acm U.S. Technology Policy Committee

July 7, 2023

RESPONSE TO REQUEST FOR INFORMATION ON NATIONAL AI PRIORITIES BY THE WHITE HOUSE OFFICE OF SCIENCE AND TECHNOLOGY POLICY[*]

The Association for Computing Machinery (ACM) is the longest-established and — with more than 50,000 American members — the largest association of individual professionals engaged in all aspects of computing in the nation. A non-lobbying and otherwise wholly apolitical organization, ACM's mission includes providing unbiased, expert technical advice to policymakers on matters of our members' wide-ranging expertise. That work is accomplished in the United States by and through ACM's U.S. Technology Policy Committee (USTPC).

USTPC commends its most recent products concerning artificial intelligence to OSTP for its general consideration. They are its Principles for the Development, Deployment, and Use of Generative AI Technologies (June 2023), *ACM TechBrief: Safer Algorithmic Systems* (January 2023), and joint Statement on Principles for Responsible Algorithmic Systems (October 2022). USTPC also offers the following specific responses to select questions posed (and as numbered) in OSTP's Request for Information in the present proceeding:[1]

Protecting rights, safety, and national security

1. What specific measures — such as standards, regulations, investments, and improved trust and safety practices — are needed to ensure that AI systems are designed, developed, and deployed in a manner that protects people's rights and safety? Which specific entities should develop and implement these measures?

AI researchers have identified that AI systems often exhibit bias and lack "resiliency," that is, poorly classify data that are outside the scope of their training data. However, neither researchers nor practitioners have developed systematic, repeatable, consensus approaches for measuring bias and resiliency. USTPC strongly recommends against policy statements, official guidance, or regulation that merely require deployed AI systems to avoid exhibiting "discrimination" or other characteristics without also providing precise technical definitions of such terms. (At a minimum, there should be a defined set of analyses to check for discrimination, or a specified procedure or methodology for assessing the quality of a discrimination analysis.) We recommend that the National Institute of Standards and Technology (NIST) embark on a program to meet this need.

[*] Principal authors of this document for USTPC were Simson Garfinkel and Jody Westby.

[1] National Priorities for Artificial Intelligence, 88 FR 34194 (May 26, 2023) [https://www.govinfo.gov/content/pkg/FR-2023-05-26/pdf/2023-11346.pdf]

Specifically, we recommend that NIST seek to develop a family of measurements for bias and resilience using a process similar to the one used to develop the Advanced Encryption Standard and that is now being used for the NIST Post-Quantum Cryptography standardization project.[2] Metrics might consider the composition of the training data, feature extraction, network design, and categorization of the AI/ML techniques in use (including techniques that might be supplemented with other approaches, such as symbolic logic and formal methods). Ideally, the design of such requirements also should be part of the public process for developing the analyses, procedures, or methodologies proposed here.

USTPC further recommends that NIST work with regulatory agencies to develop approaches for certifying AI used in safety-critical systems that go beyond testing and embrace formal methods, such as approaches for using logic to mathematically prove that computer programs satisfy specific properties. While approaches for applying formal methods to AI/ML systems are in their infancy, initial research indicates it is possible to formally validate neural networks to show that classifications are stable even after the record being classified is subjected to a significant amount of noise (even specially crafted adversarial noise).[3] One such neural network verifier is α,β-CROWN (alpha-beta-CROWN), which has been applied successfully to networks with millions of artificial neurons.

The USTPC supports appropriate AI regulation and believes that statutory mandates may be required to achieve some goals. For example, Congress could make clear, as some judicial rulings already have, that an AI-based program *itself* is not entitled to intellectual property rights or exemptions from liability separate and apart from the creator and/or user of the program.[4]

We assert that AI-generated work products should be held to the same standards and best practices as human-produced work. They should show their intermediate work and provide chains of deductions and inferences (along with citations and experimental results that back up that logic), as well as provide work in a form suitable for in-depth design and peer review. We also recommend that standards be developed for representing and presenting goals, logical deductions and inferences, citations, and experimental evidence(among other information) in a manner that permits automated review.

