Comment for *Nature Human Behaviour*

First published in Nature Human Behaviour, volume 7, pages 1818–1821 (2023) by Springer Nature

Research can help to tackle AI-generated disinformation

Stefan Feuerriegel[1,2], Renée DiResta[3], Josh A. Goldstein[4], Srijan Kumar[5], Philipp Lorenz-Spreen[6], Michael Tomz[7,3], Nicolas Pröllochs[8]

1. LMU Munich School of Management, LMU Munich, Munich, Germany
2. Munich Center for Machine Learning (MCML), Munich, Germany
3. Stanford Internet Observatory, Stanford University, Stanford, CA, USA
4. Center for Security and Emerging Technology, Georgetown University, Washington, DC, USA
5. College of Computing at Georgia Institute of Technology, Atlanta, GA, USA
6. Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
7. Department of Political Science, Stanford University, Stanford, CA, USA
8. Department of Business and Economics, University of Giessen, Giessen, Germany

Corresponding author: Stefan Feuerriegel (feuerriegel@lmu.de), LMU Munich School of Management, LMU Munich, Munich, Germany

Artificial intelligence (AI) has made it easy to create realistic disinformation that is hard to detect by humans and may undermine public trust. Some approaches used for assessing the reliability of online information may no longer work in the AI age. We offer suggestions for how research can help tackle the threats of AI-generated disinformation.

Main text:

In March 2023, images of former President Donald Trump ostensibly getting arrested circulated on social media. Former President Trump, however, did not get arrested in March. The images were fabricated using generative artificial intelligence (AI) technology (https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/). While the phenomenon of fabricated or altered content is not new, recent advances in generative AI technology have made it easy to produce fabricated content that is increasingly realistic, making it harder for people to distinguish what is real.

Generative AI refers to technologies that can be used to create original content, such as text, images, audio, and video. While most applications of these tools are benign, there is significant concern about the potential for increased proliferation of disinformation (which we refer to broadly as content spread with the intent to deceive, including propaganda and fake news). Because the content generated appears highly realistic, some of the strategies presently used for detecting manipulative accounts and content are rendered ineffective by AI-generated disinformation.

How AI disinformation differs

What makes AI-generated disinformation different from traditional, human-generated disinformation? Here, we highlight four potentially differentiating factors: scale, speed, ease of use, and personalization. First, AI makes it easier to mass produce content for disinformation campaigns, which can translate into more false stories, multiple variations of the same story, presentation in different languages, automated conversational dialogues, and more. Second, compared to manual content generation, AI technology allows disinformation to be produced in seconds. These first two factors—scale and speed—are challenges for fact-checkers who will be flooded with disinformation but still need significant time for debunking. Third, as AI tools diffuse into society more broadly, they will lower the barriers to entry for running influence operations. People can use AI tools to create realistic fake images and videos without professional expertise or time-consuming manual editing. This may democratize the troll farm. Fourth, AI technology renders it possible to launch personalized disinformation campaigns to target-specific audiences (or individuals) and their preferences or beliefs without deep knowledge of the language or culture of the target. For example, personalized disinformation may target people of different ages, political ideologies, religious beliefs, and personality types (e.g., such as extroverts or introverts), which may increase the persuasiveness of disinformation campaigns. Those already marginalized by society or who have low media literacy may be particularly vulnerable.

The scalability of AI technology could enhance the tactics of disinformation campaigns. First, tactics involving highly targeted one-to-one communication (e.g., through bots or other automated tools) may become more common. For example, scammers may create generated audio content that sounds like a distressed family member to demand ransoms from targeted victims. Second, the scale of content production may augment tactics aimed at distracting audiences and at creating the illusion of majorities (as content appears to be coming from different sources). For example, state and state-linked actors, such as Russia’s Internet Research Agency, have long leveraged hundreds of accounts to divert attention from inconvenient stories, which will now be easier with generative AI. Third, creating back-and-forth conversations in real time may help obscure the automated nature of corresponding social media accounts.

It is important to note that the relevant threat vectors are broader than social media: AI tools enable low-cost and high-volume fabrication of email campaigns[1], paid advertisements, websites, or scientific documents that provide false evidence for claims. At the extreme, the deluge of AI-generated disinformation may make it more difficult to discern the truth from the noise in online spaces, reducing broader societal trust.

