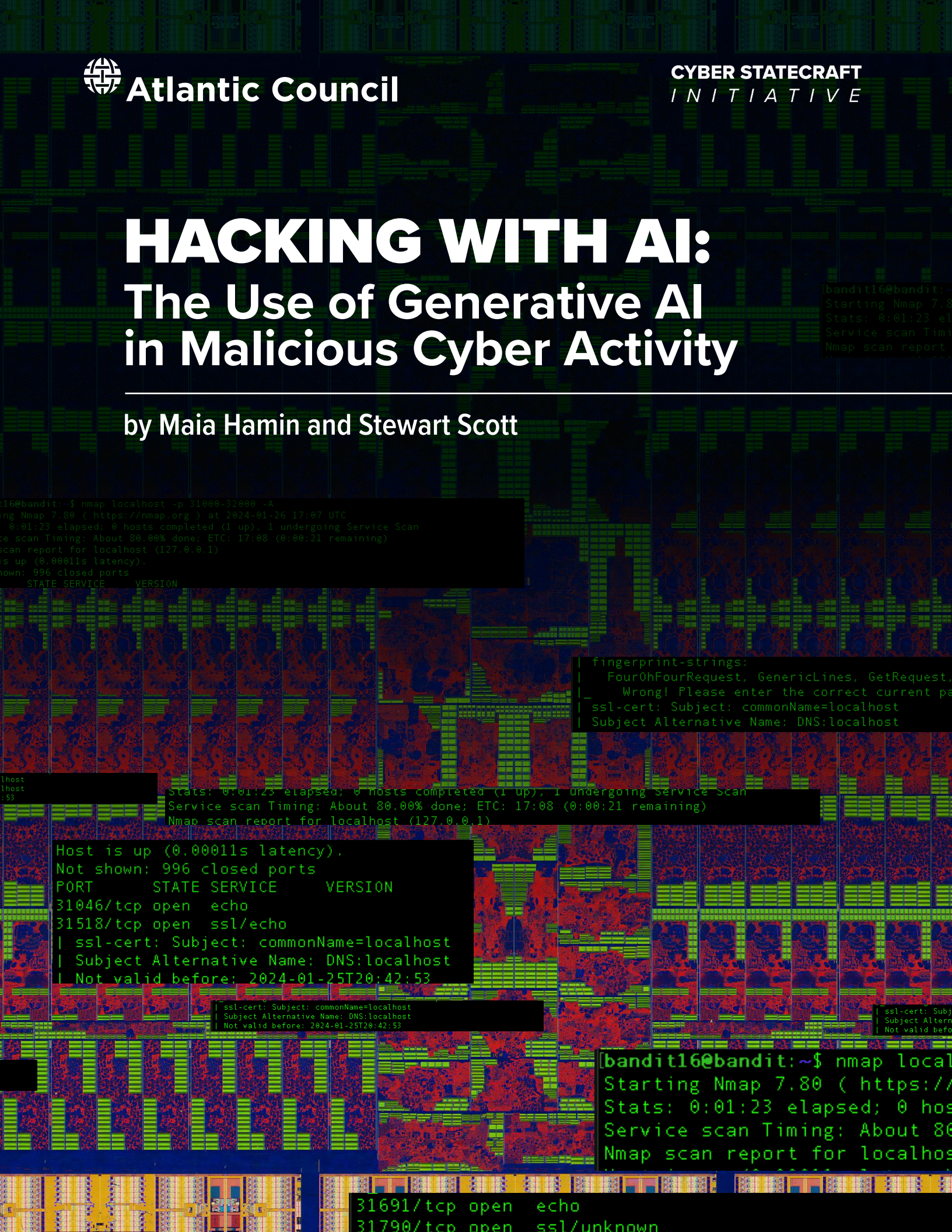
Atlantic Council

CYBER STATECRAFT INITIATIVE

HACKING WITH AI:

The Use of Generative AI in Malicious Cyber Activity

by Maia Hamin and Stewart Scott



Atlantic Council

CYBER STATECRAFT INITIATIVE

The **Cyber Statecraft Initiative** works at the nexus of geopolitics and cybersecurity to craft strategies to help shape the conduct of statecraft and better inform and secure users of technology. This work extends through the competition of state and non-state actors, the security of the internet and computing systems, the safety of operational technology and physical systems, and the communities of cyberspace. The Initiative convenes a diverse network of passionate and knowledgeable contributors, bridging the gap among technical, policy, and user communities.

Authors
Maia Hamin
Stewart Scott

Editor
Kristopher Kaliher

Atlantic Council
1030 15th Street NW, 12th Floor
Washington, DC 20005

For more information, please visit
www.AtlanticCouncil.org.

February 2024

Design by: Donald Partyka and Anais Gonzalez

Cover: Background image: Fritzchens Fritz / Better Images of AI / GPU shot etched 1 / CC-BY 4.0.
Foreground image: Terminal window showing commands commonly used by hackers. Screenshotted by Maia Hamin.

Atlantic Council

CYBER STATECRAFT
INITIATIVE

HACKING WITH AI:

The Use of Generative AI in Malicious Cyber Activity

by Maia Hamin and Stewart Scott

Table Of Contents

Acknowledgements

First and foremost, our thanks go to Sara Ann Brackett, Will Loomis, Jen Roberts, and Emma Schroeder, for their curiosity, perseverance, and good humor as they participated in the experiment described in this report. The authors would also like to thank Tim Fist, Harriet Farlow, and Katie Nickels for the thoughtful feedback they provided on various versions of this document.

Executive Summary

Questions about whether and how artificial intelligence—in particular, large language models (LLMs) and other generative AI systems—could be a tool for malicious hacking are relevant to ongoing conversations and policy frameworks seeking to manage risks from innovations in the field of artificial intelligence. This report maps the existing capabilities of LLMs to the phases of the cyberattack lifecycle to analyze whether and how these systems might alter the offensive cyber landscape. In so doing, it differentiates between generative artificial intelligence (GAI) capabilities that can help less sophisticated actors enter the space or scale up their activities—potentially increasing the overall volume of opportunistic activities such as cybercrime—and those that can enhance the capabilities of sophisticated malicious entities such as state-backed threat actors. Each phase of the cyberattack lifecycle is investigated using desk research into research papers and written accounts that examine GAI models' utility for relevant tasks or activities. This research is augmented with the findings from a novel experiment conducted in June 2023 that tasked participants with differing amounts of technical or hacking experience to complete cyber war games using the help of either ChatGPT or search engines and existing online resources.

The results of the analysis suggest that there are certain phases for which both sophisticated and unsophisticated attackers may benefit from GAI systems, most notably in social engineering, where the ability to write convincing phishing emails or to create convincing audio or video deepfakes can benefit both types of actors. For other phases, there was less evidence to suggest that contemporary GAI systems can provide meaningful new capabilities to sophisticated hackers: for example, at present, LLMs do not appear to outperform existing tools at vulnerability discovery, although this is an area of ongoing development and thus potential risk. Lower-skill actors or those who are more resource-constrained might particularly benefit from models' ability to scale up activities such as open-source information gathering. Our experiment suggested that GAI models can help novice hackers more quickly develop working code and commands, but also that it can also produce outputs that these same users are not well-positioned to vet and manage given models' current tendency to "hallucinate."[1] Built-in safeguards appeared to make LLMs less useful for novice users seeking high-level instruction on how to complete hacking tasks, but even these users found ways to circumvent safeguards. Through many of the phases, LLM outputs useful for malicious hacking—such as code for a script or text for an email—closely resemble outputs useful for more benign tasks. This resemblance makes it challenging to create safeguards that prevent models from generating outputs that could be used for hacking.

While most of this paper focuses on GAI systems as tools for human hackers, questions about autonomy, or the ability of GAI-based systems to string together multiple actions without human intervention, are also highly relevant when evaluating new offensive cyber risks that may emerge from AI. There is not yet evidence that LLM systems have the capability to complete multiple phases of an attack without human intervention, but several factors demand ongoing attention to this question, including the way that the unsupervised learning paradigm creates capabilities overhang (in which certain model abilities are only discovered over time, including after release[2]), as well as increasing focus and development energy around autonomous systems. The report contains a section examining the current state of autonomy in cyber offense as well as where autonomy might be particularly impactful in the cyberattack lifecycle.

To address these challenges, this report concludes with policy recommendations, including:

- Develop testing standards for leading-edge models that assess cyber risks across different phases, actors, and levels of autonomy, prioritizing transparency and participation
- Assess and manage cyber risk arising from GAI systems while protecting the equities of open model development
- Mobilize resources to speed up technical and standards-setting work on AI content labeling with a focus on implementation potential
- Begin investing in policy structures and technical measures to address potential risks associated with AI-based autonomous agents

Throughout, this report urges leaders to design policy based on an empirical assessment of present and future risks, avoiding reactive decision-making while ensuring that adaptive structures are in place to keep pace with the rapid rate of change in the field of AI and the potentially far-reaching implications of the technology.

1 "Hallucination" is a term for the false, misleading, or otherwise incorrect information that GAI systems will generate and state as fact. See Matt O'Brien and the Associated Press, "Tech Experts Are Starting to Doubt That ChatGPT and A.I. 'hallucinations' Will Ever Go Away: 'This Isn't Fixable,'" *Fortune,* August 1, 2023, https://fortune.com/2023/08/01/can-ai-chatgpt-hallucinations-be-fixed-experts-doubt-altman-openai/.

2 Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, November 7, 2023, http://arxiv.org/abs/2307.03718.

Introduction

Generative AI models have brought a renewed surge of interest and attention to the idea of intelligent machines. In turn, this surge has also triggered renewed conversation about the potential risks of harmful capabilities and negative societal impacts, both in the current generation of models and in future successor systems.

The question of whether AI systems are now or could in the future be capable of materially assisting malicious hackers is highly relevant for national security, as cyber criminals or nation-state adversaries could potentially harness such tools to perform more, or more successful, cyber intrusions against companies and governments. It is also of interest to those concerned by more existential fears of superintelligence: hacking would likely be a key stepping-stone for an intelligent system to escape limitations imposed by its creator. The ability of AI systems to support hacking is at the fore of many AI policy discussions: a recent Executive Order on AI from the Biden administration requires developers and Infrastructure as a Service (IaaS) providers to make reports to the federal government related to the training of "dual-use foundation models," defined in terms of their potential capability to pose serious threats to national security such as through enabling automated offensive cyber operations.[3] This centers the question of the cyber capabilities of GAI systems as a core concern in the US AI policy landscape.