[2] We caution that the NIST model developed for its Cybersecurity Framework, in which organizations determine their own risk levels and define the controls to manage risk, may result in inconsistent approaches since few organizations understand AI to the same extent as cybersecurity risk and controls. Instead, NIST may wish to adopt a more proscriptive approach, such as that developed for encryption algorithms, in which NIST decides upon standards with community input. Organizations then implement those standards.

[3] For example, see Huan Zhang, Shiqi Wang, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, and Zico Kolter, "A Branch and Bound Framework for Stronger Adversarial Attacks of ReLU Networks," Proceedings of the 39th International Conference on Machine Learning, 2022.

[4] For example, Section 230 of the Communications Decency Act states that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." Congress may wish to clarify whether an AI program is or is not "another information content provider" for purposes of that statute, thereby either precluding or allowing the provider or user of such a program from liability as "the publisher or speaker of the information" of the content generated by an AI program. Similarly, the Sony Betamax case took from patent law the idea that the manufacturer of a device "capable of substantial non-infringing uses" is not a contributory infringer. The role played by an AI program in creating what may be an infringing work is considerably larger than that of a non-creative device, such as a VCR. Congress thus also may wish to clarify that an AI program should not be deemed a contributory infringer if, like the VCR, it too can be used for both infringing and non-infringing purposes.

Finally, we believe that any regulations or best practices developed must include requirements for the regular review of systems, particularly given the virtual certainty that the ways they are used and applied over time will evolve in the field. Regulatory agencies also should make use of the Federal Advisory Committee Act to create advisory committees that reserve places specifically for AI researchers and academics who, by statute, are domain experts actively involved in teaching and research.

2. How can the principles and practices for identifying and mitigating risks from AI, as outlined in the Blueprint for an AI Bill of Rights and the AI Risk Management Framework, be leveraged most effectively to tackle harms posed by the development and use of specific types of AI systems, such as large language models?

We strongly support the notion that human beings should be alerted when they are interacting with an AI chatbot or reading an email or document that was automatically generated. Standardized labeling and watermarking approaches should be mandated rather than recommended for this purpose.[5] We note, however, that it will always be possible for a sophisticated attacker to remove such identifiers. Likewise, there is also a risk that legitimate watermarks might be embedded within content that was not AI generated in order to cause confusion or to attack the reputation of legitimate organizations. Thus, to be effective, watermark mandate legislation also must prohibit their improper use.

5. How can AI, including large language models, be used to generate and maintain more secure software and hardware, including software code incorporating best practices in design, coding, and post deployment vulnerabilities?

– <u>Assessing the effectiveness of AI/LLMs software generation and improved security</u>

Software engineers and programmers typically use dedicated tools for creating software code. These tools include features to assist with all phases of writing software. Some of these features are based on statistical analyses and AI/ML methods for classifying or predicting errors. Others use search to retrieve relevant, previously written and tested snippets of code.

Generative AI methods, including LLMs, generate code automatically. However, one cannot assume that code generation will improve the productivity of programmers and the accuracy of the code. Code generated by LLMs is not guaranteed to be "correct" or to actually execute. Thus, further effort is required by the programmer to clean up LLM-generated code and ensure that it functions as intended. The magnitude of such effort is likely to increase with the complexity of the software.

Moreover, it is not clear how to improve the accuracy of LLM-generated programs since their learning mechanisms and structure are not understood. Therefore, neither can be easily "tuned" to achieve a

[5] In the 1990s the United States faced a labeling issue related to how online companies collected and used personal information. Ultimately, the U.S. government chose not to adopt mandatory privacy labels and instead allowed the industry to develop voluntary approaches. Decades of academic research that followed showed that the industry's disclosure statements and labels were often confusing and sometimes intentionally misleading. We recommend that this experience not be repeated in the field of AI, and that instead regulatory agencies go through rulemaking processes and develop uniform labeling standards, similar to what nine federal agencies did in December 2009 when they published the "Final Model Privacy Form Under the Gramm-Leach-Bliley Act" in the *Federal Register*. https://www.federalregister.gov/documents/2009/12/01/E9-27882/final-model-privacy-form-under-the-gramm-leach-bliley-act.

specific quality level. We are still learning what the limits of LLMs are and whether adding relevant data can improve performance. Rather than employ them for automatic program generation, LLMs thus might be most useful for assisting programmers by generating snippets of required code that the programmer may then incorporate if desired.

