July 7, 2023

Institute for Security and Technology (IST)
195 41st St. PO Box # 11045 Oakland, CA 94611
info@securityandtechnology.org
501(c)(3) organization

Submitted via Federal eRulemaking Portal

Re: Request for Information, National Priorities for Artificial Intelligence

Office of Science and Technology Policy Team:

The Institute for Security and Technology (IST) appreciates the opportunity to provide feedback to the teams in the Office of Science and Technology Policy (OSTP) working on the National Artificial Intelligence Strategy that will chart a path for the United States to harness the benefits and mitigate the risks of artificial intelligence (AI).

IST drives solutions to challenges that arise at the nexus of emerging technologies and international security. We are a West Coast-based 501(c)(3) nonprofit research institute that provides unparalleled access to technologists, policy makers, and civil society organizations grappling with geopolitics, digital threats, and advanced computing. Our portfolio covers a range of issues, from cybersecurity and cybercrime, to AI governance and open source vulnerabilities, to the integration of AI in military systems and the likely impact of AI technologies on strategic stability. Our non-traditional, operationally focused approach has a bias for action, as we convene and build trust across domains, conduct applied research, and offer tangible solutions.

Overall, IST is supportive of OSTP's efforts to engage on national priorities for AI. In order to help inform the stakeholders that will rely on OSTP for guidance, we offer the following feedback that incorporates input from lines of effort across our Future of Digital Security, Geopolitics of Technology, and Innovation and Catastrophic Risk pillars. Our feedback builds off of previous and ongoing research efforts at IST to promote technically informed AI, cybersecurity, and trust and safety practices.

6. How can AI rapidly identify cyber vulnerabilities in existing critical infrastructure systems and accelerate addressing them?

Governments, security researchers, and industry experts increasingly employ AI systems to automate technical analysis and identify software vulnerabilities. The use of such AI systems can speed up threat identification and ultimately enable broader remediation of vulnerabilities that could cause harm to the critical infrastructure systems that power our everyday lives.

Use of AI is already rapidly maturing, both in offensive and defensive work-streams. There has already been great success in using AI as a tool to enhance learning and accessibility of both offensive and defensive skills. Likewise, there has been substantial AI-driven evolution in traditionally complex workflows such as reverse engineering or vulnerability identification at scale. Commercial products leveraging these opportunities have already begun to emerge - from Veracode's "Fix" product to guided malware analysis in Virus Total's Enterprise suite of tools.

However, additional research is needed to ensure that the acceleration and accessibility of traditionally less accessible skills and workflows translates to effective and safe use, and to truly understand and outline the limits of AI use in increasing offensive and defensive cyber capabilities. For example, from a defensive perspective it is well known that current models are limited in their ability to navigate complex logic vulnerabilities in code. At the same time, from an offensive perspective, AI is excellent at finding specific patterns that indicate known vulnerability classes but is currently of limited use in discovering novel or nuanced vulnerabilities.

These gaps in capability need to be identified, especially if, as expected, the trend of using AI to extend professional capabilities far exceeds the limitations of those operating it. For example, knowing that the code produced by AI is functional but not necessarily safe or resilient requires a level of expertise that may not be present in more junior developers. This in turn creates the risk of rapid code prototyping without the safety of a more critical eye to remove potentially catastrophic flaws lurking beneath the surface.

As OSTP builds out these efforts, we urge the inclusion of a range of perspectives and expertise from civil society, the private sector, and other government entities in order to develop a complete picture of the risks and opportunities involved.

AI models provide great opportunities to accelerate new learning in cyber threat research, along with opportunities to add power to both threat actors and defenders in the space. Large language models (LLMs) in particular are capable of processing information at a greater scale and dimensionality than humans. In the cyber threat context, for example, a human might use their knowledge of traditional threat actor techniques to approach a given threat, while an appropriately trained LLM might be able to rapidly identify attack tools or non-traditional attack pathways (like lateral attacks), thereby potentially identifying niche areas of risk. Further applications of AI in the identification of cyber vulnerabilities are therefore likely to arise, especially given the increasing proliferation of open source AI models.

Beyond the immediate AI platforms and models, organizations integrate AI technologies into ecosystems and workflows. We encourage OSTP to focus resources on understanding the impact of these integration activities. For example, rapid identification of vulnerabilities may not translate to increased remediation without analyzing and updating existing vulnerability response processes, which today are often bottlenecked by lagging government response times and overly taxed incident responders. Using AI to accelerate the remediation process alongside

vulnerability identification will therefore be critical to increasing the cybersecurity of critical infrastructure.

Additionally, we do not yet fully understand the limits of AI applications in cyber threat research, vulnerability identification, and remediation, and it is critical that OSTP focus resources on better understanding this nexus. As noted above, for example, some large language models are capable of producing functional code, but to date are not sophisticated enough to ensure that code is secure or safe. Further, we understand that LLMs may not only accelerate cyber threat and risk identification, but are already exacerbating human driven weaknesses in cyber defense. Complex, multi response “chatbot” driven phishing campaigns have already been identified in the wild. Likewise, AI enhanced fakes of sensitive biometric information such as voice have already been successfully used to bypass security controls. While AI did not create these flaws, it has enabled attackers to exploit them with greater ease and at much greater scale.

There are a number of different ways to approach this nexus, but we feel it is critical to reiterate that in understanding this ecosystem, OSTP and its interagency partners convene experts working across industry, government, and civil society, both in public and closed-door settings. We commend OSTP for its participation in the AI Village taking place at DEFCON 2023 as one avenue to pursue this kind of collaboration. We encourage OSTP to further expand such collaborative efforts, for example by working with a neutral arbiter to convene disparate experts and develop nuanced understandings of the opportunities and risks posed by the development and proliferation of open source artificial intelligence models. Such a convening series could outline risks and baseline current threat assessments. For example, the convenings could ask experts from industry, government, and civil society to reflect on the question, “What risks are presented in the near, medium, and long term by applications of AI in the cyber threat landscape?” Another way to frame this nexus could be to focus on AI inputs and outputs. For example, how can we use inputs like data as a control to influence the types of outputs these models produce? How can we ensure that publicly accessible data sets are secure? How can we leverage tools like software bills of material (SBOMs) to ensure the safety of training data and model weights and architecture?

There is no doubt that AI will increase the effectiveness of actors on both sides of the cyber threat ecosystem; both attackers and defenders will benefit from the integration of AI tools and the ability of some models to distill vast quantities of information and deduce patterns that are difficult for humans to identify. We must first develop a complete understanding of the range of risks and opportunities associated with this landscape, one that can only be accomplished effectively through expanded and enhanced public-private collaboration.