AI disinformation is hard to detect

Existing research shows that generative AI systems can write disinformation that is hard to detect by humans because it can mimic the style of reliable sources and communicators that are trusted by the target audience[2,3]. Research is needed to explore the vulnerabilities of individuals to AI-generated disinformation campaigns and, specifically, the extent to which capabilities such as microtargeting and one-on-one chats make disinformation campaigns more credible and persuasive. For example, future research could compare differences in behavioral outcomes between persuasive and distractive campaigns with AI-generated disinformation.

Endogenous cues that humans have previously used for judging the reliability of information (e.g., whether a text is free of grammatical errors, whether images have lighting and shadows consistent with reality) may no longer be good determinants. As a result, the importance of non-content-based, exogenous cues may increase. These cues include indications of who wrote or created content, whether it bears a watermark, how well a piece of information is connected to existing knowledge (e.g., by references to known sources), or how other people in one's network interact with it (e.g., the diversity of the readership is associated with quality)[4]. We believe that one high-value research direction is identifying and testing exogenous cues that are transparent and difficult to game with generative AI.

Behavioral mitigations

To mitigate the effects of disinformation campaigns, behavioral interventions can be employed which involve nudging the behavior of online users or boosting their competencies so they are less likely to believe or share fake content. However, it is not clear how effective previously-designed behavioral interventions will be in tackling AI-generated disinformation. For reasons discussed above, behavioral interventions based on endogenous cues such as asking people to check whether images look realistic before sharing them or asking people to check for professional writing quality may be insufficient, since generative AI can fabricate images that look authentic and text that is largely error-free. Rather, behavioral interventions aimed at exogenous cues will grow more important. We see at least three categories of behavioral interventions for tackling AI-generated disinformation.

First, adding more reliable, exogenous cues that can indicate the source and its quality may help people assess the accuracy of information. Examples include adding flags for trustworthy sources (e.g., a badge for verified users) or adding labels to AI-generated content that can warn or inform users. However, this raises several challenges, including which users to verify, how to appropriately label content without surfacing false-positives, and how to detect that content is

AI-generated in the first place. An alternative approach would be to mark the source of content upon creation. For example, some users who create AI-generated content, such as the Trump arrest photos, might voluntarily flag their own output as AI-generated to avoid future decontextualization (deliberate or otherwise). Malicious actors, however, are unlikely to follow suit.

Second, leveraging collective intelligence can be effective in addressing not only human-generated but also AI-generated disinformation. Platforms could display social cues, for instance, that provide feedback on how many people actually shared a text or what amount of time people spent reading a text compared to the overall viewership. Such social cues are generally easy to implement and are already available for many platforms. However, they, too, have limitations, as AI-generated audio, images, and videos are produced with the intent to capture attention, and adversarial actors often incorporate inauthentic engagement with their content to manipulate this very method of gauging legitimacy. Alternatively, crowdsourcing can be used to assess content that is potentially misleading as in the case of Twitter's Community Notes, which places user-generated labels on certain pieces of content. The success of such crowdsourced labels depends on the efforts of users, which may be low and thus require greater public buy-in or a platform-led incentive structure.

Third, boosting competencies can be effective in helping users evaluate or verify information without relying on a single source. Potentially valuable examples include media literacy, psychological inoculation, critical ignoring, and lateral reading[5]. In particular, media literacy training will need to be adapted to the AI era. For example, one could train people in the basics of how generative AI models work or teach distinctive characteristics in their output, so that they better understand the potential risk of AI-generated disinformation. Yet, it is unclear whether this would increase general skepticism or actually improve truth discernment and whether the AI-specific characteristics will remain consistent across models and as generative AI technology improves.

In sum, the combination of exogenous cues, which contribute to an online environment that can not be gamed as easily, and the development of competencies that rely on multiple sources, can provide a toolbox of behavioral interventions. Still, there are practical limitations behind the above behavioral interventions as the implementation generally depends on the willingness and capacity of social media platforms, which vary widely. Likewise, research is needed to understand how effective behavioral interventions will be in the era of AI-generated disinformation and in comparison to existing strategies for increasing resilience to disinformation more generally.