– Security

LLMs are not designed to learn logical structures, rules, and logical operations. They thus do not explicitly learn programming rules or principles of secure software design, implementation, and deployment. Accordingly, there is no reason to expect that LLM-generated code would be more secure than human-generated programming. In fact, some studies have shown the opposite.[6]

We note, however, that AI/ML techniques have already been used to support specific practices around software security. For example, AI can be used to learn from previous vulnerabilities and predict whether/when they may occur for a specific software installation already deployed. Thus, with AI/ML assistants, software engineers may be able to prevent security problems or to react faster when they are detected.

– Assessing the effectiveness of AI/LLMs for hardware design and hardware security

Similar to software engineering practices, there are many areas in hardware design and hardware security where advanced statistical and machine learning methods have already been used. Generative AI methods can produce designs if purposeful AI models are trained from relevant data (e.g., specifications and performance of prior designs). However, the same significant level of validation effort is required as for software to ensure that the generated design is coherent.[7] For generating textual content, LLMs like ChatGPT have already achieved a high level of coherence. But, unfortunately, the "correctness" of the generated text is not guaranteed, and verification is required.

– Creating AI models for automated code generation

AI tools, such as large language models, are being widely used by some software developers to assist in writing code. This is likely to increase in the near future. It is also likely that AI tools will soon be writing complete programs. Experience to date has shown that AI-generated code can contain specification and security errors, likely a result of similar errors appearing in the training data.

To improve the use of AI for code generation, AI developers require large, curated datasets consisting of positive and negative examples of secure software and hardware to create generative AI systems that can be deployed to improve cybersecurity. Such datasets should be open and available royalty-free so that they can support both research and commercial activities. It is thus vital that they not be encumbered by prohibitions against commercial use.

[6] See Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt, "Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants," USENIX Security 2023; also Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh, "Do Users Write More Insecure Code with AI Assistants?," November 7, 2022 [https://arxiv.org/abs/2211.03622].

[7] It is an open empirical question as to whether generative AI models could be trained and directed to learn relevant constraints and dependencies among design components and factors to ensure that the generated designs are coherent.

Such datasets are expensive to create and maintain. This work is rarely fundable by broad conventional research programs since such maintenance is not generally considered to be research. Instead, the U.S. government may wish to explore establishing standing programs to create and curate such datasets.

The U.S. government could enforce the requirements of the open-source policy M-16-21 and make it easier to find and use "the people's code" — that is, the source code already written by the U.S. government. Ideally, the government would not only release the code but also discover vulnerabilities and design their patches. This open-source code would be readily available to help train cybersecurity models. It also could be published in a standardized manner on the NIST website and pointed to by data.gov or code.gov.

Advancing equity and strengthening civil rights

10. What are the unique considerations for understanding the impacts of AI systems on underserved communities and particular groups, such as minors and people with disabilities? Are there additional considerations and safeguards that are important for preventing barriers to using these systems and protecting the rights and safety of these groups?

USTPC recommends taking a "sectoral" approach to defining and addressing these considerations with respect to AI systems, as follows:

– General Users

- Ensure use cases, user personas, and contexts take into account the diverse and specific needs of underserved communities;
- Ensure that definitions of underserved communities encompass all germane subpopulations, including persons with individual and multiple disabilities of all kinds.

– System Design

There is a tendency to regard underserved communities as passive recipients of technology that need to be protected by responsible providers against inadvertent or disproportionate harm from technology. This is particularly true of technology, such as AI, that is relatively inaccessible or where the community is underrepresented in its development. While such protection often is necessary, theorists, scientists, designers, engineers, and practitioners from within underserved communities should be proactively involved in the creation and oversight of these technologies as expert stakeholders, not as mere users.