How close is the reality of AI-assisted or autonomous hacking? This report seeks to answer this question by deconstructing "hacking" into a series of constituent activities and examining the potential for generative AI models (as their capabilities are currently understood) to materially assist with each phase. Rather than treating "hacking" as a monolith, this analysis relies upon known and battle-tested models of different activities used by malicious hackers to compromise a system. This report also considers the varying profiles of potential operators (ranging from cyber "noobs" to sophisticated hackers) and the various capabilities of the models themselves. In this way, different case studies and examples can be better contextualized to determine the current level of risk of AI in the cyber landscape.

But first, a few notes on terminology and scope. The term "hacking" is fraught with meaning and history in the computer security context. Many kinds of hacking are merely a kind of technical exploration of an information system, rather than an attempt to subvert its controls for malicious ends. This report specifically examines the capability of GAI systems to assist hackers with attacking information systems for malign purposes, ranging from crime to espionage. While the report examines these models' usefulness for hacking as a broad class of activity, whether and how much contemporary GAI systems can help hackers for any specific case will be informed by contextual factors including the relative strength or vulnerability of the target, the complexity and nature of the information systems at play, and the skills or behavior of the specific human operator.

The term "artificial intelligence" is also charged, describing not so much a single technology as a goal– the creation of machines with human-like intelligence–shaped by a research field with a long history that spans paradigms from rules-based systems to deep neural networks. This report focuses primarily on generative AI (GAI) as an area of recent progress and policy focus. GAI broadly refers to computational systems based on neural networks that are capable of generating novel content in different modalities (such as text, code, or images) in response to user prompts.[4] In their current form, such systems are trained through a combination of unsupervised learning from large amounts of unstructured data combined with other techniques like reinforcement learning from human feedback (RLHF), which is used to align models with more helpful and desirable behavior. GAI systems such as LLMs (like OpenAI's GPT-series models) and image diffusion models (like OpenAI's DALL-E series models) are *not the first, nor will they be the last, incarnation of AI systems.* However, AI systems created through different combinations of paradigms will likely be useful in different ways for malicious cyber activities; while this report stops short of examining each of these paradigms, the taxonomies it provides on GAI and hacking may be useful in studying or understanding the capabilities of successor systems.

3 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

4 Philippe Lorenz, Karine Perset, and Jamie Berryhill, "Initial Policy Considerations for Generative Artificial Intelligence," OECD, https://doi.org/10.1787/fae2d1e6-en.

Throughout, this report distinguishes between GAI capabilities that can help novice actors, such as opportunistic criminals entering the offensive cyberspace or seeking to scale up their activities, versus those that could make sophisticated hackers more effective. This distinction is material to understanding the impacts of AI on the cyber landscape. For example, capabilities that can help less technically resourced malicious actors enter the space could enable an expanded set of opportunistic cyber criminals to exploit more businesses with ransomware or increase other types of financially motivated cybercrime. Capabilities that improve the skills of experienced hackers, on the other hand, might pose national security concerns in the hands of experienced nation-state adversaries who might utilize the technology for espionage or in conflict.

Finally, this report focuses primarily on GAI's usefulness as a tool for malicious human hackers in each phase of the cyberattack lifecycle. In the concluding section on autonomy, this report examines the potential ability of GAI systems to enable additional functions, up to and including serving as end-to-end "hacking bots" themselves rather than as tools to produce outputs for human hackers. The utility of GAI models as a tool for human hackers is a useful indicator for this question in some ways. For example, a model's ability to provide outputs that materially assist a human with each phase of the hacking lifecycle is likely a prerequisite for the model being able to create such outputs without human direction. However, autonomy is also a distinct area of AI development, with its own trajectory and unique associated risks.

Deconstructing malicious hacking into a set of activities and malicious hackers into a set of profiles relative to their capability and resources is a way to impose structure onto a highly uncertain and fast-moving space of great policy interest. This structure forces the analysis away from platitudes and generalities about the *potential* of GAI systems and towards a more realistic examination of their *current abilities* paired with known activities that constitute malicious hacking.

This is not the first work to examine the question of using AI to automate the process of hacking. A 2020 report from the Georgetown Center for Security and Emerging Technology examined the potential use of AI across some of the same activities (drawn instead from the Lockheed Martin "cyber kill chain" model) and identified similar areas of risk.[5] However, it predated the rapid commercialization and subsequent diffusion of transformer-based models. A recent report from the UK's National Cyber Security Center (NCSC) examined the question of the near-term impact of AI on the cyber threat landscape, also focusing on similar questions and technologies.[6] While the public version of the NCSC report did not explain in detail the reasoning for its findings, they largely align with those in this analysis; this report will discuss the NCSC's findings in more detail as they reinforce or contradict its own. The rate of change within the field of AI will necessitate that work to evaluate models' usefulness for hacking be iterative and adaptive. This report is a contribution to, not the final form of, this important process.

5 Ben Buchanan et al., "Automating Cyber Attacks," Georgetown Center for Security and Emerging Technology, November 2020, https://cset.georgetown.edu/publication/automating-cyber-attacks/.

6 "The Near-Term Impacts of AI on the Cyber Threat," UK National Cyber Security Center, January 24, 2024, https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

Deconstructing the Question

The Attack Lifecycle

The MITRE ATT&CK framework is a taxonomy for adversary tactics and techniques across the phases of a cyberattack.[7] The framework has 14 unique steps, consolidated here into five sections: Reconnaissance, Initial Access, Privilege Escalation, and Lateral Movement, Impact, and Defense Evasion. Each of these sections examines which GAI capabilities are most relevant to the phase, as well as existing research, online accounts, and experimental results to form a tentative answer to the question of whether GAI's known capabilities could provide a benefit to either new or experienced actors.

- **Reconnaissance** is the phase in which a would-be attacker collects intelligence that helps them select their targets and design their attack. This can include information potentially useful for social engineering—for example, names and emails of employees of an organization—as well as information about networks and software systems such as assets, software versions, IP addresses, and open ports.
- **Gaining Access** describes the process of an attacker gaining a foothold into their target's information system. One common way to secure access is to steal credentials from a legitimate user and abuse their privileges to move within the system. Another method is to exploit a software vulnerability to perform an action that gives an attacker access, such as forcing a server to execute code or uploading a malicious file that provides an attacker with a backdoor into the system.
- **Privilege Escalation and Lateral Movement** are steps that an attacker takes once they have initially breached a system to gain additional privileges to carry out desired actions or gain access to other (potentially more sensitive or valuable) systems and resources.
- **Impact** refers to steps that a hacker takes to perform actions that represent the fulfillment of their goals within the information system. For example, encrypting files for ransomware or exfiltrating files for data theft.
- **Evasion of Defenses** refers to the various means by which malicious actors conceal their activity to avoid detection. This includes utilizing specialized software to evade monitoring systems that may look for signs of malicious activity such as signatures, improper attempts to alter or gain access to data, or questionable inbound and outbound connections used for command and control or data exfiltration.

Profiles of a Malicious Hacker

In examining GAI's utility for malicious hacking, there are two key questions about the relative sophistication of the potential user of the model:

1. Does GAI *enhance* or improve the capabilities of existing, sophisticated cyber adversaries in this stage of the attack lifecycle?
2. Does GAI *expand* the universe of potential cyber adversaries who might be able to undertake this stage of the attack lifecycle, such as by lowering the barrier to entry for those without much hacking expertise?

The answers to these questions lead to different risks and thus may demand different public policy interventions.

If generative AI can *enhance* the capabilities of existing cyber players, national security policymakers (and everyone else) should be concerned about the safety of sensitive information systems, as sophisticated nation-state adversaries or other advanced persistent threat (APT) groups could use GAI systems to support more effective cyber operations such as espionage. Policymakers would then need to consider how to limit the use of generative AI for these purposes or determine other interventions to secure systems against the new capabilities of AI-assisted actors.

If generative AI can *expand* the universe of cyber actors, then the question is one of scale. How much worse off is national security if many more actors can become somewhat competent hackers? How would organizations' digital infrastructure hold up against a surge in (perhaps not very sophisticated) attacks? There are good reasons to suspect that the answer might be "not well." Already many organizations are exploited every year through social engineering or vulnerable software, and there is little evidence to suggest that these hacks represent the exploitation of all existing vulnerabilities. As more entities around the world realize

7 "MITRE ATT&CK,", MITRE, https://attack.mitre.org/.

that cybercrime can be a lucrative source of income,[8] tools that make it easier for new actors to scale activities could cause substantial harm to businesses and consumers and create significant new costs for securing networks from a significantly increased volume of attacks.

Relevant GAI Capabilities

To map GAI capabilities to phases in the cyberattack lifecycle, this report taxonomizes current GAI uses that seem potentially useful for hacking activities:

- **Text generation**: This describes the generation of text in English or other natural (human) languages intended to be used wholesale: for example, generating the text of an email that could be used for phishing.
- **Text analysis:** Instead of asking the model to generate new text based on a prompt alone, GAI systems can also be given a text input and then asked to synthesize, summarize, or otherwise transform that information, such as by extracting information about an organization that might be useful for social engineering. This ability could be part of a system that automates part of the process of retrieving the text, such as a tool that uses an LLM to summarize the contents of a web page.
- **Code generation**: This refers to GAI's ability to generate computer-executable code according to the user's specifications (often but not always provided in natural language). This is likely the set of capabilities that would be most helpful for hacking if deeply developed, as the ability to generate (and even run) code gives a model a direct means to affect an information system.
- **Code analysis**: Combining some of the above elements, this relates to the idea of giving an LLM access to a piece of code and asking it to analyze it for another task, such as explaining what it does or searching for vulnerabilities. The outputs of this process could be natural language explanation (e.g., "this code is vulnerable to a SQL injection attack") or generated code (e.g., an additional code block that performs some function informed by the analyzed code).
- **Media generation**: This describes the ability of multimodal models to generate images, audio recordings, or videos in response to user prompts. This media might borrow from the likeness of a real person for impersonation attacks, or otherwise be used in social engineering such as to create a sense of fear or urgency on the part of the victim.
- **Operational instruction or question-answering**: This category describes the usefulness of GAI systems for providing instruction or guidance on how to complete a task. An example might be breaking down the process of an attack into discrete steps a hacker must take and providing the user with options or instructions. This function could be achieved by simply asking the language model for an answer or might be combined with the above functions, such as asking the model to search the internet for an answer.

