Response to CISA Request for Information on Secure by Design AI Software

19th February 2024

US Policy Team
policy@futureoflife.org

Request for Information (CISA-2023-0027-0001) on "Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software"

Organization

Future of Life Institute

Point of Contact

Hamza Tariq Chaudhry, US Policy Specialist. hamza@futureoflife.org

About the Organization

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government. We thank the Cybersecurity and Infrastructure Security Agency (CISA) for the opportunity to respond to this request for information (RfI) regarding their recent white paper, 'Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software.'

Executive Summary

The Future of Life Institute (FLI) has a long-standing tradition of thought leadership on AI governance toward mitigating the risks and maximizing the benefits of AI. As part of this effort, we have undertaken research and policy work at the intersection of AI, cybersecurity and secure software design. The principles outlined in CISA's Secure by Design white paper offer a tractable foundation for ensuring the security of traditional software systems. However, as the RfI suggests, there are security considerations unique to AI that are not covered by, or necessitate reinterpretation of, these principles. Focusing on AI as software, we advocate for four core principles (Protect, Prevent, Strengthen, and Standardize) for actions taken by CISA when ensuring adherence by developers to secure by design principles:

1. **Protect** advanced AI models developed in the United States from theft by malicious state and non-state actors, and from manipulation by these actors.
2. **Prevent** advanced AI systems from being used to launch AI-powered cyberattacks, both targeted at other kinds of software and also at the AI software itself.
3. **Strengthen** requirements that must be met before integrating advanced AI into cyber-defense systems, to ensure that cyber-defenses are not vulnerable to data poisoning, bias and other AI-derived harms.
4. **Standardize** ontologies and terminology unique to AI to inform the safe development, deployment, and governance of AI in the context of cybersecurity.

In line with these principles, we offer the following contributions:

1. **Offer a Framework for a 'Secure By Design' Technical Solution for AI systems.** The RfI is clear that 'AI is software and therefore should adhere to secure by design principles.' Using advanced AI for formal verification and mechanistic interpretability and relying on prior innovations such as cryptographic guarantees, we offer a framework for 'provably safe' AI systems, providing necessary conditions to make them secure by design.
2. **Analysis of, and Recommendations to Mitigate, Harms Posed to Software at the Intersection of AI and Cybersecurity.** The white paper extensively discusses the complexity of guarding against and responding to software vulnerabilities, and the RfI poses several questions regarding these issues. As advancements in AI have accelerated, the cyber threats posed to software underpinning our digital and physical infrastructure have also increased. In our policy contribution to this effort, we offer analysis of the risks posed to software by AI-cyber threats, alongside recommendations to mitigate them to protect and strengthen software security. This includes recommendations for the national cybersecurity strategy and guidance for the integration of AI in national security systems.
3. **Ensuring Transparency and Accountability from AI Products:** In keeping with a fundamental principle of the Secure by Design framework, which directs developers of software – including AI – to develop 'safe and secure products', we offer recommendations to ensure that any development of advanced AI software is transparent and that developers are held accountable for the advanced AI they produce. This includes suggestions for licensing, auditing, and assigning liability for resulting harms.
4. **Foster the Development of Common Ontologies and Terminology:** In order for software systems to be safe-by-design, they must be verifiable against technical specifications. However, these technical specifications and their expression through common ontologies have yet to be standardized. We recommend that CISA support the standardization of these ontologies and terms.

1. Technical Framework for 'Secure By Design' AI systems

Background - Summary of Research Findings

In September 2023, FLI founder and President Dr. Max Tegmark published a paper on provably safe AI systems, in co-authorship with AI safety pioneer Dr. Steve Omohundro.[1] Here, we condense the findings of that paper into a secure by design technical framework for AI systems.

The paper proposes a technical solution to designing secure AI systems by advancing the concept of provably safe AI systems. This framework has five components:

1. A Provably Compliant System (PCS) is a system (hardware, software, social or any combination thereof) that provably meets certain formal specifications.
2. Proof-carrying code (PCC) is software that is not only provably compliant, but also carries within it a formal mathematical proof of its compliance, i.e., that executing it will satisfy certain formal specifications. Because of the dramatic improvements in hardware and machine learning (ML), it is now feasible to expand the scope of PCC far beyond its original applications such as type safety, since ML can discover proofs too complex for humans to create.
3. Provably Compliant Hardware (PCH) is physical hardware the operation of which is governed by a Provable Contract.
4. Provable Contracts (PC) govern physical actions by using secure hardware to provably check compliance with a formal specification before actions are taken. They are a generalization of blockchain "Smart Contracts" which use the cryptographic guarantees to ensure that specified code is correctly executed to enable blockchain transactions. Provable contracts can control the operation of devices such as drones, robots, GPUs and manufacturing centers. They can ensure safety by checking cryptographic signatures, zero-knowledge proofs, proof-carrying code proofs, etc. for compliance with the specified rules.
5. Provable Meta-Contracts (PMC) impose formal constraints on the creation or modification of other provable contracts. For example, they might precisely define a voting procedure for updating a contract. Or they might encode requirements that provable contracts obey local laws. At the highest level, a PMC might encode basic human values that all PCs must satisfy.