- Research evidence shows that when underserved communities have control over and input into the creation of technology products, this positively affects not only the design questions asked but also the methods of analysis employed and the conclusions reached, which benefits not only those communities but the products and field as a whole.

– Data

- Ensuring that AI systems are trained on representative and inclusive data that includes data from underserved communities, for example, data for job applicants who are neurodivergent or have other kinds of disabilities.

- Upholding the highest standards of ethical data collection and use, such as informed consent and confidentiality. Currently, there are vastly different ethical and legal standards for research funded with federal dollars and by private funds. Since both researchers and data move between these two communities, it would be useful to broaden ethical and regulatory frameworks, and to adopt best practices for data labeling and provenance, consistent with those that are already included in cybersecurity and privacy best practices and standards.

- Involving diverse stakeholders and community participants in research design and oversight, as exemplified by the requirement to have outside community representatives on the boards of institutional review boards under the Common Rule.

- Increasing funding for research on the responsible collection and management of data from and within underserved communities.

– Modeling and Training

- Increasing funding for the development of AI systems that specifically address the use cases, and user characteristics and needs, within underserved communities.

– Tools and Visualization

There is an immediate need for digital and AI literacy tools, especially for underserved communities, that more broadly focus on the untrustworthy and possibly biased nature of the outputs of AI systems.

– Unique Needs of Minors

- More research and legal safeguards are needed to ensure that the privacy of minors is protected when interacting with AI systems.

- Where appropriate, effective, and informed adult consent to minors' use of AI systems should be required. However, in many cases minors may require access to AI systems in ways that necessarily circumvent parental consent and knowledge — such as AI systems designed to determine if a minor is experiencing parental abuse. Opportunities should be provided for minors to receive guidance and advice from adult experts and laypersons in appropriate and validated forums in which confidentiality is possible.

- Safeguards are needed to protect minors from harm related to the content of social media searches and outputs of LLMs, as well as from the long-term impacts of AI systems. Such safeguards would be similar to those provided by laws regulating children's toys.

- There is an immediate need for minor-focused digital literacy tools geared to the potentially untrustworthy nature of the outputs of AI models. Such tools should be available in diverse languages, media, and formats that are accessible, age appropriate, and sensitive to the concerns and needs of minors.

11. *How can the United States work with international partners, including low- and middle-income countries, to ensure that AI advances democratic values and to ensure that potential harms from AI do not disproportionately fall on global populations that have been historically underserved?*

The U.S. can help create and co-fund programs for low- and middle-income countries to develop local AI expertise, by fostering online access for vetted students to AI courses, training them to participate in AI research projects, and incentivizing their best and brightest students and young professionals to remain in their countries rather than move to the U.S. or Europe. Such "locally sourced AI" would help create local AI experts and centers in these countries, which would allow them to be more responsive to local needs. The U.S. could help create and co-fund fellowships to have students, researchers and professors from these countries spend part of their time at U.S. research labs while still based in their home countries.

Specifically, we also recommend that the U.S.:

- Develop interagency programs and collaborations with international organizations (e.*g.*, the OECD and appropriate arms of the UN), development banks, and other donor agencies and organizations to fund and develop safeguards and best practices that will help detect and prevent AI systems from causing harm related to disinformation and misinformation campaigns, especially during elections and civil or natural emergencies.

- Establish collaboratives to fund research in support of developing generative AI systems to respond to the challenges faced by least developed countries (LDCs) and developing countries (e.g., inadequate electricity supplies, and lack of computing facilities, training, or other related resources).

- Adopt special safeguards to be employed in developing countries and LDCs when AI-assisted tools and data are used for disaster response (and/or in communities experiencing stress, disruption, or upheaval) with a focus on ensuring transparency, accountability, and fairness of the outcomes of such AI uses.

- Foster collaborations with foundations, industry, academia, and donor agencies and organizations to fund programs in support of building AI-ready workforces, AI research, and data ecosystems in LDCs and developing countries. Concerted efforts also are needed to build the capacity of local governments, business, educational institutions, and civil society, particularly by establishing collaborative projects or public-private partnerships.