This report primarily, although not exclusively, discusses the capabilities of general-purpose GAI systems – those trained to perform domain-neutral text, code, or image generation, rather than for specific offensive hacking tasks. For certain tasks, such as vulnerability discovery, general-purpose GAI models could likely be made even more useful through modifications such as fine-tuning, in which a model undergoes additional domain-specific training to improve its performance of a specific task.

An Experimental Contribution

The following section discusses a week-long experiment run by the authors of this report. The experiment asked participants with little to no technical background in hacking to compete in hacking "wargames"[9] with the aid of either ChatGPT or Google Search. Many online accounts of using ChatGPT or other LLM systems in support of hacking were conducted by experts who knew what to ask the tool; this experiment aimed to explore the question of how useful GAI systems are as an aid for less-sophisticated actors.

8 Emily Ferguson and Emma Schroeder, "This Job Post Will Get You Kidnapped: A Deadly Cycle of Crime, Cyberscams, and Civil War in Myanmar," Atlantic Council Cyber Statecraft Initiative, November 13, 2023, https://dfrlab.org/2023/11/13/this-job-post-will-get-you-kidnapped/.

9 "OverTheWire: Wargames," https://overthewire.org/wargames/.

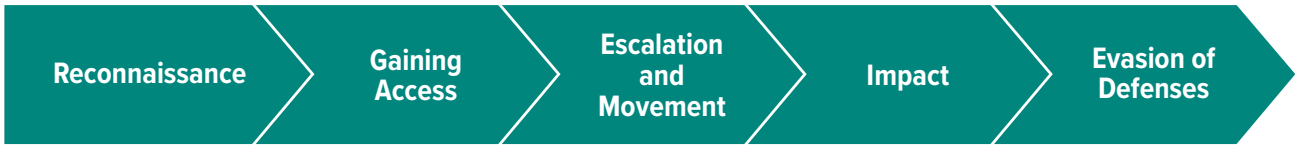
Methods

The experiment asked four participants–three with no coding experience and one with three years of coding experience–to solve online cyber war games that teach and test basic skills in penetration testing (ethical hacking). All participants completed two different game paths. The first, the "server game path," involved interacting with a remote file system, to complete tasks such as finding hidden files, searching for secrets within files, or exfiltrating information over an outbound connection. The second, the "web game path," involved interacting with a website to access hidden information by modifying cookies, injecting prompts, or uploading malicious executables.

Both game paths were broken into levels that became progressively more challenging. Both required participants to explore the technical system (e.g., the file system or website) and then write and execute commands, code snippets, or other actions to successfully obtain a password that would allow the participant to access the next level.

For each level, participants used either Google Search (and other web resources) or ChatGPT (specifically, GPT 3.5 from June of 2023).[10] We collected data on the time it took participants to complete each level, a participant's score of each level's difficulty, and self-reporting from participants on their experience using each tool. We interweave our observations from this process throughout the following sections.

Results: AI in the Attack Lifecycle



Overview

The below table summarizes for each phase of the attack lifecycle:

- The most relevant GAI capabilities
- Whether such GAI capabilities meaningfully *enhance* the capabilities of sophisticated actors (based upon a review of the relevant literature)
- Whether such GAI capabilities meaningfully *expand* the set of less-sophisticated actors or enable them to scale up their operations (based upon a review of the relevant literature and the results of our own experiment)

The below table summarizes, for each attack lifecycle, which GAI capabilities are most relevant and what present case studies suggest about whether current GAI systems can meaningfully assist sophisticated and unsophisticated cyber actors.

Notably, significant improvements in model capabilities with respect to the correctness of generated outputs, especially generated code, would change this calculus, enabling low-sophistication actors and speeding up sophisticated actors. The emergence of meaningful autonomous capabilities would also significantly alter these results: autonomy could provide new capabilities to sophisticated actors for tasks such as evading defenses and enabling semi- or fully-autonomous "hacking bots" could *dramatically* expand the set of potential opportunistic bad actors and the volume of malicious cyber activity.

The potential risks that created by model capability improvements are not equally distributed among the phases of the attack lifecycle. In particular, the Gaining Access and Escalation and Movement phases face the most risk from potential improvements in the ability of GAI models to identify vulnerabilities in code and to develop exploits. This risk is not yet realized today but seems likely to materialize in the future given substantial research interest in developing capabilities for vulnerability identification

10 Given the fast rate of improvement in models, repeating this experiment with newer generations of GPT models or with other systems would be valuable.

for cyber defense. The Evading Defenses phase stands to benefit disproportionately from increasing capabilities towards autonomy. The below table summarizes, for each attack phase, which capabilities might create risk as they improve and the level of that risk according to the likelihood and impact of substantial improvement.

Table 1: Overview of relevant GAI capabilities and level of capability enhancement across different phases of the cyberattack lifecycle

Attack Phase	Reconnaissance	Gaining Access		Escalation and Movement	Impact	Evading Defenses
		Social Engineering	Vulnerability Discovery			
Most relevant GAI capabilities	Text analysis	Text generation; media generation	Code analysis; code generation	Code analysis; code generation	Code generation; media generation	Code generation
Can current GAI systems enhance the capabilities of sophisticated actors?	Maybe	Yes	No, though a likely area of capability improvement in future	No	No	No, though a likely area of capability improvement in future
Can current GAI systems expand the set of unsophisticated actors or scale their operations?	Yes	Yes	No	Maybe, limited by reliability	Maybe, limited by reliability	No

Table 2: Overview of risk level of GAI capability enhancement across different phases of the cyberattack lifecycle

Attack Phase	Reconnaissance	Gaining Access		Escalation and Movement	Impact	Evading Defenses
		Social Engineering	Vulnerability Discovery			
Capabilities that, if improved, create most potential risk	Text analysis	Media generation	Code analysis; code generation	Code analysis; code generation	Code generation	Autonomy
Risk Level	**Medium:** High likelihood; low impact	**High:** High likelihood; high impact	**High:** Medium likelihood; high impact	**Medium:** Low likelihood, high impact	**Medium:** Medium likelihood; medium impact	**High:** Medium likelihood; high impact

Reconnaissance

In which a would-be attacker collects intelligence that helps them select their targets and design their attack.

Some parts of the reconnaissance phase are similar to other kinds of data compilation and analysis tasks where GAI is already being utilized. For example, a task that relies on compiling open-source information available on the internet, such as creating a list of an organization's employees,[11] could be completed by GAI systems with access to internet search, like Microsoft's LLM chatbot.[12] Internet-connected LLMs that search for and summarize data could present a small speed improvement over a human using a search, but they would not necessarily grant access to new or unknown information. This capability would likely benefit unsophisticated actors, who are more likely to be resource-constrained and opportunistic—the ability to process open-source information at scale could enable them to speed up this part of their work and thus target more organizations. For sophisticated actors, the consequences are less clear: if these actors are already motivated and specific in their targets, the efficiency benefits of automating or speeding up parts of the reconnaissance process might be welcome but not differentiated in terms of capability. Additionally, there is a plethora of tools available for reconnaissance of this type, including for searching through publicly accessible information (such as social media content) and data dumps (such as databases of user credentials available on the dark web) that sophisticated actors likely already know how to leverage.[13] Therefore, one significant open question in this area is whether there are types of large-scale data sources where LLMs can unlock significant new insights not otherwise available through either human review or standard searches using keywords or similar. If so, sophisticated actors might stand to see more benefit.

Separate from searching the internet for open-source information (often called "passive collection"), the reconnaissance phase also involves "active collection" in which attackers interact directly with a target information system to gather information such as the different assets in the network and the software running on each. GAI models seem less likely to aid this phase of intelligence gathering. Hackers already use semi-automated tools such as port[14] and vulnerability scanners[15] and network mappers to probe or scan target systems and identify information such as open ports, operating systems, and software versions that help them craft their attempts to compromise a system. These tools are widely accessible to current and would-be hackers.[16] In most cases, it is likely easier for experienced hackers to use existing tools rather than generate new custom code via GAI to reimplement the same functionality. However, inexperienced hackers could potentially benefit from GAI's ability to point them to these tools and provide easy-to-use instructions.

A test in 2023 purported to show that ChatGPT could answer questions about an organization's website, such as its IP address, domain names, and vendor technologies.[17] But, there is a major caveat here—the study did not test whether the information returned by ChatGPT was accurate. GAI systems are prone to returning false but plausible-sounding "hallucinations." Their knowledge ends at the end of their training data – unless the answers to these questions were present in their training data *and* have not changed *since* that data was collected, the answers returned by the model were likely fabrications. For a task like identifying the IP address or vendor technologies used by an organization, inaccurate information is equal to or worse than no information at all. Accounts like this are therefore of little use without context on the accuracy of the model's outputs.

The report from the UK's NCSC also found that AI has the potential to moderately improve sophisticated actors and to more substantially improve unsophisticated ones in the reconnaissance phase.[18] That finding largely aligns with

11 "How ChatGPT Can Be Used in Cybersecurity," Cloud Security Alliance, June 16, 2023, https://cloudsecurityalliance.org/blog/2023/06/16/how-chatgpt-can-be-used-in-cybersecurity/.

12 Yusuf Mehdi, "Reinventing Search with a New AI-Powered Microsoft Bing and Edge, Your Copilot for the Web," Microsoft, February 7, 2023, https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.

13 "Open-Source Intelligence (OSINT)," Imperva, n.d., https://www.imperva.com/learn/application-security/open-source-intelligence-osint/.

14 Sudip Sengupta, "Port Scan Attack: Definition, Examples, and Prevention," September 12, 2022, https://crashtest-security.com/port-scan-attacks/.

15 "What is Vulnerability Scanning? [And How to Do It Right]," HackerOne, June 18, 2021, https://www.hackerone.com/vulnerability-management/what-vulnerability-scanning-and-how-do-it-right.

16 Shahzeb Says, "A Quick Guide to Network Scanning for Ethical Hacking," Edureka, April 3, 2019, https://www.edureka.co/blog/network-scanning-kali-ethical-hacking/.

