The Computing Community Consortium's Response to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Written by: David Danks (University of California, San Diego), Catherine Gill (Computing Community Consortium), Daniel Lopresti (Lehigh University), Rajmohan Rajaraman (Northeastern University), Michela Taufer (University of Tennessee, Knoxville), Ufuk Topcu (University of Texas, Austin), Matthew Turk (Toyota Technological Institute at Chicago), and Holly Yanco (University of Massachusetts, Lowell).

This response is from the Computing Research Association (CRA)'s Computing Community Consortium (CCC). CRA is an association of nearly 250 North American computing research organizations, both academic and industrial, and partners from the professional societies. The mission of the CCC is to bring together the computing research community to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges.

Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations or of the National Science Foundation, which funds the CCC through a cooperative agreement.

Assignments 1 and 2:

Artificial Intelligence technologies and applications are advancing so quickly, and the widespread adoption of generative AI is so new, that expecting current technological solutions to manage their risks and other shortcomings is unrealistic. Reliable content authentication is a complex problem, and there are no simple one-size-fits-all solutions. Understanding that, in the most general sense, it is impossible to detect AI generated content consistently and accurately is an important step in creating a practical framework for agencies to follow. Research in this area will certainly advance, and AI detection tools may continue to improve, but it is our opinion that depending on traditional software testing paradigms to detect generated content would be unwise. Even as detection tools improve, AI generation tools will improve as well, so detection will always be playing catch-up, very much like the "cat and mouse" games we have seen with cybersecurity for decades. Given the rapid advances and changes in the development of AI, it is important that NIST considers approaches and solutions that are

flexible, responsive, and adaptable in the years to come. As experts in the field, we are highly skeptical about claims regarding near-term technical solutions to these very challenging problems.

We encourage evaluating technologies that are proposed for mitigating the risks of generative AI (watermarking programs, programs to detect content generation, etc.) with measured caution. These techniques are all young, relatively untested, and constantly evolving, and so are still error prone. Independent testing for biases and identifying edge cases needs to be done before any of these programs can be considered reliable. OpenAI themselves discontinued their AI detection tool because its accuracy rate was so low.[1] We are not aware of any detection tools for AI generated content that are anywhere near accurate enough to be considered trustworthy.[2] Watermarking technologies and other technologies that add metadata to documents also have faults. For instance, current research indicates that watermarking and metadata can both be removed from documents or graphics, or forged onto content.[3] NIST likely already has plans to do so, but we encourage involving the Cryptography and Cybersecurity research communities in developing and evaluating authentication mechanisms for AI generated content.

Instead of relying only on technology, for the time being we also need to rely on the humans who are deploying, using, and regulating AI technologies. Limits should be placed on when it is considered acceptable to use a generative AI program. Users should also be made aware of the risks associated with generative AI, as well as situations where it is likely to fail. Users who are trained on warning signs will be more likely to notice bias in systems and maliciously generated content (deep fakes, doctored documents, etc.). Looking for models to adapt from elsewhere, we note that significant time and investment have been put into cybersecurity training for end users to limit the number of scenarios in which they expose themselves and their organizations to ransomware and other cybersecurity risks. We recommend that the lessons learned from cybersecurity be researched, modified, and extended to generative AI, and that generative AI training be supported and encouraged across US businesses and government agencies. Regular training will make users more comfortable and secure when using AI and generative AI systems.

---

[1] https://arstechnica.com/information-technology/2023/09/openai-admits-that-ai-writing-detectors-dont-work/

[2] https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00146-z

[3] https://eprint.iacr.org/2023/1776

We regard education as critical to successfully addressing the challenges we will face with AI. Effective workforce development should build on the design and deployment of comprehensive and interdisciplinary curricula in high schools, community colleges, and higher education institutions, ensuring the education of professionals capable of leading in developing, deploying, and governance of AI technologies. The curriculum should combine technical AI and machine learning knowledge with ethics, legal studies, data governance, and cybersecurity. The curriculum should be modular and provide a foundation in AI and ML principles, advanced systems, and practical application skills while integrating ethical considerations, legal frameworks, and policy analysis to navigate the complex implications of AI in society. It must also include data ethics and management modules, emphasizing the importance of ethical data use, privacy, and security. Real-world challenges (through project-based learning, case studies, and industry internships) should complement theoretical knowledge, promoting an understanding of AI's social impacts and ethical dilemmas. It will also be important to distinguish where existing methodologies break down, so that we are not misled by a false sense of confidence in AI systems. This will include teaching the limits of pre-deployment testing and the emergence of AI bad behavior after the fact, emphasizing the need for ongoing evaluation and oversight of fielded AI systems.

For this effort to be beneficial in the long term, a standardized feedback mechanism should be established so that government agency employees and the public can share feedback with NIST when they notice bias or are adversely affected by an AI system. A national AI incidence database should be established to track these reports to help aid AI system development down the line. This database and the organization supporting it could be similar to the CERT technical division at Carnegie Mellon University. The CERT division works to “improve the security and resilience of computer systems and networks” and handles a wide variety of cybersecurity threats.[4] To do this, CERT partners with government, law enforcement, industry, and academia to thwart cyberattacks. The CERT website also allows users to report vulnerabilities, which are added to their databases and investigated.

Evaluating every AI and generative AI system that every government agency will employ and every user will report to, however, would be an enormous undertaking. Given the monumental nature of this task, we recommend that NIST utilize all available resources whenever possible. We believe the US Artificial Intelligence Safety Institute Consortium would be a good resource for this purpose. The AI Safety Institute

[4] https://www.sei.cmu.edu/about/divisions/cert/index.cfm

Consortium could also be very helpful in testing out different programs and processes that will be suggested to NIST in these RFI responses.

Assignment 3:

We believe it is vitally important to choose accurate and descriptive terms when referring to AI. Poor term designation can cause policy makers and the public to become confused or to dismiss important concerns. The term “hallucination” for instance is an example of poor appellation. The term “hallucination” in AI refers to when an AI program generates a response which is false or misleading, and often includes inventing false content. This term is dangerous because it is vague and misleading: AI does not hallucinate, it fails and gives inaccurate answers. The term can also add to the tendency to anthropomorphize AI, which can create improper assumptions and expectations. We recommend that NIST initiates a focused dialogue with computing experts to identify similar terms across languages in referring to AI so that important conversations can take place using a shared understanding of these complex technical concepts.