-based disinformation campaigns. While the technological advances in each of these various areas are laudable, researchers have yet to produce AI systems sophisticated enough to detect coordinated campaigns designed to manipulate how groups of people feel about what they already believe, which is the motivation for campaigns involving memes.

This kind of AI analysis is on another level entirely from existing systems and technologies. It's a method called semantic analysis, a methodology aimed at mapping the meaning of the disinformation campaigns themselves. For semantic analysis, it isn't enough to detect whether a post contains a manipulated image, faked audio clip, or hate-speech. Algorithms need to be able to identify coordinated multimodal (i.e. text/image/video) campaigns deployed across platforms so as to inflame the emotional landscape around an audience's beliefs. AI systems will need to understand history, humor, symbolic reference, inference, subtlety, and insinuation. Only through such novel technologies will researchers be able to detect large-scale campaigns designed to use multimedia disinformation to amplify or magnify what a group of people feel about their preexisting beliefs.

This is a much more difficult task than simply identifying manipulated multimedia, particular words in a hate-speech lexicon, or new instances of a known "fake news" story. Rather this requires developing the capacity for machines and algorithms to better grasp the complex and multifaceted layers of human meaning making. Systems capable of parsing the complicated layers of meaning deployed in shallowfakes like memes represent the very cutting-edge of AI systems and are the forefront of the foreseeable future of technological development in this space. In many ways this represents the transition from existing AI perceptual systems to nascent AI cognitive systems. The enormous difference in complexity and computing power between these cannot be overstated.

Just the beginning

Systems capable of detecting deepfakes don't actually do much to help counter the proliferation of disinformation campaigns deploying shallowfakes designed to magnify and amplify an audience's preexisting beliefs. Likewise, systems focused on identifying problematic text also are inadequate to the task of parsing memes and shallowfakes. But imagine for a moment that AI researchers are able to develop semantic analysis

systems, and that it becomes possible to detect these coordinated disinformation campaigns as they occur. What then? AI won't be enough. The social media companies and policymakers will need to look at interventions, including software development, media literacy and education, and even new social norms. In other words, society needs a broad-spectrum approach to sufficiently prepare for the disinformation campaigns that are becoming increasingly common.

Such an approach should include the following elements: (a) policy-level responses that more carefully consider the complex relationship between disinformation, democracy, and freedom of speech; (b) information sharing agreements designed to coordinate the sharing of information across government agencies as well as social media platforms for the rapid identification and deceleration of disinformation campaigns in real time; (c) media literacy education campaigns that educate and prepare users to identify trustworthy sources of information and fact-check or further analyze sources of information that seem questionable.

Some responses to disinformation might involve not a technological fix to remove content, but rather techniques to help users know what they're consuming online. App developers should consider developing a "disinformation engagement metric" similar to screen-time counters and app-specific engagement statistics trackers. These would help users know more about the volume of disinformation that they are encountering. There are multiple hurdles related to this, but as the threat of emotional manipulation through disinformation continues to grow, policymakers and developers alike will need to develop new tools to help users navigate a rapidly evolving landscape.

As more of the human population gains reliable and fast access to the internet, an increasing percentage of people will become susceptible to campaigns intended to manipulate, magnify, and amplify their preexisting notions and emotional dispositions. Realizing when this is occurring doesn't just require technological systems capable of identifying deepfakes, but rather systems with the ability to identify coordinated shallowfake campaigns across platforms.

But beyond the necessary advances in technology, we also need a multi-faceted response that integrates policy-level decisions, content moderation strategies, information sharing agreements, the cultivation of new social norms around disinformation sharing online, the development of new disinformation consumption/interaction tools at the software-level, and even social initiatives intended to help people interact in civic life rather than just in an online forum.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Walter Scheirer was supported by the Notre Dame Institute for Advanced Study. Michael Yankoski and Tim Weninger were supported by USAID grant number 7200AA18CA00054. Portions of this work were supported by the Defense Advanced Research Projects Agency and the Air Force Research Laboratory under agreement number [FA8750-16-2-0173].

Notes on contributors

Michael Yankoski is a postdoctoral research associate at the University of Notre Dame. He is also a scholar of peace studies and ethics, and earned his PhD from the Kroc Institute for International Peace Studies at the University of Notre Dame.

Walter Scheirer is an associate professor in the Department of Computer Science and Engineering at the University of Notre Dame. His research is in the area of artificial intelligence, with a focus on visual recognition, media forensics, and ethics. He is also a Kroc Institute Faculty Fellow.