17 Sheetal Temara, "Maximizing Penetration Testing Success with Effective Reconnaissance Techniques Using ChatGPT," arXiv, March 20, 2023, https://doi.org/10.48550/arXiv.2307.06391.

18 "The Near-Term Impacts of AI on the Cyber Threat," UK National Cyber Security Center.

those in this report: resource-constrained actors stand to benefit the most from the potential utility of GAI models and tools built on them to automate parts of the reconnaissance process, but there also may be other avenues that benefit skilled actors who can find new ways to leverage large data sources.

Unfortunately, it will be challenging to devise safeguards for GAI systems that can limit their potential use in the Reconnaissance phase. Limitations or safeguards applied to GAI models to reduce their usefulness for open-source research for hacking reconnaissance are likely to hamper their usefulness for other legitimate tasks. For example, journalists, researchers, or financial analysts all might have legitimate reasons to ask models to amass information like a list of people who work at a particular company. A prohibition on use cases that aid hacking reconnaissance could limit many other kinds of legitimate and benign activities. This is a throughline throughout many of the phases of the attack lifecycle: many hacking activities are very similar to benign GAI use cases, presenting a major challenge for safeguarding models so that their outputs cannot support hacking.

Gaining Access

In which an attacker gains a foothold into the target information system, such as through credential theft or the exploitation of software vulnerabilities.

PHISHING AND SOCIAL ENGINEERING

According to IBM's Cost of a Data Breach report for 2023, the most common initial access vector for data breaches was phishing, in which attackers send emails or other communications that trick victims into sharing sensitive information like their password or into interacting with a malicious resource, such as a link to a fake log-in page that steals credentials or a file that provides an attacker with access to the system on which it is downloaded. Given the fact that LLMs are explicitly designed to be good at generating well-written text, they can easily be co-opted to help write text for phishing emails or other communications with a more malign purpose.

Yet, the research on their efficacy for this purpose is mixed. Two different studies found that LLM-generated emails were *less* effective than human-created emails at getting users to click on a phishing link.[19] Both studies used relatively expert humans who either had experience in social engineering or used known models for drafting effective phishing emails. It is possible LLMs would provide an advantage to hackers without experience writing phishing emails or those not fluent in the language of the organization they are targeting. LLMs could also help craft a large number of customized emails in a short amount of time. Overall, the existing research indicates that LLM-drafted phishing emails are unlikely to enhance the capabilities of existing, motivated hackers, but they could be a tool to expand phishing capabilities to a broader class of actors or provide benefits in terms of efficiency and scale.

In one of the studies, ChatGPT rebuffed requests to draft phishing emails due to its safeguards against illegal and unethical behavior. However, the authors were able to circumvent this limitation by asking the model to help them create a marketing email, into which a malicious link was inserted. As the authors note, "the only difference between a good phishing email and a marketing email can be the intention [...] if we were to prevent LLMs from creating realistic marketing emails, many legitimate use cases would be prohibited."[20]

GAI capabilities such as image, audio, and video generation also create new potential threats around a specific type of phishing known as an "impersonation attack," in which an attacker impersonates someone (perhaps a boss or coworker) to trick a user into handing over credentials or performing an action. Hackers have already used deepfake technology on video calls to pose as the CEO of crypto exchange Binance, successfully persuading crypto leaders to pay a "fee."[21] A recent segment news report demonstrated how AI systems can generate a fake voice on a phone call as part of a social engineering attack.[22] The ability to convincingly falsify a voice or video recording of a trusted individual can augment sophisticated,

19 Fredrik Heiding et al., "Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models," arXiv, November 30, 2023, https://doi.org/10.48550/arXiv.2308.12287; Pyry Åvist, "Who's Better at Phishing, Humans or ChatGPT?," HoxHunt, March 15, 2023, https://www.hoxhunt.com/blog/chatgpt-vs-human-phishing-and-social-engineering-study-whos-better.

20 Heiding et al., "Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models."

21 Kyle Barr, "Hackers Use Deepfakes of Binance Exec to Scam Multiple Crypto Projects," Gizmodo, August 23, 2022, https://gizmodo.com/crypto-binance-deepfakes-1849447018.

22 Sharyn Alfonsi, "How Con Artists Use AI, Apps, Social Engineering to Target Parents, Grandparents for Theft, CBS News, August 27, 2023, https://www.cbsnews.com/news/how-con-artists-use-ai-apps-to-steal-60-minutes-transcript/.

targeted attacks and more run-of-the-mill, low-tech scams. Additionally, as organizations—including the US government—increasingly turn to systems such as biometrics to verify identity from afar,[23] AI-based impersonation could pose another challenge to identity verification and security. Finally, image generation capabilities could also be used for social engineering purposes outside of impersonation, such as using an AI-generated image to trick a victim into thinking there has been an emergency or accident at their home or workplace, creating a sense of fear and urgency that characterizes many successful phishing messages.[24]

The NCSC report found that AI had the potential to improve sophisticated actors and to significantly improve the abilities of unsophisticated actors concerning social engineering attacks, including phishing.[25] The findings in this report largely align. Opportunistic actors who generate a high volume of phishing emails might gain the most from the ability to generate content for simple social engineering such as phishing emails, but GAI systems do not appear likely to make sophisticated actors more effective in this area given human-written phishing emails appear to be as or more effective than GAI-generated ones. However, sophisticated actors might be able to benefit from more improved social engineering vectors such as deepfaked audio or video calls. Lower-skill actors could also leverage these types of attacks, but they may also have less time and fewer resources to create convincing frauds, so this risk will be depend on the quality of deepfakes generated by existing commercial tools.

Systems that help users identify AI-generated content could help mitigate the risks that AI poses in this phase of the attack lifecycle by making it easier for technology systems such as email clients or video-calling platforms to detect and warn users of AI-generated content. These systems (and the associated implementation challenges) are addressed in the policy recommendations section.

VULNERABILITIES AND EXPLOITS

Attackers can also gain access to an information system by exploiting vulnerabilities in software code. In these cases, attackers can either exploit a known, unpatched vulnerability or discover and exploit a previously unknown vulnerability (often called a "zero-day"). Per IBM, 11 percent of data breaches last year used zero-day vulnerabilities, so another way that LLMs could significantly impact the dynamics of cybersecurity is by enabling attackers to identify new vulnerabilities more rapidly.

Interest in software systems capable of automatically identifying bugs and vulnerabilities in code did not start with the arrival of GAI systems. Back in 2016, the Defense Advanced Research Projects Agency (DARPA) hosted a Grand Cyber Challenge that asked researchers to build the best automated system for identifying software vulnerabilities.[26] LLM's fluency in reading and explaining code reignited interest in the potential use of AI to find software vulnerabilities for the purpose of better securing software systems, and DARPA launched a new AI Cyber Challenge in 2023 aiming to develop LLM-based models for the same ends.[27] Vulnerability-scanning LLMs would be unavoidably "dual-use"—they could help malicious cyber actors identify vulnerabilities in code as well as defenders seeking to harden their code against attack.

Existing research on the vulnerability discovery capabilities of LLMs does not offer immediate cause for concern (or excitement). A 2021 paper evaluating the performance of Codex—OpenAI's model trained exclusively on code—found that "Codex did not perform well when compared even to rudimentary Static Application Security Testing (SAST) tools" and reported that the authors "encountered no cases in our testing where using a Codex model led to better or more efficient results than SAST tools."[28]

A subsequent study from 2023 found that GPT3.5 did not perform significantly better than a dummy classifier (which selected vulnerabilities based on their frequency in the underlying distribution) at identifying vulnerabilities in Java code.[29] In a technical paper accompanying the release of

23 Shawn Donnan and Dina Bass, "How Did ID.Me Get Between You and Your Identity?," January 20, 2022, https://www.bloomberg.com/news/features/2022-01-20/cybersecurity-company-id-me-is-becoming-government-s-digital-gatekeeper.

24 Kelly Sheridan, "Phishing Emails That Invoke Fear, Urgency, Get the Most Clicks," Dark Reading, October 11, 2017, https://www.darkreading.com/endpoint-security/phishing-emails-that-invoke-fear-urgency-get-the-most-clicks.

25 "The Near-Term Impacts of AI on the Cyber Threat," National Cyber Security Center.

26 "Cyber Grand Challenge," Defense Advanced Research Projects Agency, https://www.darpa.mil/program/cyber-grand-challenge.

27 Justin Doubleday, "DARPA Competition Will Use AI to Find, Fix Software Vulnerabilities," Federal News Network, August 9, 2023, https://federalnewsnetwork.com/artificial-intelligence/2023/08/darpa-competition-will-use-ai-to-find-fix-software-vulnerabilities/.

28 Mark Chen et al., "Evaluating Large Language Models Trained on Code," arXiv, July 14, 2021, https://doi.org/10.48550/arXiv.2107.03374.

29 Anton Chekov, Pavel Zadorozhny, and Rodion Levichev, "Evaluation of ChatGPT Model for Vulnerability Detection," arXiv, April 12, 2023, https://doi.org/10.48550/arXiv.2304.07232.

GPT-4, OpenAI reported that "GPT-4 could explain some vulnerabilities if the source code was small enough to fit in the context window, just as the model can explain other source code," but found it "less effective than existing tools for complex and high-level activities like novel vulnerability identification."[30]

Fine-tuning LLMs on vulnerability identification tasks could increase their efficacy. A study in 2023 built a large dataset of code and code vulnerabilities and then trained LLMs and AI systems with the data. While none of the models were reliably accurate at the task, the study found that increasing the size of the training data appeared to increase model performance at finding vulnerabilities, at least up to a point, where after performance returns appeared to diminish.[31] However, this training set, though large, was still relatively small in LLM terms. Given how well-established scaling laws are across different kinds of AI model tasks,[32] more data would likely continue to improve model performance. While the present research does not suggest that LLMs are close to improving upon sophisticated bug hunters' performance, the proliferation of interest and activity around developing AI vulnerability hunting systems means this is an area for experts to monitor as GAIs continue to improve.