Bolstering democracy and civic participation

15. What are the key challenges posed to democracy by AI systems? How should the United States address the challenges that AI-generated content poses to the information ecosystem, education, electoral process, participatory policymaking, and other key aspects of democracy?

The possibility that generative AI systems can be used to create disinformation that does harm to democratic processes has been widely discussed, but deserves continued attention, particularly in the areas of voter suppression, voter manipulation through deepfakes and disinformation, and voter education.[8] Less attention, however, has been focused on the less obvious risk to democracy posed by AI "hallucinations," such as the inclusion of fake citations in legal briefs, as recently demonstrated by attorneys in the southern district of New York who submitted a brief containing nonexistent case

[8] See, e.g., Mekela Panditharatne and Noah Giansiracusa, "How AI Puts Elections at Risk — and the Needed Safeguards," the Brennan Center, June 13, 2023, https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards.

citations entirely invented by ChatGPT.[9] As the number of such so-called hallucinatory references increases, it will become harder and harder for agencies, legal professionals, academics, and individuals to distinguish real from fabricated text or citations, thus ultimately undermining confidence in the American legal system and the rule of law.

USTPC thus recommends that the U.S. government foster the development and use of techniques for referencing and data finding that employ unique identifiers, such as Digital Object Identifiers (DOIs). Indeed, we urge that every document published by the U.S. government be assigned a DOI. The government should also work on expanding the use of DOIs to any such document that might be cited, including everything published in the Federal Register, government datasets, federal judicial filings, and all public comments filed in officially noticed administrative proceedings. In addition, to create unambiguous attribution, USTPC recommends that the government assign Open Researcher and Contributor Identifier (ORCID) numbers to the named authors of all government-funded published research and reports.

USTPC also suggests that the U.S. can better address the challenges AI-generated content poses to the information ecosystem, education, electoral process, participatory policymaking, and other key aspects of democracy by ensuring that personnel within the executive and legislative branches of government have the requisite expertise to guide government officials and members of Congress and are sufficiently compensated to attract and retain them in greater numbers.

Additional input

29. Do you have any other comments that you would like to provide to inform the National AI Strategy that are not covered by the questions above?

> *(a) How can cybercriminals utilize AI to commit cybercrimes, stealthily penetrate systems, and adapt to mask their presence, create malware, present false findings for forensic investigations, etc., and what do governments and private sector entities need to do to detect, deter, and counter these activities?*

Cybercriminals are using AI to significantly change the threat environment by introducing new methods of deception and stealth operations that make attacks easier to conduct and even harder to detect. Phishing attacks increased 47% last year, and Zscaler claims the use of AI helped make these attacks more personalized and credible.[10] The cybercriminal community collaborates with its own AI experts to create malware, identify paths of attack in a system, and pinpoint insecure code that can be exploited.

Although some cybersecurity vendors are incorporating AI into their solutions, they have an uphill battle to counter the advanced uses of AI already deployed by cybercriminals. Increased information sharing among all stakeholders on how AI is being used in cyberattacks will be critical to detecting and mitigating their impact. The National Science Foundation and partner agencies recently announced

[9] Larry Neumeister, "Lawyers Tell Angry New York Judge That A.I. Tricked Them into Citing Fake Cases in Court Filing," *Fortune*, June 9, 2023, https://fortune.com/2023/06/09/lawyers-angry-new-york-judge-ai-tricked-them-citing-fake-cases-court-filing-chatgpt/.

[10] *Zscaler ThreatLabz 2023 Phishing Report*, Zscaler, 2023, https://info.zscaler.com/resources-industry-reports-threatlabz-phishing-report.

that $140 million has been allocated to fund seven new research institutes devoted to AI, including one at the University of California, Santa Barbara, which will research how AI can be used to defend against cyberthreats.

This is a good first step, but the private sector needs help understanding how AI is being used in cyberattacks, the risk to operations, what it can do to counter AI-driven attacks, and how to manage incidents with an AI component. Organized collaboration between the private sector; federal, state, and local governments; law enforcement; information sharing and analysis organizations; and cybersecurity researchers on detecting and countering AI-influenced attacks is needed to raise awareness and develop approaches that can be documented and shared.