Tim Weninger is the Frank M. Friemann Collegiate Associate Professor of Engineering with the Department of Computer Science and Engineering at the University of Notre Dame, Notre Dame, IN, USA. His current research interests include the intersection of social media, data mining, and network science in which he studies how humans create and consume networks of information.

ORCID

Michael Yankoski http://orcid.org/0000-0002-9654-6993
Walter Scheirer http://orcid.org/0000-0001-9649-8074
Tim Weninger http://orcid.org/0000-0003-3164-2615

References

Buts, J. 2021. "How Anti-vaxx Memes Replicate through Satire and Irony." *The Conversation*, January 18. https://theconversation.com/how-anti-vax-memes-replicate-through-satire-and-irony-153018

Dawkins, R. 2016. *The Selfish Gene*. Oxford, UK: Oxford University Press.

Denham, H. 2020. "Another Fake Video of Pelosi Goes Viral on Facebook." *The Washington Post*, August 3. https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/

Eavis, P. 2021. "What Is GameStop, the Company, Really Worth? Does It Matter?" *The New York Times*, February 1. https://www.nytimes.com/2021/02/01/business/gamestop-how-much-worth.html

Flatley, D., and E. Wasson. 2021. "Congress Targets Stocks Mania as Ocasio-Cortez Rips Robinhood." *Bloomberg*,

January 28. https://www.bloomberg.com/news/articles/2021-01-28/brown-says-senate-panel-to-hold-hearing-amid-gamestop-frenzy

Funk, C., and A. Tyson. 2021. "Growing Share of Americans Say They Plan to Get a COVID- 19 Vaccine – Or Already Have." *Pew Research Center*, March 5. https://www.pewresearch.org/science/2021/03/05/growing-share-of-americans-say-they-plan-to-get-a-covid-19-vaccine-or-already-have/

Guarnera, L., O. Giudice, and S. Battiato. 2020. "DeepFake Detection by Analyzing Convolutional Traces." In Proceedings of the IEEE/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Virtual, June 14–19. https://openaccess.thecvf.com/content_CVPRW_2020/html/w39/Guarnera_DeepFake_Detection_by_Analyzing_Convolutional_Traces_CVPRW_2020_paper.html

Hwang, T., and C. Watts. 2020. "Deepfakes are Coming for American Democracy. Here's How We Can Prepare." *The Washington Post*, September 10. https://www.washingtonpost.com/opinions/2020/09/10/deepfakes-are-coming-american-democracy-heres-how-we-can-prepare/

Mittal, T., U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. 2020. "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues." In Proceedings of the 28th Association for Computing Machinery International Conference on Multimedia (MM '20). Virtual, October 12–16. https://dl.acm.org/doi/10.1145/3394171.3413570

Newmyer, T., and M. Zapotosky. 2021. "Wall Street Regulators Signal Tougher Approach to Industry after GameStop Frenzy." *The Washington Post*, February 14. https://www.washingtonpost.com/business/2021/02/14/sec-gamestop/

PeaceTech Lab. n.d. "PeaceTech Lab Lexicons." PeaceTech Lab. https://www.peacetechlab.org/toolbox-lexicons

Simonite, T. 2020. "What Happened to the Deepfake Threat to the Election?" *Wired*, November 16. https://www.wired.com/story/what-happened-deepfake-threat-election/

Technologies & International Development Lab. n.d. "Enhancing Civic Engagement, Strengthening Democracy, and Monitoring Elections. What Role Does Social Media Play in Elections?" Georgia Tech. http://tid.gatech.edu/dtd.html

Theisen, W., J. Brogan, P. B. Thomas, D. Moreira, P. Phoa, T. Weninger, and W. Scheirer. 2020. "Automatic Discovery of Political Meme Genres with Diverse Appearances." In Proceedings of the International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media (ICWSM). https://arxiv.org/abs/2001.06122

Verdoliva, L. 2020. "Media Forensics and DeepFakes: An Overview." *IEEE Journal of Selected Topics in Signal Processing* 14 (5): 910–932. doi:10.1109/JSTSP.2020.3002101.

Woolley, S. C., and P. N. Howard, eds. 2018. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford, UK: Oxford University Press.

Yahoo Finance. 2021. "GameStop Corp. (GME)" Yahoo Finance, March 24. https://finance.yahoo.com/quote/GME/history/

Yankoski, M., T. Weninger, and W. Scheirer. 2020. "An AI Early Warning System to Monitor Online Disinformation, Stop Violence, and Protect Elections." *Bulletin of the Atomic Scientists* 76 (2): 85–90. doi:10.1080/00963402.2020.1728976.