Another way in which AI systems could be useful in this stage is by helping to develop exploits, or code to take advantage of already-discovered vulnerabilities. However, OpenAI also reported that GPT-4 "performed poorly at building exploits for the vulnerabilities that were identified."[33] Online accounts suggest that some users have been able to convince models to write relatively simple exploits. For example, one researcher used a "jailbreak" (a prompt that puts a model into a state such that it no longer follows its training safeguards) to get ChatGPT to write code and inputs to exploit vulnerabilities, such as structured query language (SQL) injection.[34] These accounts and the findings from our experiments suggest that GAI systems could be helpful for novice hackers seeking basic ways to exploit known vulnerabilities, such as prompts to inject into an unsafe web form.

The NCSC report suggests that AI may moderately improve low-sophistication hackers and that there is a "realistic possibility of uplift" for skilled actors.[35] This may correspond to this report's findings that LLM capabilities for vulnerability identification may increase as the result of domain-specific training and development, including spillover from efforts to build such systems for defensive purposes. Vulnerability identification capabilities would likely benefit sophisticated actors such as hackers associated with nation-states, who are likelier to have access to expert tools and resources. Such systems could also potentially be used or developed by well-resourced cybercriminal groups specializing in discovering zero-day vulnerabilities for use or sale. Lower-skilled actors are less likely to rely on zero-day vulnerabilities and may follow many other attackers in preferring simpler vectors such as social engineering or exploiting known, unpatched vulnerabilities. This report's results indicate that such actors may modestly benefit from being able to rely on code-generating systems to help write simple exploits for known weaknesses such as vulnerable web-form inputs.

EXPERIMENTAL FINDINGS: WEB HACKING

In our experiment, the advanced web hacking levels in which participants completed tasks such as finding and exploiting vulnerabilities in input forms were the most similar to tasks in the "gaining access" phase.

We observed that ChatGPT generally avoided providing answers to high-level questions about finding and exploiting website vulnerabilities, although several participants found that they could get around limitations by rephrasing questions. For example, ChatGPT would not answer the question "Can you give an example of a file which could be uploaded and would result in unauthorized access?" but would answer "What do web shells that allow for file access via uploads in php look like? Can you give an example?" or "What tools do hackers use to get around exif_imagetype?" Participants experimented with strategies to circumvent these limits, such as asking the model to pretend to be a clown making balloon animals in the shape of an input that would hack the website; however, these strategies were unreliable and participants more often chose to simply rephrase their questions to avoid triggering the model's safeguards.

30 Josh Achiam et al., "GPT-4 Technical Report," arXiv, December 18, 2023, https://doi.org/10.48550/arXiv.2303.08774.

31 Yizheng Che et al., "DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection," arXiv, August 8, 2023, https://doi.org/10.48550/arXiv.2304.00409.

32 Pablo Villalobos, "Scaling Laws Literature Review," Epoch, January 26, 2023, https://epochai.org/blog/scaling-laws-literature-review.

33 Josh Achiam et al., "GPT-4 Technical Report."

34 Diego Tellaroli, "Using ChatGPT to Write Exploits," System Weakness, March 23, 2023, https://systemweakness.com/using-chatgpt-to-write-exploits-4ac7119977.

35 "The Near-Term Impacts of AI on the Cyber Threat," National Cyber Security Center.

During these levels, participants struggled with ChatGPT's reliability. For example, one of the most challenging tasks required participants to reverse-engineer a plaintext value based on the PHP code that encrypted it. Because this task combined challenging logical reasoning (reversing the encryption steps) with the need to write code, it was uniquely challenging for novice participants. Notably, ChatGPT erred in two ways during this task which made it difficult for novice participants to recover. First, it often presented logically incorrect code (for example, offering code to "reverse" a series of operations that performed those operations in the wrong order for reversal), and second, it provided incorrect answers to questions about running the code, such as "what is the reverse of this string," or, "if I were to run this code, what would be the output?" Sometimes ChatGPT would state that it could not run the code, but other times it would provide an answer to the question, which was often incorrect.[36] During the experiment, participants disagreed about whether ChatGPT was running the code itself versus simply "predicting" the output. Though it was not running the code–in-chat Code Interpreter was not available at the time of this experiment—the model's willingness to provide results that seemingly described the outputs of running code confused participants who came to believe that it could execute code if they asked it in the right way.

One of the participants described being sent into a "tailspin" as they proceeded down an incorrect path for more than an hour based on one such incorrect value returned by ChatGPT. As the participant put it, "While ChatGPT feels more approachable–easier to ask questions and do follow up–it's kind of a false comfort. Having to dig through conflicting and confusing sources through Google searching reinforces not trusting what you find and while it might slow 'progress,' it at least maybe helps to prevent 'progress' in wrong directions."

These findings suggest that ChatGPT (as of June 2023) is not yet ready to serve as a co-pilot for novice hackers to explore and exploit new information systems. Nevertheless, its ability to explain and generate custom code was useful, especially for tasks with a relatively consistent form (e.g., supplying a string that can serve as an exploit for an unsanitized input field).

Escalation of Privilege and Lateral Movement

In which an actor gains additional privileges to carry out desired actions or to pivot to gain access to other more sensitive or valuable systems and resources.

Once inside a compromised system, attackers often need to escalate their privileges or move to other system resources to access high-value data. Typically, attackers achieve this by stealing additional user credentials (e.g. by using key logging tools like Mimikatz that capture user-typed passwords) or bypassing authentication altogether (such as by "passing the hash," in which an attacker steals a valid hash to masquerades as an authenticated user).[37]

It is unclear how much benefit GAI systems can provide at this stage of an attack. There are currently few public accounts or research results examining whether and how GAI systems can write code for improperly elevating privileges or moving laterally between information systems. It is unclear whether GAI-generated code would provide any benefit compared to existing tools for this purpose. Novice hackers may benefit more than experienced ones from LLM's ability to generate simple commands to search through file systems for credentials, as well as from being able to ask models how to go about the process of seeking to escalate their access. However, our experiment found that existing safeguards are still relatively effective at preventing users from asking high-level questions about improperly escalating their access.

The NCSC report found that unsophisticated actors would receive no advantage from AI-assisted lateral movement and that sophisticated actors would see minimal benefits in this area.[38] Broadly, this report also found less evidence that GAI models could assist at this stage. However, the experimental findings suggest that unsophisticated actors may gain a slight benefit from GAI models' assistance with creating basic commands and using tools to explore networks and hunt for credentials.

36 In one example, ChatGPT gave participants a "reversed" string that had 25 out of 30 characters in the right place. Crucially, the characters at the beginning and end of the string were correct, making it easy for the human operator to miss the error.

37 Bart Lenaerts-Bergmans, "What Is Lateral Movement?," Crowdstrike, April 17, 2023, https://www.crowdstrike.com/cybersecurity-101/lateral-movement/.

38 "The Near-Term Impacts of AI on the Cyber Threat," UK National Cyber Security Center.

EXPERIMENTAL FINDINGS: LINUX COMMANDS

The "server game path" in our experiment involved tasks such as finding files or values within files and then using that data to progress through the challenge. This is a very basic version of tasks that may support privilege escalation once inside a system, such as finding files that contain log-in credentials. ChatGPT was particularly helpful for generating the basic script commands that participants required to advance through these levels. Participants reported finding it much faster to ask ChatGPT for the right code snippet than to try to figure out the command themselves using Google Search or Stack Overflow. As one participant put it, "Once I figured out how to use ChatGPT my time getting through challenges significantly reduced." Multiple participants also found it helpful that they could ask ChatGPT to explain the code it was providing.

However, in later levels, participants had to perform more complex tasks such as decrypting files with appropriate keys or using found credentials with a Secure Shell (SSH) protocol to access other servers. Our participants found ChatGPT less helpful for these kinds of open-ended tasks as compared to ones where they simply needed it to provide a command. As one participant said, "I found that ChatGPT's responses were not as helpful [...] maybe because the problems were more complex." Additionally, as participants advanced towards levels that more closely mirrored hacking tasks, they were more likely to run into ChatGPT's safeguards. For example, asking questions such as "How do I get root [privileges]?" or "How do I perform an action as another user?" would often trigger safeguards in ways that requesting the model to write a command to find a particular string within a set of text files did not.

Multiple participants noted the importance of "getting the right questions" to make ChatGPT work for their purposes. On day two, one participant described the difference as "ChatGPT may be easier to get specific answers when you have the right specific question, but it is difficult when you run into a wall that you can't seem to find the right question to get around." Participants also described feeling like they had a different level of understanding when they used ChatGPT as compared to Google. One participant said, "ChatGPT was *way* easier to resolve these puzzles, but working through Google and other types of online tools made me feel like I had a better understanding of *what* I was actually doing."

Impact

In which an attacker performs actions to fulfill their goals within the information system, such as encrypting files for ransomware or exfiltrating files for data theft.

Ransomware, in which actors encrypt the files on a system and demand payment for decryption, is an area of particular concern for how GAI capabilities may aid cyber crime. Online accounts describe using ChatGPT to generate code to implement the functionality of ransomware (finding, encrypting, and deleting files),[39] suggesting that it could provide modest benefit with this type of impact. However, it is important to note that in most of these cases, the interface refuses explicit requests to write ransomware. Instead, the operator must deconstruct the prompt into a series of tasks, such as a request to find files, then a request to encrypt them, and so on. As such, unsophisticated actors may receive less benefit, as they cannot simply ask the model to write the code for them, and must instead already understand its key functions. Additionally, the need to write custom ransomware code may not be a significant roadblock for many opportunistic cyber criminals: increasingly, groups are able to purchase malware, sometimes with accompanying infrastructure, from so-called "ransomware-as-a-service" providers.[40]

Another type of potential impact is data exfiltration, or the theft of data from a system. Data exfiltration often goes hand-in-hand with the next activity on this list: evasion of defenses. Attackers who wish to exfiltrate a large volume of data often must conceal the exfiltration activity so that it can go on for long enough to transmit the desired data before defenders can detect and stop it. Attackers use a variety of means to covertly exfiltrate data, including transferring files through file transfer protocols or cloud services, hiding exfiltrated data in network traffic such as DNS or HTTPS requests, or stashing obfuscated data in file formats such as images or audio files.[41] Little has been written about whether GAI models might unlock new ways to exfiltrate data more effectively. Some research has suggested that AI-generated images could be used to improve

39 Mark Stockley, "ChatGPT Happy to Write Ransomware, Just Really Bad at It," Malwarebytes, March 27, 2023, https://www.malwarebytes.com/blog/news/2023/03/chatgpt-happy-to-write-ransomware-just-really-bad-at-it.