Regulatory and technological mitigations

On the regulatory side, effective online governance is required that balances the importance of free speech against harms that may arise from AI-generated disinformation. Countries have begun developing relevant legal frameworks that prohibit or restrict the use of AI for generating content that is deceptive or manipulative[6], though it is not clear how effective such regulation will be. Likewise, there are regulatory discussions that seek to limit the size of generative AI models. However, such regulations may be sidestepped by malicious actors, and current generative AI models may already be sufficient for producing high-quality content.

AI tools have also been developed that can be used to identify and flag AI-generated content in general or AI-generated disinformation in particular. However, these tools have notable shortcomings, such as surfacing false-positives. As generative AI models evolve, they will probably outpace detection tools, creating a continuous cat-and-mouse game between generators and detectors. This highlights the need for continuous research and updates to keep detection tools effective and accurate—and to be aware when they are no longer reliable.

Technical solutions involve, for example, usage restrictions or fact-sensitive AI models as complements to regulation[7]. Further, watermarking could ease detection by embedding hidden signals in content to indicate that the content was produced by AI; these machine-readable signals enable social media platforms, which often serve as distribution sites, to recognize the content as AI-generated. Companies that produce AI tools for image and video generation could incorporate watermarks voluntarily, as a form of self-regulation, even if not required by law. However, there are enormous challenges behind coordinating watermarks across multiple digital platforms and across a growing number of AI tools (e.g., through standardization).

Another technical solution is algorithmic amplification. While AI-generated disinformation may be more persuasive, its effects are partially determined based on its reach. Depending on the extent to which algorithmic curation focuses on endogenous and exogenous cues, particularly as watermarking and provenance efforts progress, algorithmic curation may amplify or deamplify the exposure to AI-generated disinformation. To this end, algorithmic curation that rewards cues of epistemic quality is desirable. For example, if the diversity of the readership is considered as a cue for algorithmic curation[8], it could make it hard for disinformation to flourish—whether it is generated by AI or not.

Finally, research could enable AI technology itself to help write counter-responses to disinformation (e.g., a correction to fake news)[9], which could be shared on social media (e.g., via bots or other automated tools) and may help prevent users from falling for or sharing it. However, more research from behavioral science is needed to understand the effectiveness and ethics of AI-generated counter-responses, as well as the potential risks. A potential benefit of such counter-responses is that they are scalable, work in real-time, and can be deployed by both social media platforms and third parties (e.g., non-profit organizations or journalists) that seek to promote the accuracy of content posted on digital platforms.

Call for research

Generative AI technology has many positive applications, yet a negative externality is the democratization of disinformation content production: increasing its volume, velocity, and potential persuasiveness while decreasing its cost. As we argue above, AI-generated disinformation challenges existing detection and mitigation strategies employed by platforms and humans alike, and, therefore, we call for more research to update detection and mitigation strategies in light of AI-generated disinformation (Figure 1). Eventually, to inform policy responses, we need rigorous, causal evaluations of how AI-generated content affects perceptions and behavior, and of how detection and mitigation strategies could help.

