Cybersecurity and Artificial Intelligence: Problem Analysis and US Policy Recommendations

10th October 2023

Hamza Chaudhry & Landon Klein
policy@futureoflife.org

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

In March of this year, we released an open letter that sparked a global debate on AI safety.

Domain Definition

Cybersecurity refers to the wide array of practices concerning the attack and protection of computer systems and networks. This includes protection from attacks by malicious actors that may result in unauthorized information disclosure, theft, or damage to hardware, software, or data, as well as protection from the disruption or misdirection of services that rely on these systems. The National Cybersecurity Strategy Implementation Plan (NSCIP) published by the White House in July 2023 recognizes cybersecurity as critical to American national security interests, economic innovation, and digital empowerment.

Problem Definition

Numerous reports have pointed to the ways that artificial intelligence (AI) systems can make it easier for malevolent actors to develop more virulent and disruptive malware.[1][2] 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 and also enhancing techniques such as phishing and ransomware. As powerful AI systems are increasingly empowered to develop the set of tasks and subtasks to accomplish their objectives, autonomously-initiated hacking is also expected to emerge in the near-term.

The threats posed to cybersecurity in convergence with artificial intelligence can be broadly divided into four categories:

1. AI-Enabled/Enhanced Cyberattacks on Critical Infrastructure and Resources

An increasing proportion of US critical infrastructure, including those pieces relevant to health (hospital systems), utilities (including heating, electrical supply and water supply), telecommunications, finance, and defense are now 'on the grid', 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. Several AI systems have already demonstrated some success in exploiting such vulnerabilities. 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 of possible attempts that may occur. In addition, 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. AI-Enabled Cyber-Manipulation of High-Value Persons

Phishing refers to the fraudulent practice of sending communication (e.g., emails, caller-ID spoofed and deep-fake voice phone calls) purporting to be from reputable sources, to extract information. Advanced AI systems, in particular large language models, have

1 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.

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

demonstrated considerable effectiveness in powering phishing attacks, both by enabling greater efficiency and volume in launching these attacks, and by tailoring them to hyper-target and more effectively deceive individuals. As these abilities scale, they could be used to launch spearfishing attacks on individuals in leadership positions within organizations critical to national-security interests. The attacker could then manipulate that individual into revealing high-value information, compromising access protections (e.g. passwords) for sensitive information or critical systems, or taking decisions detrimental to national-security interests. Beyond deception, this manipulation could include blackmail techniques to compel harmful actions. Generative AI systems could also facilitate spearfishing attacks targeted at leaders of geopolitical adversaries in order to trick them into destructive 'retaliatory' action.

3. Cyber-vulnerabilities in Labs Developing Advanced AI Systems

The companies developing the most advanced AI systems in the world are primarily based within the United States and the United Kingdom. These AI systems 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). Strategic competitors and adversaries may steal these technologies without taking the considerable effort to innovate and develop them, damaging the competitiveness of the U.S and exacerbating risks from malicious use. These actors could also remove the safeguards from these powerful models which normally protect against access to dangerous information such as how to develop WMDs. In a straw poll, a majority of top cybersecurity experts expressed concerns that the top AI labs are ill-equipped to protect these critical technologies from cyber-attacks.

4. Integration of Opaque and Unreliable AI-Enabled Cybersecurity Systems

There has been growing discussion around using AI systems to enhance cybersecurity and cyber-defense. This comes with its own set of dangers, especially with opaque AI systems whose 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 or failing to detect intrusions. In addition, the model weights of the systems themselves can be stolen using querying techniques designed to find loopholes in the model. These systems could also counter-attack beyond their operators' intentions, targeting allied systems or risking escalation with adversaries.

Policy Recommendations

In light of the significant challenges analyzed in the previous section, considerable attention from policymakers is necessary to ensure the safety and security of the American people. The following policy recommendations represent critical, targeted first steps to mitigating these risks:

- **Minimum Cybersecurity Requirements for Advanced AI Developers:** Only a handful of AI developers, primarily based in the United States, are presently developing the world's

most advanced AI systems, with significant implications for American economic stability and national security. In order to safeguard these AI systems from malicious state and non-state actors, minimum cybersecurity requirements should be adopted for those developing and maintaining them, as is the case with high-risk biosafety labs (BSLs) and national nuclear laboratories (NNLs). These standards should include minimum criteria for cybersecurity personnel numbers, red-team tests, and external evaluations.

- **Explicitly Focus on AI-Enabled Cyberattacks in National Cyber-Strategies:** Artificial intelligence 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.[3] 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.

- **Establish 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 Cybersecurity[4], EO 13800 on Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure[5], and the Framework for Improving Critical Infrastructure Cybersecurity published by NIST[6], AI-concsious 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.

More general oversight and governance infrastructure for advanced AI systems is also essential to protect against cyber-risks from AI, among many other risks. We further recommend these broader regulatory approaches to track, evaluate, and incentivize the responsible design of advanced AI systems:

- **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 that could exacerbate cyber-risk. In order to adequately mitigate cybersecurity risks, it is essential to know what systems are being developed and who has access to them. Requiring registration for the acquisition of large amounts of computational resources for training advanced AI systems, and for

3 "Too often, we are layering new functionality and technology onto already intricate and brittle systems at the expense of security and resilience. The widespread introduction of artificial intelligence systems—which can act in ways unexpected to even their own creators—is heightening the complexity and risk associated with many of our most important technological systems." National Cybersecurity Strategy, March 2023, p.2.

4 Office of the Press Secretary. (February, 2013). Executive Order -- Improving Critical Infrastructure Cybersecurity. The White House.

5 Executive Office of the President. (May, 2017). Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure. National Archives.

6 National Institutes of Standards and Technology. (2018). The Framework for Improving Critical Infrastructure Cybersecurity. Department of Commerce.

carrying out the training runs themselves, would help with 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 systems that can facilitate cyber-attacks falling into the hands of malicious actors.

- **Establish a Robust Pre-deployment Auditing and Licensure Regime for Advanced AI Systems:** Advanced AI systems that can pose risks to cybersecurity, or may be integrated into cybersecurity or other critical functions, are not presently required to undergo independent assessment for safety, security, and reliability before being deployed. Requiring licensure before advanced AI systems are deployed, contingent on 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 ensure that systems cannot be readily used or modified to threaten cybersecurity.

- **Clarify Liability for Developers of AI Systems Used in Cyber-attacks:** 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 advanced AI systems and systems expected to be used in critical contexts. 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.