Taking these components together, provably compliant systems form a natural hierarchy of software and hardware. If a GPU is PCH, then it should be unable to run anything but PCC meeting the GPU's specifications. As far as software is concerned, PCH guarantees are analogous to immutable laws of physics: the hardware simply cannot run non-compliant code. Moreover, a PCC can be often be conveniently factored into a hierarchy of packages, subroutines and functions that have their own compliance proofs. If a provable contract controls the hardware that PCC attempts to run on, it must comply with the specification. Compliance is guaranteed not by fear of sanctions from a court, but because it is provably physically impossible for the system to violate the contract.

Implications for Secure by Design AI

Due to the black box nature of AI systems, some AI experts argue that it is nearly impossible to fully secure an AI system through technical means alone.[2][3]

1 Max Tegmark and Steve Omohudro. (2023). Provably safe systems: the only path to controllable AGI. arXiv preprint arXiv:2309.01933.

2 Mike Crapps. (March, 2023). Making AI trustworthy: Can we overcome black-box hallucinations? TechCrunch.

3 W.J. von Eschenbach. (2021). Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), 1607-1622.

By applying and building on the research of Dr. Tegmark and Dr. Omohundro, however, developers can build technical components into AI systems that create a pathway to verifiably secure systems. Hence, this line of research serves as proof of concept that securing AI systems by design is technically feasible. Coupled with thoughtful policy mechanisms to strengthen the security of AI systems, we believe this type of technical solution can be effective in ensuring secure by design AI systems. We look forward to engaging with CISA in the future to expand this research project and integrate it with 'secure by design' guidance offered by CISA to AI software developers.

2. Problem Analysis and Recommendations to Mitigate Harms Posed to Software at the Intersection of AI and Cybersecurity

Numerous reports have pointed to the ways that AI systems can make it easier for malevolent actors to develop more virulent and disruptive malware, and can lower the barrier of technical expertise necessary for motivated individuals to carry out cyberattacks.[4][5] AI systems can also help adversaries automate attacks on cyberspaces, increasing the efficiency, creativity and impact of cyberattacks via novel zero-day exploits (i.e. previously unidentified vulnerabilities), targeting critical infrastructure, better automating penetration scans and exploits, and enhancing techniques such as phishing and ransomware. As AI systems are increasingly empowered to plan and execute self-selected tasks to achieve assigned objectives, we can also expect to see the emergence of autonomous hacking activities initiated by these systems in the near future. All of these developments have changed the threat landscape for software vulnerabilities. This policy contribution first summarizes these threats, and then provides recommendations that could help companies, government entities and other actors protect their software.

Threat Analysis

1. **Threat to Software Underpinning Critical Infrastructure.** An increasing proportion of US critical infrastructure, including those pieces relevant to health (e.g. hospital systems), utilities (e.g. heating, electrical supply and water supply), telecommunications, finance, and defense are now 'on the grid' – reliant on integrated online software – leaving them vulnerable to potential cyberattacks by malicious actors. Such an attack could, for instance, shut off the power supply of entire cities, access high-value confidential financial or security information, or disable telecommunications networks. AI systems are increasingly demonstrating success in exploiting such vulnerabilities in the software underpinning critical infrastructure.[6] Crucially, the barrier to entry, i.e. the level of skill necessary, for conducting such an attack is considerably lower with AI than without it, increasing threats from non-state actors and the number and breadth of possible attempts that may occur. Patching these vulnerabilities once they have been exploited takes time, which means that painful and lasting damage may be inflicted before the problem is remedied.
2. **Cyber-vulnerabilities in Labs Developing Advanced AI Software.** As the RfI outlines, there is a need to ensure that AI is protected from vulnerabilities just as is the case with traditional software. The "Secure by Design" white paper advocates for software developers to "take ownership of their customer's security outcomes." This responsibility should also apply to AI developers, compelling them to address AI-specific cyber vulnerabilities that affect both product safety for customers and wider societal concerns. The most advanced AI systems in the world – primarily being developed in the United States – are very likely to be targeted by malicious state and non-state actors to access vital design information (e.g., the model weights underpinning the most advanced large language models). Because developing these systems is resource intensive and technically complex, strategic competitors and adversaries may instead steal these technologies without taking the considerable effort to innovate and develop them, damaging U.S. competitiveness and exacerbating risks from malicious use. Once model weights are obtained, these actors could relatively easily remove the safeguards from these powerful models, which normally protect against access to dangerous

4 Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities. Artificial Intelligence Review, 54(5), 3849-3886.