References

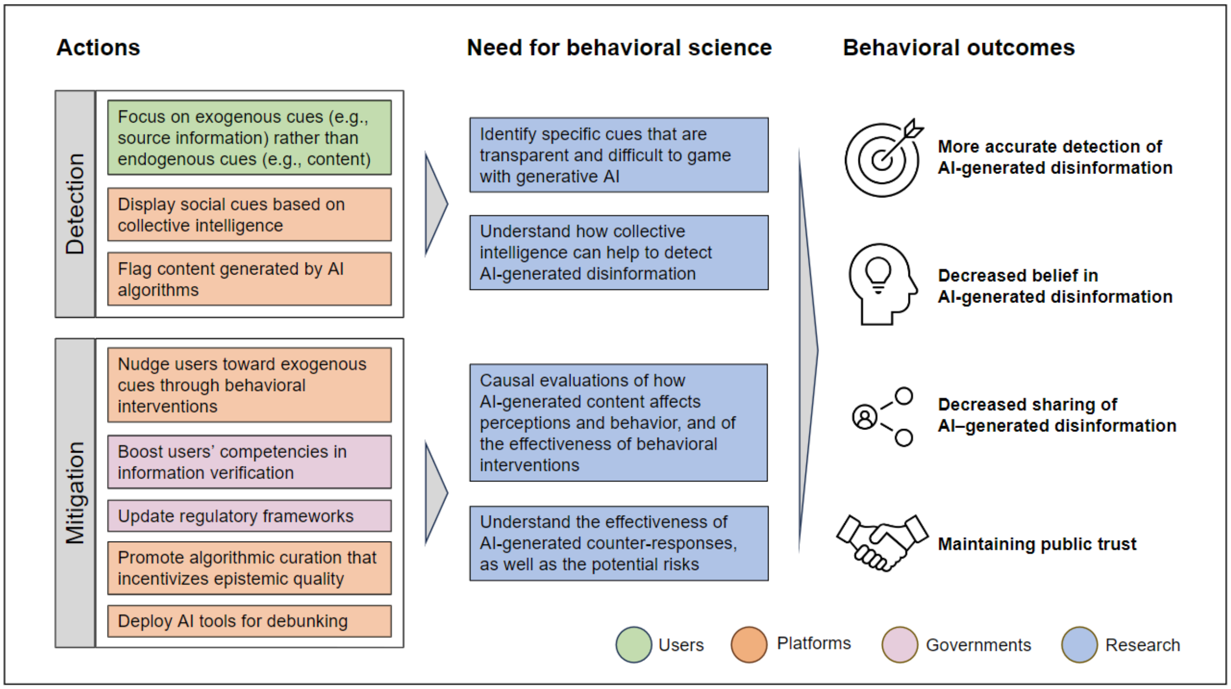
1. Kreps, S., & Kriner, D. L. The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. *New Media & Society* (2023).
2. Spitale, G., Biller-Andorno, N. & Germani, F. AI model GPT-3 (dis)informs us better than humans. *Sci. Adv.* **9**, eadh1850 (2023).
3. Kreps, S., McCain, R. M. & Brundage, M. All the news that’s fit to fabricate: AI-generated text as a tool of media misinformation. *J. Exp. Political. Sci.* **9**, 104–117 (2022).
4. Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nat. Hum. Behav.* **4**, 1102–1109 (2020).
5. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest* **21**, 103–156 (2020).
6. Hine, E., & Floridi, L. New deepfake regulations in China are a tool for social stability, but at what cost? *Nat. Mach. Intell.* **4**, 608–610 (2022).
7. Goldstein, J. A., et al. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv:2301.04246* (2023).
8. Bhadani, S., Yamaya, S., Flammini, A., Menczer, F., Ciampaglia, G. L. & Nyhan, B. Political audience diversity and news reliability in algorithmic ranking. *Nat. Hum. Behav.* **6**, 495–505 (2022).
9. He, B., Ahamad, M. & Kumar, S. Reinforcement learning-based counter-misinformation response generation: A case study of COVID-19 vaccine misinformation. *ACM The Web Conference* (2023).

Competing interests

The authors declare no competing interests.

Display item

Figure 1. Behavioral science can promote detection and mitigation strategies for humans in tackling AI-generated disinformation.



Actions

Detection

- Focus on exogenous cues (e.g., source information) rather than endogenous cues (e.g., content)
- Display social cues based on collective intelligence
- Flag content generated by AI algorithms

Mitigation

- Nudge users toward exogenous cues through behavioral interventions
- Boost users' competencies in information verification
- Update regulatory frameworks
- Promote algorithmic curation that incentivizes epistemic quality
- Deploy AI tools for debunking

Need for behavioral science

- Identify specific cues that are transparent and difficult to game with generative AI
- Understand how collective intelligence can help to detect AI-generated disinformation
- Causal evaluations of how AI-generated content affects perceptions and behavior, and of the effectiveness of behavioral interventions
- Understand the effectiveness of AI-generated counter-responses, as well as the potential risks

Behavioral outcomes

- More accurate detection of AI-generated disinformation
- Decreased belief in AI-generated disinformation
- Decreased sharing of AI–generated disinformation
- Maintaining public trust

● Users ● Platforms ● Governments ● Research