NEWS RELEASE

Contact: Jim Ormond
ACM Media Relations
212-626-0505
ormond@hq.acm.org

Adam Eisgrau
ACM Director of Global Policy and Public Affairs
202-580-6555
eisgrau@hq.acm.org

Can We Ensure That Systems for Detecting Generative AI Are Accurate and Fair?

Influential Association of Computer Scientists and Software Engineers Issues Statement on Principles for the Development and Use of Systems to Detect Generative AI Content

New York, NY, October 18, 2023 – With the public interest in generative AI technologies increasing every day, some of the most pressing issues revolve around questions such as "Is generative AI being used to create fake images and videos?" and "Are students using generative AI to write papers and cheat on exams?" For these reasons, there is growing demand for systems that can detect whether an image, audio file, or written work has been created by a human or an AI system.

Although AI detection systems are beginning to proliferate, there have been no industry standards or government regulations to make sure these systems are accurate or fair. Because the impact of these systems on individuals can be so significant, the Association for Computing Machinery's US Technology Policy Committee (ACM USTPC) has issued a "Statement on Principles for the Development and Use of Systems to Detect Generative AI Content."

The introduction to the new USTPC Statement highlights various scenarios in which systems to detect generative AI developed content would be desirable. For example, employers wanting to know if generative AI was used to fill out a job application, or media companies trying to determine if comments posted on platforms were left by humans or chatbots.

At the same time, the Statement notes that "Demand for such systems, however, is no measure of their fairness or accuracy." The committee goes on to explain that "no such presently available detection technology is sufficiently reliable on which to exclusively base critical, potentially life- and career-altering decisions…"

The statement provides a technical context as to why the fairness and accuracy of existing generative AI detection systems cannot be guaranteed, and sets out six specific principles and recommendations:

- **Low risk of false rejections and human-driven appeal process:** 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.
- **High stakes submissions**: 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. 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.
- **Codes of conduct**: 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.
- **Contesting outcomes**: Consistent with past USTPC statements, 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.
- **Appropriate training**: Human content evaluators should, on an ongoing basis, be provided with appropriate training on the right methods and tools to validate submitted content.
- **Increased funding**: 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.

"In principle, detecting generated text and images that are generated by AI is an open-ended problem," explained Simson Garfinkel, lead author of the statement and Chair of the USTPC Digital Governance Subcommittee. "Although it might be possible to build a system that can detect *today's* AI-generated content, such a detector could be used to build tomorrow's AI generation system that evades such detection. This statement is released to add a voice of technical expertise to the moral panic over the use of generative AI. We are saying that text and images produced by generative AI systems cannot be reliably detected today. We also encourage all institutions to defer from deploying systems that purport to automatically detect and discard materials because they were allegedly created by a generative AI system."

"This new Statement is part of an ongoing series that the ACM US Technology Committee publishes to inform the public about new technologies and their impacts on society," added Larry Medsker, Chair, ACM US Technology Policy Committee. "Recently, we have been especially active in giving timely input to address new developments in AI. In this vein, USTPC members have published "Principles for the Development, Deployment and Use of Generative AI," and "Statement on Principles for Responsible Algorithmic Systems." All of our policy products are available online and our members are available to lend their expertise in policy forums or to the media when needed."

In addition to lead author Simson Garfinkel, primary additional contributors to the "Statement on Principles for the Development and Use of Systems to Detect Generative AI Content" 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.

About the ACM US Technology Policy Committee

ACM's US Technology Policy Committee (USTPC) serves as the focal point for ACM's interaction with all branches of the US government, the computing community, and the public on policy matters related to information technology. The Committee regularly educates and informs Congress, the Administration, and the courts about significant developments in the computing field and how those developments affect public policy in the United States.

About ACM

ACM, the Association for Computing Machinery, is the world's largest educational and scientific computing society, uniting computing educators, researchers, and professionals to inspire dialogue, share resources, and address the field's challenges. ACM strengthens the computing profession's collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence. ACM supports the professional growth of its members by providing opportunities for life-long learning, career development, and professional networking.

###