Newhouse Impact Journal

Volume 1 | Issue 1 Article 5

2024

In a Digital World of Generative AIDetection Will Not be Enough

Jason Davis
Syracuse University

Recommended Citation
Davis, Jason (2024) "In a Digital World of Generative AIDetection Will Not be Enough," *Newhouse Impact Journal*: Vol. 1: Iss. 1, Article 5.
DOI: http://doi.org/10.14305/jn.29960819.2024.1.1.01
Available at: https://surface.syr.edu/newhouseimpactjournal/vol1/iss1/5

In a Digital World of Generative AI Detection Will Not be Enough

Jason Davis
Syracuse University, USA

Recent and dramatic improvements in artificial intelligence (AI) driven by large language models (LLM), image generators, audio, and video have fed an exponential growth in generative AI applications and accessibility.[1] The disruptive ripples of this rapid evolution have already begun to fundamentally impact how we create and consume content on a global scale.[2] And while the use of generative AI has and will continue to enable massive increases in the speed and efficiency of content creation, it has come at the cost of uncomfortable conversations about transparency and the erosion of digital trust.[3]

While the problem of mis and disinformation is not new, the speed and scale at which it currently propagates is. Over the last decade the rise of social media platforms and their development of hugely successful algorithmic amplification strategies has led to increasing challenges associated with the damaging spread of fake news.[4] While the use of these platforms and their ability to automate the targeting and distribution of mis and disinformation has already created significant damage, they have remained limited in some sense by the need for human content creation. The addition of generative AI offers the potential to remove this human capacity driven bottleneck and magnify the problem exponentially.

The resulting digital landscape becomes one in which a significant portion of all content is fully, or at least partially synthetic and existing detection strategies designed to use binary classifications (i.e. synthetic or human generated) offers rapidly diminishing down selection value in terms of threat evaluation and human actionable response. In a recent study by Europol "experts estimate that as much as 90 percent of online content may be synthetically generated by 2026."[5] Even this staggering numerical transformation represents an oversimplification which still assumes a binary distribution of media. In reality, this increasing digital ocean will not remain some bimodal distribution of wholly real or synthetic media but rather a continuum of hybridized content that stretches fully between the two.

To address the rapidly shifting mix of digital content a more holistic approach is required where detection remains an important part of digital tool development to combat mis and disinformation but is insufficient in and of itself. Rather, a more comprehensive approach is required that includes **D**etection (is it synthetic or human generated), **A**ttribution (is it coming from the source it says it's coming from or if synthetic, which generative tool or model was used to create it), and **C**haracterization (is it malicious or benign, what is the intent, who is the target audience).[6] These two additional layers represent progressively more challenging tasks for AI/ML driven tools, so it is not surprising that simple binary detection strategies that focus primarily on leveraging statistical differences in non-human readable meta data, pixel level artifacts or token prediction remains the current focus of most tool development.

The complexity of this task grows even higher when considering the paths to user adoption

for these tools by both trained human analysts making critical national security decisions about a potential threat or the general public as simple consumers of digital content. Both will require human interpretable evidence that can support AI/ML driven recommendations or conclusions. Otherwise, the result is a black box problem being solved with a black box solution and requires humans to relinquish any role in the decision-making process by placing unconditional trust in any detection system that is applied. This represents a challenging approach if the goal is to achieve rapid and widespread adoption in a population where digital trust is already at a premium. Given the rapid rate at which generative AI and its applications are advancing, approaches requiring gradual validation with a secondary focus on the development of trust in digital transparency tools suggests an ever-widening gap. If this is the case, short of the global throttling of the development of Generative AI capabilities, the co-development of AI/ML driven digital transparency safeguards and human explainable evidence frameworks becomes critical.[7] The alternative is a black box arms race between generative AI and its detection which, at best, represents a partial solution that gets us no closer to advancing protections for the public against malicious disinformation content. To have any chance at actually diminishing the societal impact of digital disinformation in an age of generative AI, approaches strategically designed to assist human decision making must move past simple detection and provide more robust solutions.

To achieve this, algorithm developers must consider beyond what model driven patterns can be exploited using how much and which training data to achieve the required statistical thresholds for detection, attribution, and characterization determinations. They must also consider how to integrate, communicate, and display these various AI/ML decision making processes to human users so they can retain some ability to self-validate any particular conclusion.

References

1 Black, J., & Fullerton, C. (2020). Digital deceit: fake news, artificial intelligence, and censorship in educational research. Open Journal of Social Sciences, 8(07), 71.; Gozalo-Brizuela, Roberto, and Eduardo C. Garrido-Merchán. "A survey of Generative AI Applications." arXiv preprint arXiv:2306.02781 (2023).; Gozalo-Brizuela, R., and Garrido-Merchan, E.C. Chatgpt is not all you need. A state of the Art Review of Large Generative AI Models, 2023. https://doi.org/10.48550/arXiv.2301.04655

2 Yongqiang Ma, Jiawei Liu, Fan Yi, (2023) Is This Abstract Generated by AI? A Research for the Gap between AI-generated Scientific Text and Human-written Scientific Text. https://www.researchgate.net/publication/367431889_Is_This_Abstract_Generated_by_AI_A_Research_for_the_Gap_between_AI-generated_Scientific_Text_and_Human-written_Scientific_Text; Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, Xiaozhong Liu. (2023) AI vs. Human--Differentiation Analysis of Scientific Content Generation. https://doi.org/10.48550/arXiv.2301.10416

3 Wardle, C. (2017, February 17). Fake news. It's complicated. Medium. Retrieved from https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79; Abu Arqoub, O., Abdulateef Elega, A., Efe Özad, B., Dwikat, H., & Adedamola Oloyede, F. (2022). Mapping the scholarship of fake news research: A systematic review. Journalism Practice, 16(1), 56-86.

4 Yang, J. & Luttrell, R. (2022). Digital misinformation & disinformation: The global war of words. In J.H. Lipschultz, K.Freberg, and R. Luttrell, (Eds.), The emerald handbook of computer-mediated communication and social media. Emerald Publishing Limited, Bingley, pp. 511-529. https://doi.org/10.1108/978-1-80071-597-420221030; X. Zhou, R. Zafarani (2020) A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53 (5) (2020), pp. 1-40

5 Europol Innovation Lab: FACING REALITY? LAW ENFORCEMENT AND THE CHALLENGE OF DEEPFAKES, 2022. DOI: 10.2813/08370; Schick, Nina, Deepfakes: The Coming Infocalypse: What You Urgently Need To Know, Twelve, Hachette UK, 2020.

6 Luttrell, R., Davis, J. Smith, P., & Welch, C. "Authenticity in Synthetic Media: The Theory of Content Consistency and AI Algorithmic Multi-Modal Fake News Detection" in progress, October 15, 2023. Syracuse University.; Davis, J. Luttrell, R., Smith, P. & Hong, N. "Unmasking Manipulated Media." Presentation presented at AEJMC, August, 2023, Washington, D.C.

7 Yang, J. & Luttrell, R. (2022). Digital misinformation & disinformation: The global war of words. In J.H. Lipschultz, K.Freberg, and R. Luttrell, (Eds.), The emerald handbook of computer-mediated communication and social media. Emerald Publishing Limited, Bingley, pp. 511-529.