5 Menn, J. (May, 2023). Cybersecurity faces a challenge from artificial intelligence's rise. Washington Post.

6 Office of Intelligence and Analysis. Homeland Threat Assessment 2024. Department of Homeland Security.

information such as how to develop WMDs. Several top cybersecurity experts have expressed concerns that the top AI labs are ill-equipped to protect these critical technologies from cyber-attacks.

3. **Integration of Opaque, Unpredictable and Unreliable AI-Enabled Cybersecurity Systems.** Partly to guard against exploitation of vulnerabilities, there has been increasing interest in the potential use of AI systems to enhance cybersecurity and cyber-defense. This comes with its own set of threats, especially with opaque AI systems for which behavior is extremely difficult to predict and explain. Data poisoning – cases where attackers manipulate the data being used to train cyber-AI systems – could lead to systems yielding false positives, failing to detect intrusions, or behaving in unexpected, undesired ways. In addition, the model weights of the systems themselves can be largely inferred or stolen using querying techniques designed to find loopholes in the model. These systems could also autonomously escalate or counter-attack beyond their operators' intentions, targeting allied systems or risking serious escalations with adversaries.

In summary, software vulnerabilities are under greater threat of covert identification and exploitation due to AI-powered cyberattacks. At the same time, integration of AI into cybersecurity systems to guard software presents unique threats of its own. Finally, the state of the art AI software being developed within leading labs within the United States is itself under threat from malicious actors.

Recommendations for Threat Mitigation

To mitigate these problems, we propose the following recommendations:

1. **Industry and governmental guidance should focus explicitly on AI-enabled cyber attacks in national cyber strategies:** AI goes completely unmentioned in the National Cybersecurity Strategy Implementation Plan published by the White House in July 2023, despite recognition of cyber risks of AI in the National Cybersecurity Strategy itself. AI risks need to be integrated explicitly into a broader cybersecurity posture, including in the DOD Cyber Strategy, the National Cyber Incident Response Plan (NCIRP), the National Cybersecurity Investigative Joint Task Force (NCIJTF) and other relevant plans.
2. **Promulgate Guidance for Minimum Standards for Integration of AI into Cybersecurity Systems and Critical Infrastructure:** Integrating unpredictable and vulnerable AI systems into critical cybersecurity systems may create cyber-vulnerabilities of its own. Minimum standards regarding transparency, predictability and robustness of these systems should be set up before they are used for cybersecurity functions in critical industries. Additionally, building on guidance issued in accordance with EO 13636 on Improving Critical Infrastructure Cybersecurity4, EO 13800 on Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure5, and the Framework for Improving Critical Infrastructure Cybersecurity published by NIST6, AI-conscious standards for cybersecurity in critical infrastructure should be developed and enforced. Such binding standards should account in particular for risks from AI-enabled cyber-attacks, and should be developed in coordination with CISA, SRMA and SLTT offices.

3. Ensuring Transparency and Accountability from AI Products

We ask that CISA and DHS consider the following recommendations to guarantee the transparent and accountable development of secure AI. In addition, these recommendations would ensure that developers take responsibility for software security and do not impose unfair costs on consumers, a fundamental principle of the Secure by Design framework. To protect and strengthen AI systems, we recommend that CISA:

1. **Require Advanced AI Developers to Register Large Training Runs and to "Know Their Customers":** The Federal Government lacks a mechanism for tracking the development and proliferation of advanced AI systems, despite there being a clear need expressed by agencies including CISA to guarantee security of AI software. In addition, these advanced AI systems could exacerbate cyber-risk for other kinds of software. In order to mitigate cybersecurity risks, it is essential to know what systems are being developed and what kinds of actors have access to them. Requiring registration for the acquisition of large amounts of computational resources for training advanced AI systems, and for carrying out the training runs themselves, would help with tracking and evaluating possible risks and taking appropriate precautions. "Know Your Customer" requirements, similar to those imposed in the financial services industry, would reduce the risk of powerful AI systems falling into the hands of malicious actors.
2. **Establish, or Support the Establishment of, a Robust Pre-deployment Auditing and Licensure Regime for Advanced AI Systems:** In order to ensure the security of AI software, it must first be guaranteed that AI systems do not behave in dangerous and unpredictable ways. Advanced AI that can pose risks to cybersecurity, may be integrated into a system's critical functions, or may be misused for malicious attacks are not presently required to undergo independent assessment for safety, security, and reliability before being deployed. Additionally, there are presently no comprehensive risk assessments for AI systems across their extensive applications and integrations. Requiring licensure before potentially dangerous advanced AI systems are deployed, contingent on credible independent audits for compliance with minimum standards for safety, security, and reliability, would identify and mitigate risks before the systems are released and become more difficult to contain. Audits should include red-teaming to identify cyber-vulnerabilities and to ensure that systems cannot be readily used or modified to threaten cybersecurity.
3. **Clarify Liability for Developers of AI Systems Used in Cyber-attacks:** In order to encourage transparency, accountability and generally protect software from AI-powered cyberattacks, it is critical to establish a liability framework for developers of AI systems that could conceivably be used to exploit cyber-vulnerabilities. At present, it is not clear under existing law whether the developers of AI systems used to, e.g., damage or unlawfully access critical infrastructure would be held liable for resulting harms. Absolving developers of liability in these circumstances creates little incentive for profit-driven developers to expend financial resources on precautionary design principles and robust assessment. Because these systems are opaque and can possess unanticipated, emergent capabilities, there is inherent risk in developing systems expected to be used in critical contexts as well as advanced AI systems more generally. Implementing strict liability when these systems facilitate or cause harm would better incentivize developers to take appropriate precautions against cybersecurity vulnerabilities, critical failure, and the risk of use in cyber-attacks.

4. Foster the Development of Common Ontologies and Terminology

The lack of standardized ontologies, terminology, and comprehensive risk management frameworks complicates the security landscape for AI systems, which present novel and amplified challenges compared to traditional software.[7] In order for software systems to be safe by design, they must be verifiably compliant with technical specifications, and technical specifications are expressed using ontologies, i.e. graphical schema representing the entity types, properties, relationships, and constraints within one or more domains of knowledge. Furthermore, the general purpose nature of many machine learning systems, which inherently have a wide range of applications, renders the assessment of their risks particularly challenging. To standardize these shared approaches we recommend that CISA:

1. **Induce and support the development of shared ontologies at the intersection of AI and cybersecurity8:** These should be developed within and across industries, government, and nations so that broader and deeper networks of compatible and provable security can more easily flourish. Likewise, development of crosswalks, bridge ontologies, and ontology alignment faculties would also aid such an ecosystem.[9]
2. **Support the standardization of terminology relevant to AI and cybersecurity:** AI security approaches have often borrowed terms, frameworks, and techniques from related fields like cybersecurity, hardware, and system safety engineering.[10] While this can occasionally be appropriate, it often leads to misinterpretations that prevent the effective use of established risk mitigation strategies. Formal definitions for what constitutes, e.g., audits, system requirements and safety requirements should be established within the context of AI and cybersecurity to avoid conflation with other fields and inform downstream management.[11]

7 While the NIST AI RMF may constitute a standardized RMF, we believe it still requires considerable iteration to fill gaps in AI risk management.

8 A shared ontology – or a shared schematic representation of concepts and terminologies across different contexts – is often developed to help collaborate on workflows. For instance, a shared biomedical ontology could help computer systems and decision-makers collate and analyze information across several different biomedical websites. In this context, it would help different actors working with a wide variety of systems in diverse contexts to effectively cooperate on AI and cybersecurity issues.

9 Crosswalks effectively function as translators in cases where complex networks of systems and data employ different terminologies and classifications for concepts. Crosswalks provide mappings to allow translation between these different schemes. A bridge ontology can serve a similar function, representing the construction of a bridge between different ontologies. All of these efforts feed into ontology alignment, the practice of ensuring correspondence between concepts in different ontologies.

10 Heidy Khlaf. (March, 2023). Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems. Trail of Bits.

11 We commend current efforts in this regard such as the NIST glossary of terms as a starting point. (See Trustworthy and Responsible AI Resource Center. Glossary. NIST. https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary) We request that this glossary be expanded and more widely adopted and applied to serve the function of effectively standardizing terminology. CISA can play a critical role here by incorporating interpretations of cybersecurity terms in AI contexts where their meaning may be more ambiguous due to distinctions between AI and traditional software.

Closing Remarks

We appreciate the thoughtful approach of CISA to the development of the Secure by Design Software framework and are grateful for the opportunity to contribute to this important effort. We hope to continue engaging with this project and subsequent projects seeking to ensure AI software does not jeopardize the continued safety, security, and wellbeing of the United States.