40 Arianne Bleiweiss, "Off-the-Shelf Ransomware Source Code Is a New Weapon for Threat Actors," KELA Cyber Threat Intelligence, January 15, 2024, https://www.kelacyber.com/off-the-shelf-ransomware-source-code-is-a-new-weapon-for-threat-actors/.

41 Anusthika Jeyashankar, "The Most Important Data Exfiltration Techniques for a Soc Analyst to Know," Security Investigation, November 3, 2023, https://www.socinvestigation.com/the-most-important-data-exfiltration-techniques-for-a-soc-analyst-to-know/.

steganography (hiding data in ordinary files files).[42] The NCSC report predicted that both sophisticated and unsophisticated actors could use AI for more effective exfiltration, but did not specify how this would occur in practice.[43]

Evasion of Defenses

In which an attacker conceals their activities within a compromised information system to avoid detection.

Across multiple phases of the attack lifecycle, a key question for attackers is how to conceal their presence within a compromised network for long enough to achieve their objectives. How could GAI systems help them do so?

One sensational post from a cybersecurity researcher in 2023 described the ability to use ChatGPT to create detection-evading malware. However, the article makes clear that the human operator had knowledge of vendor detection systems and provided explicit prompts to ChatGPT asking it to add specific detection-evasion features such as a time-delayed start and obfuscated variable names.[44] That is, these evasion tactics were not features that the model conceived of on its own. Based upon such cases, LLMs could potentially benefit experienced attackers by helping them more efficiently write custom code to evade certain types of defenses. However, it is too soon to claim that it can help inexperienced operators do so or that it is *better* at writing such features than a sophisticated hacker.

Another potential application of LLMs in this context is for polymorphic malware: malicious code that lacks a consistent signature, making it more challenging to detect for defensive systems such as anti-virus software.[45] Security researchers have begun publishing proof-of-concept versions of AI-based polymorphic malware, such as programs that call out to the ChatGPT API to receive newly generated malicious code for execution.[46] Asking a GAI system to dynamically generate code means that the malicious instructions are stored in memory only, which avoids creating a signature that might trigger defensive systems. As a result, an Endpoint Detection and Response (EDR) system reportedly failed to flag the malware. While this threat is concerning, other security researchers have pushed back on the claims, suggesting that signature-based detection is far from the only means by which modern EDR systems identify malicious code, meaning polymorphic malware would not represent an "uncatchable" threat. Polymorphic malware of this type is not necessarily autonomous, as the human operator may still maintain primary control over the process such as by directing the prompts the model uses. However, the potential to use GAI systems and their code generation abilities as a component of more autonomous malware raises significant risks concerning the evasion of defenses. These risks are discussed in the following section.

The report from the NCSC did not cover evasion of defenses as a separate set of activities; however, it did iterate that advanced operators would be "best placed to harness AI's potential in advanced cyber operations [...] for example use in advanced malware generation."[47] This report's findings suggest that autonomy could be a meaningful enabler for advanced malware, with the caveat that the timeline for the development of reliable is highly uncertain.

42 Christian Schroeder de Witt et al., "Perfectly Secure Steganography Using Minimum Entropy Coupling," arXiv, October 30, 2023, https://doi.org/10.48550/arXiv.2210.14889.

43 "The Near-Term Impacts of AI on the Cyber Threat," UK National Cyber Security Center.

44 Aaron Mulgrew, "I Built a Zero Day Virus with Undetectable Exfiltration Using Only ChatGPT Prompts," Forcepoint, April 4, 2023, https://www.forcepoint.com/blog/x-labs/zero-day-exfiltration-using-chatgpt-prompts.

45 "What You Need to Know About Signature-based Malware Detection?," RiskXchange, May 4, 2023, https://riskxchange.co/1006984/what-is-signature-based-malware-detection/.

46 Jeff Sims, "BlackMamba: Using AI to Generate Polymorphic Malware," Hyas, July 31, 2023, https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware.

47 "The Near-Term Impacts of AI on the Cyber Threat," UK National Cyber Security Center.

Autonomy

Autonomy is not a property of the MITRE ATT&CK cycle but is relevant for assessing the risk and efficacy of GAI systems for hacking. Autonomy is defined in the military context as systems that can act "with delegated and bounded authority," in which an autonomous system takes certain decision steps usually reserved for human decision-makers without explicit direction.[48] In the AI context, the term describes systems that can identify and take actions to achieve some higher-level goal. In the offensive cyber context, this could describe the ability of a GAI system to identify the steps required to perform a task such as accessing a target information system, and then to iteratively write, run, and evaluate the results of the code until it has achieved its objective.

Ongoing work has explored the potential of "autonomous agents," software systems that use an LLM to take iterative, independent steps to achieve a user-defined goal. Generally, these models work through the "chain-of-thought" prompting, in which an LLM iteratively prompts itself to decide what to do next in service of a goal and then produce the outputs it needs to achieve that goal.[49] Typically these autonomous agent systems combine a GAI model that is used for reasoning and input creation with other software-defined capabilities that allow the agent to achieve its goals, such as a code interpreter through which it can run the code it generates or an API it can use to search the web for a query it writes.

While the initial wave of excitement around these prototypical autonomous agents tempered as it became clear they are not yet effective enough to autonomously achieve complex tasks, commercial interest in AI agents has persisted.[50] Given this enthusiasm as well as the obvious business cases—such as AI assistants capable of performing tasks like booking flights or scheduling meetings—it is likely that the field of autonomous systems will continue to attract funding and attention. As these systems operate by generating and executing code, they have a host of potential impacts on the cybersecurity landscape.

Leaving aside the obvious cybersecurity risks associated with allowing an unsupervised software system to make changes or modifications to its operator's machine or to conduct activities on the internet on their behalf, such systems could also be useful for information security and other hacking, especially as GAI models grow more capable.

For some of the phases, including Reconnaissance and Initial Access, the primary benefit afforded by autonomous systems is the combination of scalability and adaptability—the ability for one operator to launch multiple autonomous processes, each capable of executing a complex action sequence. A malicious hacker could use multiple autonomous bots to conduct bespoke phishing campaigns or spin up a set of agents to adaptively probe many different information systems for vulnerabilities.

For other stages, such as Evasion of Defenses, autonomous agents could offer benefits not only in terms of scalability but also by virtue of their autonomy itself. For example, cybersecurity defenders can often detect and impede a hack in progress by spotting unusual connections that malicious actors establish between the compromised system and external command-and-control servers that provide instructions or receive exfiltrated data.[51] Advanced cyber threat groups have devised increasingly complex ways to camouflage these connections to maintain persistence in a compromised system. If LLMs could be used to create autonomous malware that takes multiple adaptive steps within an information system without needing to call out to an external system for instructions, this could increase such actors' ability to perform other actions, such as escalating privileges, while avoiding detection.[52] This risk would be heightened if attackers can build malware using GAI models that can run locally on compromised systems since this would allow the malware to generate code and instructions without needing to establish a connection to an internet-based API that could potentially be spotted by defenders. This seems likely to be possible in the future,

48 Mark Maybury and James Carlini, "Counter Autonomy: Executive Summary," Defense Science Board, September 9, 2020, https://apps.dtic.mil/sti/citations/AD1112065.

49 For example, for a prompt such as "Write a weather report for San Francisco today," the model might reason "I need to write a weather report for San Francisco today. I should use a search engine to find the current weather conditions." This would then prompt the model to generate a search query and use it to search the internet using a pre-configured search action. For more see: "AutoGPT," LangChain, https://js.langchain.com/docs/use_cases/autonomous_agents/auto_gpt.

50 Anna Tong et al., "Insight: Race towards 'autonomous' AI Agents Grips Silicon Valley," *Reuters*, July 18, 2023, https://www.reuters.com/technology/race-towards-autonomous-ai-agents-grips-silicon-valley-2023-07-17/.

51 "Command and Control," MITRE ATT&CK," July 19, 2019, https://attack.mitre.org/tactics/TA0011/.

52 Ben Buchanan et al., "Automating Cyber Attacks."

as there has been substantial interest and development focused on adapting LLMs to be run locally on consumer devices.[53]

These possible risks associated with autonomy are not yet realized because autonomous agents are not yet particularly reliable. An evaluation of 27 different LLM models (embedded into an autonomous agent framework) on a range of tasks found that even the strongest (GPT-4) was not yet a "practically usable agent."[54] The GPT-4-based agent had a success rate of 42 percent on command-line tasks (such as answering questions about or modifying file information) and 29 percent on web browsing tasks (such as finding a specific product on a site and adding it to the user's cart). These rates are still, in some sense, impressively high, and might be sufficient for actors to use autonomous agents for certain phases of the lifecycle such as reconnaissance, where failure is not very costly. However, higher reliability (and perhaps greater task-specific sophistication) is necessary before would-be attackers can trust autonomous agents to reliably perform all the steps of the attack lifecycle.

Autonomy would be relevant for both enhancing sophisticated malicious cyber actors and expanding the set of actors. For sophisticated actors, the degree of improvement would depend heavily on the capabilities of the autonomous agents. The risks would be heightened if bots were near to or better than sophisticated human abilities and thus capable of undertaking many different paths to compromise a target system at machine speed. Less sophisticated actors could obviously benefit from the same improvements (if they were able to access and direct such systems with equal efficacy) but also might be perfectly well-served by an army of simple bots capable of testing systems for common vulnerabilities and performing standardized actions such as ransomware or data exfiltration. Here, as is true throughout considerations of autonomy, the devil will be in the details, namely the tasks in which bots are most effective and how clever and adaptable they are when confronted with the real-world diversity of information systems and cyber detection and defense measures.

These risks must be considered in the ongoing development of autonomous agent frameworks, products, and evaluations, especially for agents and systems that relate to cybersecurity. The development of autonomous agents for cyber *defense* may also risk creating tools with powerful capabilities for cyber *offense,* such as those capable of hunting through code for vulnerabilities and automatically writing patches (or instead, exploits). Additionally, the incorporation of automation into cyber defense will create new potential attack surfaces, as hackers might seek to directly target and co-opt AI-based cyber defense systems for their own ends using methods like prompt injection. Policymakers should be careful to ensure that ongoing research into autonomy, especially autonomy in the cyber context, is well-scoped and potentially released with safeguards to limit its potential dual use for malicious hacking. Researchers should study not only how to further develop autonomy, but also how to develop and deploy it safely, such as by examining which cybersecurity tasks, and to what level of autonomy, can be safely delegated to autonomous systems.