(b) *How should the U.S. collaborate with other nation-states and multilateral to facilitate a global response to and regulation of AI?*

U.S. leadership will be critical in guiding a globally harmonized approach to AI that is palatable to U.S. business, respectful of civil liberties, and furthers important U.S. national and economic security interests. Such a result can only be achieved by diplomatic efforts and public-private sector involvement in multilateral forums. Conversely, unilateral actions by the U.S. are likely to be ignored or contravened by other governments and will weaken U.S. influence in global discussions. In a globally connected world, inconsistent legal frameworks create enormous burdens and costs for organizations. The use of AI and its explosive proliferation — coupled with a flurry of regulatory action by governments — makes a globally harmonized approach to regulating AI a necessity.

According to Stanford University's 2023 AI Index, 37 AI-related pieces of legislation were adopted globally in 2022, and 123 AI-related bills have been passed since 2016.[11] The European Parliament recently approved its Artificial Intelligence Act[12] and will now begin working with member states on the final text.[13] Most recently, on June 21, 2023, U.S. Senate Majority Leader Chuck Schumer announced a new legislative process for the rapid development of a SAFE Innovation Framework.[14] Numerous U.S. government agencies also are actively analyzing possible ways to regulate AI, and at least 17 states have introduced AI-related legislation.[15]

[11] Shana Lynch, "2023 State of AI in 14 Charts," Stanford University Human-Centered Artificial Intelligence, April 3, 2023, https://hai.stanford.edu/news/2023-state-ai-14-charts.

[12] See Adam Satariano, "Europeans Take a Major Step Toward Regulating A.I.," *New York Times*, June 14, 2023, [https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html].

[13] "MEPs Ready to Negotiate First-Ever Rules for Safe and Transparent AI," *European Parliament News*, June 14, 2023, https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai.

[14] See "Majority Leader Schumer Delivers Remarks to Launch SAFE Innovation Framework for Artificial Intelligence at CSIS," press release, June 21, 2023 [https://www.democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis].

[15] Blair Levin and Larry Downes, "Who Is Going to Regulate AI?" *Harvard Business Review*, May 19, 2023, https://hbr.org/2023/05/who-is-going-to-regulate-ai.

USTPC also notes that multiple proposals for international coordination have been made to date. Twelve members of the EU Parliament recently called on European Commission President Ursula von der Leyen and U.S. President Joe Biden to convene a global summit on the regulation of AI.[16] British Prime Minister Rishi Sunak has scheduled such a summit for the fall of 2023,[17] and the Group of Seven digital ministers agreed to endorse risk-based regulation of AI.[18] Japan's digital minister called for all governments to come together to discuss AI and its risks to democracy and to work toward harmonization of regulatory approaches."[19]

Industry has also articulated the need for effective global coordination. OpenAI CEO Sam Altman, for example, recently suggested that an international body similar to the International Atomic Energy Agency be formed to oversee AI.[20]

[16] Martin Coulter and Supantha Mukherjee, "EU Lawmakers Call for Summit to Control 'Very Powerful' AI," Reuters, April 17, 2023, https://www.reuters.com/technology/eu-lawmakers-call-political-attention-powerful-ai-2023-04-17/.

[17] See "UK to Host First Global Summit on Artificial Intelligence," press release, June 7, 2023 [https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence].

[18] "G7 Digital Ministers Call for 'Risk-Based' Artificial Intelligence Regulation," WIONews, May 13, 2023, https://www.wionews.com/business-economy/g7-digital-ministers-call-for-risk-based-artificial-intelligence-regulation-587826.

[19] Leo Lewis, "AI Will Test Faith in Democracy, Tokyo Warns," *Financial Times*, Special Report G7: Japan, May 18, 2023, https://www.ft.com/content/6a6b91ca-62d0-43ac-a1c5-717ee218a2e6.

[20] Jon Gambrell, "OpenAI CEO Suggests International Agency Like UN's Nuclear Watchdog Could Oversee AI," Associated Press, June 6, 2023, https://apnews.com/article/open-ai-sam-altman-emirates-10b15d748212be7dc5d09eabd642ff39.