STATEMENT ON PRINCIPLES FOR THE DEVELOPMENT AND USE OF SYSTEMS TO DETECT GENERATIVE AI CONTENT

The ACM US Technology Policy Committee (USTPC)[1] notes that the dramatic increase in the availability, proliferation, and use of generative artificial intelligence technology in all sectors of society has created concomitant growing demand for systems that can reliably detect when a document, image, or audio file contains information produced in whole or in part by a generative AI system. Specifically, for example:

- Educational institutions want systems that can reliably detect when college applications and student assignments were created with the assistance of generative AI systems;
- Employers want systems that can detect the use of generative AI in job applications;
- Media companies want generative AI systems so that they can distinguish human comments on their articles from responses generated by chatbots; and
- Government agencies need to tell human letters and comments from responses that were algorithmically generated.

Demand for such systems, however, is no measure of their accuracy[2] or fairness.[3] Indeed, the Committee finds and cautions that ***no such presently available detection technology is sufficiently reliable on which to exclusively base critical, potentially life- and career-altering decisions*** in the contexts and use cases cited above, or any other. Accordingly, while AI detection systems may provide useful preliminary assessments, their outputs should never be accepted as proof of AI-generated content.

---

[1] The Association for Computing Machinery (ACM), with more than 100,000 members worldwide, is the world's largest educational and scientific computing society. ACM's US Technology Policy Committee (USTPC), currently comprising more than 200 members, serves as the focal point for ACM's interaction with all branches of the U.S. Government, the computing community, and the public on policy matters related to information technology. This statement's principal author for USTPC is Simson Garfinkel, Chair of the Committee's Digital Governance Subcommittee. Primary additional contributors include Committee members Houssam Abbas, Andrew Appel, Harish Arunachalam, Ricardo Baeza-Yates, David Bauman, Ravi Jain, Carl Landwehr, Larry Medsker, Neeti Pokhriyal, Arnon Rosenthal, and Marc Rotenberg.

[2] USTPC notes that the apparent accuracy of some detectors heralded in early 2023 is unlikely to be sustainable as AI systems evolve and multiply. Indeed, Open AI discontinued its AI writing detector in July 2023, less than six months after its initial release, due to the program's "low rate of accuracy." "OpenAI discontinues its AI writing detector...," *Ars Technica* (July 26, 2023).

[3] See, *e.g.*, Liang, W., *et al*, "GPT detectors are biased against non-native English writers," *ScienceDirect* (July 14, 2023).

ACM Technology Policy Office
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/publicpolicy

Technical Context

Generative AI detection systems are not now, and are unlikely in the foreseeable future, to be able to detect AI-generated content with sufficient accuracy to be fairly relied upon in matters of high risk or consequence, such as determining and penalizing cheating academically or falsifying a resume. This is true for a number of reasons:

1) There are fundamentally two approaches for identifying computer-generated output. The first is to embed a signal, called a "watermark," in the output for later detection by a specialized reader. The second approach is to build a detector that assesses a random file to rate the *likelihood* that it was produced by a generative process. Although it is possible to embed watermarks that are hard for a human to notice yet easy for a machine to detect, there are currently no watermarks that cannot be readily removed from text or images through straightforward editing or image manipulation.[4]

2) Reliably detecting the output of generative AI systems without an embedded watermark is beyond the current state of the art, which is unlikely to change in a projectable timeframe. That is true because that objective is and will remain a perpetual "moving target." While it may be possible to create a system that detects the output of a *specific* generative AI system, a follow-on system could avoid detection by repeatedly making subtle changes to the data until the use of AI is no longer detectable.