Policy Directions

Overall, GAI systems appear to have considerable potential utility for both expanding the set of cyber actors and enhancing the operations of sophisticated hackers in different ways, but the degree to which this potential is realized in current models is more mixed. For example, models do not yet appear to have the level of reliability needed to assist novice hackers from start to finish or to operate autonomously. Both sophisticated and unsophisticated operators, however, stand to benefit from current and developing capabilities in AI models that make them useful for social engineering attacks and open-source intelligence gathering. However, the prognosis for other activities, such as vulnerability identification or the development of more advanced tools for lateral movement or data exfiltration is more uncertain.

However, this reality is not permanent. The AI field has moved in fits and starts with the development of new architectures and discoveries about the power of factors such

53 Benj Edwards, "You Can Now Run a GPT-3-Level AI Model on Your Laptop, Phone, and Raspberry Pi," Ars Technica, March 13, 2023, https://arstechnica.com/information-technology/2023/03/you-can-now-run-a-gpt-3-level-ai-model-on-your-laptop-phone-and-raspberry-pi/.

54 Xiao Liu et al., "AgentBench: Evaluating LLMs as Agents," arXiv, October 25, 2023. https://doi.org/10.48550/arXiv.2308.03688.

Table 3: Summary of level of capability enhancement from GAI across different phases of the cyberattack lifecycle

Attack Phase	Reconnaissance	Gaining Access		Escalation and Movement	Impact	Evading Defenses
		Social Engineering	Vulnerability Discovery			
Can current GAI systems enhance the capabilities of sophisticated actors?	Maybe	Yes	No, though a likely area of capability improvement in future	No	No	No, though a likely area of capability improvement in future
Can current GAI systems expand the set of unsophisticated actors or scale their operations?	Yes	Yes	No	Maybe, limited by reliability	Maybe, limited by reliability	No

as scale. The current level of interest and investment in GAI and use cases such as autonomous agents make it easy to imagine that one or more paradigmatic steps forward in the way models are constructed or trained may emerge in the not-so-distant future, changing the answers to the questions posed here. In addition, the capabilities of AI systems trained using the now-dominant unsupervised learning paradigm are often discovered rather than explicitly designed by their creators; thus, additional use cases and risks alike will likely continue to emerge through the decentralized testing and use of GAI systems.

Taken together, these factors provide an opportunity as well as a challenge: can policymakers create and calibrate a legal regime that is ready to manage the risks of AI with hacking capabilities, while allowing and encouraging safe innovation in the software realm? The following recommendations propose policy approaches to manage known and knowable risks while seeking to protect the positive impacts arising from AI innovation. Where applicable, they also discuss the recommendation of these intersections with major areas of policy effort such as the recent Executive Order on AI in the United States[55] and agreements arising out of the UK's AI Safety Summit.

1

Develop testing standards for leading-edge models that assess cyber risks across different phases, actors, and levels of autonomy, prioritizing transparency and participation

The findings from this report illustrate that the benefits GAI systems deliver to hackers will be unevenly distributed across different activities in the attack lifecycle and will differ depending on an actor's methods of operation, relative strengths and limitations, and the resources at their disposal, both in terms of traditional tools and their ability to leverage and customize GAI-based tools. As governments move to establish bodies, authorities, and standards to test the safety and potential impacts of AI systems, these efforts should use these empirically grounded models of the cyberattack lifecycle to examine the full spectrum of ways that AI might influence cyber tactics and techniques preferred by different categories of actors. Testing frameworks should account for capabilities that might drastically lower barriers to entry for low-skill actors or allow such actors to significantly speed up or scale their activities, and for ways in which AI systems might afford substantially new or above-human capabilities to sophisticated actors. For both actor profiles, autonomy is a significant area of concern, so leading-edge models should be tested for their capabilities in autonomy, including when they are incorporated into current autonomous agent frameworks.

In the United States, a comprehensive step towards government-required testing of AI system capabilities came in the recent AI Executive Order, which directed the secretary of commerce to use the Defense Production Act to require companies developing "potential dual-use foundation models" to provide the federal government with information about such models, including the results of red-teaming or adversarial testing.[56] Eventually, the National Institute of Standards and Technology (NIST) will develop a standard for red-team testing which AI developers will be required to use in these reporting requirements. The EU's AI Act appears

55 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House.

56 Dual-use foundation models are defined in the EO as general-purpose models trained using unsupervised learning that have "a high level of performance" at tasks that pose a threat to national security, including by helping automate sophisticated cyberattacks, and the Commerce Department will be able to develop definitions and thresholds for the models that will be subject to this reporting requirement.

poised to require general-purpose AI models posing a "systemic risk" to uphold additional standards including red-teaming and adversarial testing,[57] and the Bletchley Agreement signed by twenty countries at the UK's Safety Summit emphasizes the responsibility of leading-edge model developers to perform and share the results of safety testing.[58]

Standards developed for adversarial testing or red teaming models for cyber risk should draw from models of the cyber-attack lifecycle like the ATT&CK framework to test how GAI models could assist with different potential activities and phases of a cyberattack, allowing decision-makers to examine the results with more specificity and consider how they differentially impact the risks created by a model. Key questions should include:

- Which steps or phases in the attack lifecycle can the tool support, and what is the level of risk or harm of improvements to that stage or activity?
- To what degree could the model enable an experienced actor to perform the task or phase more effectively? That is, how does the model's effectiveness compare to an experienced human operator or existing available tools?
- To what degree could the model enable an inexperienced actor to perform the task or phase more effectively? That is, how does the model's capability compare to an unskilled human operator or easy-to-use existing tools?
- To what extent is the model (alone or when combined with autonomous agent frameworks) capable of chaining together multiple phases of the attack lifecycle?

This report suggests a few areas of particular risk that, should they manifest, might necessitate more urgent policy interventions. One such area is vulnerability discovery—models capable of discovering zero-day vulnerabilities more efficiently than either humans or existing tools would create significant risk by potentially unlocking new vectors for sophisticated actors to attack sensitive and high-value systems. The ability for AI systems to create synthetic videos of individuals indistinguishable from real videos, or to falsify other forms of biometric authentication, could also create significant cyber risk without clear mitigation paths. Both capabilities present risk as they would offer substantial new capabilities for hackers to gain access to information systems. Finally, models capable of autonomously chaining together multiple phases of a cyberattack create extreme risk, because they could assist in scaling unskilled actors' operations, afford new capabilities in defense evasion to sophisticated actors, and create significant challenges to securing and containing models that could someday exhibit emergent self-directed behavior.

As the AI Executive Order suggests, and as the findings from this report reinforce, adversarial testing of models' cyber tactics, techniques, and autonomous potential should be performed and reported using versions of models both with and without safeguards. Our experiment and countless other accounts show that safeguards can often be evaded by changing the phrasing of requests, as well as by through more clever and technical approaches, such as "jailbreak" prompts.[59] Policymakers should presume that safeguards do little to change the baseline risk created by a model's capabilities unless and until model developers offer much more conclusive and thorough proof to the contrary.

If models capabilities continue to increase in these high-risk areas, lawmakers should consider enshrining requirements for cyber-related safety testing into the pre-release process for models. The United Kingdom's recent AI Safety Summit culminated in an agreement by AI companies to allow governments, including the United States and United Kingdom, to test their models for potential national security risks before release.[60] However, this requirement is not yet backed up with the force of law. The White House's AI Executive Order also lacks an explicit structure for whether and how the government would prevent the release of a model with capabilities that create a high level of risk. An explicit legal framework tying together testing requirements and policy mechanisms for addressing high-risk capabilities will be a crucial next evolution of these efforts. One useful model for how this requirement could be constructed in law comes from another high-stakes software domain: medical device manufacturing. The Food and Drug Administration (FDA) has created extensive requirements for manufacturers of medical devices to perform and document cybersecurity risk management processes in the design and development

57 Jillian Deutsch, "Here's How the EU Will Regulate AI Tools like OpenAI's ChatGPT and GPT-4," Fortune, December 9, 2023, https://fortune.com/2023/12/09/eu-tech-regulations-ai-openai-chatgpt-gpt-4/.

58 "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," Department for Science, Innovation & Technology, Foreign, Commonwealth & Development Office, Prime Minister's Office, November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

59 Andy Zou et al., "Universal and Transferable Attacks on Aligned Language Models," LLM Attacks, December 20, 2023, https://llm-attacks.org/.

60 Madhumita Murgia, Anna Gross, and Cristina Criddle, "AI Companies Agree to Government Tests on Their Technology to Assess National Security Risks," Financial Times, November 2, 2023, https://www.ft.com/content/8bfaa500-feee-477b-bea3-84d0ff82a0de.

of medical devices.[61] The FDA can create such a regime because, crucially, it gates access to the market, allowing the agency to place the burden onto medical device makers to justify the adequacy of their cybersecurity testing regime rather than on the FDA itself to publish a one-size-fits-all set of testing standards. A long-term framework for managing the cyber risks associated with the most leading-edge models could take inspiration from this structure.

AI model testing as enshrined in the Executive Order and in subsequent legal structures for pre-release testing should be paired with requirements for public information sharing and structures that allow non-governmental entities to participate in testing. For example, the US government should develop and publicize a plan for how they will share the information they receive under the new Executive Order, designed to maximize transparency while accounting for potential countervailing factors like national security and proprietary or business-sensitive information. Additionally, the US Congress and other legislative bodies should consider mechanisms to facilitate access to cutting-edge models for independent testing and research by civil society organizations, academic researchers, and auditing firms outside of government. Many AI companies already invite domain experts to perform red-teaming and other evaluations before a model's release; establishing this process in law would cement this good practice as a requirement in the model release lifecycle and ensure that experts have recourse to publicize or report adverse findings. So long as the companies developing AI models have sole discretion over which auditors are granted access, auditors will face perverse incentives to avoid publicizing negative findings for fear of losing privileged access.