3) False negatives also are a problem. Cutting, pasting, and editing AI-generated output will reduce the likelihood that an AI detector will recognize such output. Users sufficiently clever to use these techniques will be able to avoid detection.[5]

4) Finally, AI detectors operate as "black boxes." Developers and vendors offering AI detection services often do so with proprietary tools not subject to independent analysis at any point in their development cycle or after deployment. Although such tools provide an assessment based on the input received, the techniques used by the tool to make those determinations remain opaque as its results cannot be proven or its "logic" tracked.[6]

In light of the foregoing — as generative AI detection technologies are increasingly employed in education, the workplace, media, and government — every effort by designers, developers, vendors, and institutions using such technology must be made, consistent with ACM's Code of

---

[4] Roger Montti, "How the ChatGPT Watermark Works and Why It Could Be Defeated," *Search Engine Journal* (December 30, 2022) [https://www.searchenginejournal.com/chatgpt-watermark/475366/#close]

[5] Doraid Dalalah, Osama M.A. Dalalah, "The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT," *The International Journal of Management Education*, Volume 21, Issue 2, 2023. https://doi.org/10.1016/j.ijme.2023.100822.
[https://www.sciencedirect.com/science/article/pii/S1472811723000605]

[6] Saurabh Bagchi, "What is a black box? A computer scientist explains what it means when the inner workings of AIs are hidden," *The Conversation*, May 22, 2023. https://theconversation.com/what-is-a-black-box-a-computer-scientist-explains-what-it-means-when-the-inner-workings-of-ais-are-hidden-203888

ACM Technology Policy Office
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/publicpolicy

Ethics and Professional Conduct stipulates Responsibility,[7] to "avoid harm." To that end, USTPC offers these guiding principles and recommendations:

Principles and Recommendations

- The use of systems for detecting AI-generated images and other media that automatically flag submissions for rejection should be acceptable only if such detection systems have an exceedingly low risk of false rejections and provided that a human-driven appeal process is provided.[8]
- It is generally not appropriate to automatically reject textual submissions in high-stakes circumstances that are classified as being produced by a generative AI system, even if a process for appealing such rejections is provided. While some AI detection systems may be useful to produce preliminary assessments of high-stakes submissions, humans must always be the final arbiters of their accuracy. Examples of high-stakes submissions include (but should not be limited to) classroom assignments, and applications for admission to an educational institution, credit, or employment.[9]
- Entities using generative AI detection systems should adopt guidance – such as codes of conduct, employee handbooks, and enforceable honor codes – requiring those affiliated with the entity to comply with the AI policies of the organization. Violations of such codes should result in appropriate sanctions.
- Consistent with past USTPC statements,[10] individuals should have the opportunity to contest outcomes whenever an adverse decision about them is made, in whole or in part, in reliance upon the output of an AI system.[11] This should be especially true for AI systems that purport to detect the use of generative AI. Organizations should be prepared to establish an appeal process for any proposed deployment of AI systems that could produce adverse consequences.
- Human content evaluators should, on an ongoing basis, be provided with appropriate training on the right methods and tools to employ to validate submitted content.
- Increased public and private sector funding for research on how to develop better detection mechanisms, conduct impact analyses, perform user research, and related matters would be prudent and beneficial.

---

[7] See https://www.acm.org/code-of-ethics.

[8] An acceptable system might be based, for example, on the detection of durable watermarks.

[9] When the labor available to read and assess submissions is insufficient to handle the volume of submitted material, there may be appropriate uses of AI to assist human readers. Among the appropriate uses are AI tools to detect and flag incoherent, irrelevant, or nonsensical material, or to sort submissions by similarity or topic. Such tools focus on content and not on apparent authorship, and they annotate without making decisions about disposition.

[10] *Joint Principles for the Development, Deployment, and Use of Generative AI Technologies*, ACM Technology Policy Council, Europe/US Technology Policy Committees (June 27, 2023) [https://www.acm.org/binaries/content/assets/public-policy/ustpc-approved-generative-ai-principles]

[11] Indeed, the European Union's GDPR goes further to require that individuals adversely affected by an automated decision have a right to human review of that determination.

ACM Technology Policy Office
1701 Pennsylvania Ave NW, Suite 200
Washington, DC 20006

+1 202.580.6555
acmpo@acm.org
www.acm.org/publicpolicy