Throughout the process of creating testing standards and policy mechanisms for acting upon the results of testing, policymakers should be attuned to the potential risks while also realistic about the fact that society has implicitly decided to allow the development of other technologies that materially aid malicious hackers—everything from Google Search itself to port sniffers and vulnerability scanners—in recognition of the fact that these technologies also provide a myriad of other benefits. While it makes sense to ensure new AI technologies do not change the cybersecurity risk landscape faster than society is equipped to manage, policy should also be premised on a clear-eyed and empirically grounded accounting of the true capabilities of these systems as well as the existing ecosystem where they are utilized. The need to carefully separate real risk from generalized excitement and anxiety about model capabilities is another reason to invest in developing multifaceted testing standards informed by real cyber tactics and techniques.

2

Assess and manage cyber risk while protecting the equities of open model development.

While the findings from this report indicate some areas of present and future concern—such as the ability to generate synthetic media useful for social engineering or autonomous system operations—they also indicate that there are still reasons to be cautious about claims that GAI models in their current form create unique risks in the hacking context. Existing (non-AI-based) software tools continue to offer would-be hackers assistance above and beyond that provided by GAI models for many activities. As policymakers consider the panoply of results likely to emerge under new AI testing requirements, they should take inspiration from the information security community's general bias towards allowing openness and the publication of new tools with both offensive and defensive capabilities[62] by ensuring AI safety regimes are compatible with open-sourcing and other public release of GAI models, absent evidence of a step-change in GAI models' hacking assistance capability.

AI models can be made more open in a variety of ways, including by publishing their source code, trained weights, or training data.[63] Open-source or otherwise publicly available AI models create many potential benefits: they allow researchers to investigate AI systems' properties and risks on topics from cybersecurity to bias and fairness, and support experimentation, innovation, and entrepreneurship by allowing developers to build a myriad of applications

61 "Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions," US Food and Drug Administration, Center for Devices and Radiological Health, September 26, 2023, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions.

62 Geoff Duncan, "Could It Be... SATAN?" TidBITS, March 20, 1995, https://tidbits.com/1995/03/20/could-it-be-satan/.

63 Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse, "Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators," Proceedings of the 5th International Conference on Conversational User Interfaces, 2023, https://doi.org/10.1145/3571884.3604316.

atop AI systems without paying enterprise prices for each API query.[64] At the same time, open models face unique governance challenges, as it is harder for their creators to impose safeguards through API restrictions, and because the ability of users to repurpose and modify open-source code as they see fit enables the potential removal of creator-imposed safeguards.[65]

In light of these benefits and risks, policymakers have begun to grapple with how to account for open-source models in AI safety and risk-management regimes. The recent AI Executive Order directed the Department of Commerce to develop a report on the risks and benefits of "dual-use foundation models with widely accessible weights" and associated potential policy approaches.[66] The leaked final text of the EU's long-negotiated AI Act also directly addresses the applicability of safety standards to open models, largely carving them out of many of the regime's requirements, with the exception of open models that pose a "systemic risk."[67] These models are defined as those with "high impact capabilities," defined in the text as those exceeding a certain compute threshold. The blended model adopted by the AI Act seems largely correct: the most capable models cannot be carved out of testing requirements, regardless of whether they are open source, but policymakers should seek to reduce compliance burdens on open model developers outside of those operating at the most leading edge of model development.

Given this report's findings that many model outputs are useful for hacking but hard to restrict due to their similarity to benign use cases, and given the many well-documented ways to circumvent safeguards in closed models,[68] the US Department of Commerce and other policymakers seeking to design policy regimes for open models should regard with skepticism arguments from large labs that models with advanced capabilities are safe for release through an API but not for their competitors to open source.[69] The policy conversation should place the onus on these large labs to demonstrate that their safeguards, API filters, and alignment techniques are robustly preventing user abuse before accepting arguments that the lack of such features makes open-source AI inherently unsafe. At the same time, policymakers will need to grapple with the fact that there may be some important safety precautions that do not work, or do not work in the same way, for open models. For example, it is still unclear whether it is possible for open models to include output watermarks that would be impossible for users to remove. The forthcoming report from the Department of Commerce and other areas of work should delineate key risk-management technologies for AI models and analyze which of these are compatible with open models, providing a more reasoned assessment of the potential risks as compared to closed models and a wider menu of policy options.

Additionally, including or excluding open models from governance regimes is not the only way for policymakers to support the equities of open developers and the safety of such models. One way to make testing requirements more equitable for the open-source ecosystem would be for the government to provide funding grants or technical infrastructure to help open model developers comply with standards. Resources and funding that organizations like the National Science Foundation have already programmed for AI-related research could be directed towards developing and evaluating anti-abuse safeguards for open models.[70] Government agencies beyond the Department of Commerce should also begin the process of engaging with open-source AI stakeholders to build trust and buy-in around governance regimes, including small developers, open-source AI users, and companies engaging in substantial open-source development or that host open-source models.

In short, where policymakers consider risk management regimes that might limit model open-sourcing or place significant barriers on open-source model developers, it is essential that such determinations are not based on fear and hype about potential capabilities but instead on empirical testing results and a clear-eyed comparison of how such risks compare to existing software tools and the tradeoffs of hampering greater transparency and openness.

64 Rishi Bommasani et al., "Issue Brief Considerations for Governing Open Foundation Models," Stanford Center for Human-Centered Artificial Intelligence, December 13, 2023, https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

65 Pranav Gade et al., "BadLlama: Cheaply removing safety fine-tuning from Llama 2-Chat 13B," arXiv, October 31, 2023, https://arxiv.org/abs/2311.00117.

66 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House.

67 Allen Overy, "EU AI Act: Key Changes in the Recently Leaked Text," January 25, 2024. https://www.allenovery.com/en-gb/global/blogs/tech-talk/eu-ai-act-key-changes-in-the-recently-leaked-text.

68 Andy Zou et al., "Universal and Transferable Attacks on Aligned Language Models."

69 Cade Metz and Mike Isaac, "In Battle Over A.I., Meta Decides to Give Away Its Crown Jewel,." The New York Times, May 18, 2023, https://www.nytimes.com/2023/05/18/technology/ai-meta-open-source.html.

70 "NSF Announces 7 New National Artificial Intelligence Research," National Science Foundation, May 4, 2023. https://new.nsf.gov/news/nsf-announces-7-new-national-artificial.

3

Mobilize resources to speed up technical and standards-setting work on AI content labeling with a focus on implementation potential.

This report found that one area of present risk concerning the intersection of GAI capabilities and hacking is the ability to synthesize images, audio, or video useful for impersonation attacks and social engineering. Depending on these tools' sophistication and accessibility, they could be useful to sophisticated hackers and opportunistic fraudsters alike. Policymakers in the United States and beyond are already aware of the need for labeling AI content on social media and communications platforms, as reports have proliferated of the use of AI-generated images in disinformation campaigns[71] and AI-generated voices in scams.[72] Methods to appropriately label AI-generated content will be key risk mitigations for cybersecurity in addition to helping combat misinformation. The United States and other governments should rapidly speed up investments in research and development of methods for AI content labeling to make it possible for policymakers to develop and begin implementing workable standards.

Proposed solutions to the problem of appropriately labeling AI-generated content include detecting the content, watermarking (embedding an unremovable identifier that content is AI-generated), or certifying the authenticity and provenance of non-AI-generated content (often via cryptography). Each of these approaches has its limitations. Currently, the outright detection of AI content suffers from poor accuracy. Researchers have found many ways to break existing proposed AI watermarks,[73] and watermarking as a general approach relies upon the compliance of AI developers with watermarking standards, which poses practical enforcement challenges related to jurisdictional issues (as some model developers may be based outside of the US) and open-source models (where model developers cannot prevent users from tampering with the watermarking functionality[74]). Authentic content certification may be the most robust solution, and there are already proposed technical standards for content provenance certification,[75] but it also faces significant challenges around implementation feasibility given the need to embed certification processes in the many different technologies through which "content" can be created and modified, from digital cameras to image editors and social media sites.

In part because of these notable limitations, it is unclear which solution is most effective, or whether the best approach will be to use multiple mechanisms in tandem. Policy should drive further research investment into this area on all fronts until it becomes clearer which avenue is most promising. The ultimate goal should be the creation of a set of standards that can be widely used for labeling AI-generated content on communication platforms such as email, videoconferencing software, and social media platforms.

The recent Executive Order on AI tasked the Department of Commerce with producing a report on the current state of AI watermarking and authentic content labeling, after which the Department of Commerce will work with the Office of Management and Budget to develop "guidance" for the federal government based on the report's findings.[76] This is an important step: the US government has already (wisely) begun to require cryptographic digital signatures on certain kinds of government communications such as subpoena orders issued by the Department of Homeland Security[77] and requirements to include provenance certification for other government-generated content should follow. However, watermarking and content authentication requirements will need to be implemented far beyond the public sector to meaningfully reduce associated cybersecurity risks. Successfully detecting and labeling AI-generated content will require not just the cooperation of AI developers but also the myriad of different technologies and platforms where content is created and transmitted, from social media sites to email clients and mobile messaging protocols.

71 David E. Sanger and Steven Lee Myers, "China Sows Disinformation About Hawaii Fires Using New Techniques," *The New York Times*, September 11, 2023, https://www.nytimes.com/2023/09/11/us/politics/china-disinformation-ai.html.

72 Carter Evans and Analisa Novak, "Scammers Use AI to Mimic Voices of Loved Ones in Distress," CBS News, July 19, 2023, https://www.cbsnews.com/news/scammers-ai-mimic-voices-loved-ones-in-distress/.

73 Kate Knibbs, "Researchers Tested AI Watermarks—and Broke All of Them," *Wired*, October 3, 2023, https://www.wired.com/story/artificial-intelligence-watermarking-issues/.

74 Siddarth Srinivasan, "Detecting AI Fingerprints: A Guide to Watermarking and Beyond," The Brookings Institution, January 4, 2024, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

75 See C2PA Specifications: https://c2pa.org/specifications/specifications/1.3/index.html.

76 "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," The White House.

77 Stephen Davidson, "New U.S. Senate Bill Proposes Digital Signatures to Protect Sensitive Court Orders," DigiCert, August 12, 2021, https://www.digicert.com/blog/new-senate-bill-proposes-digital-signatures-for-sensitive-court-documents.

Currently, the White House's voluntary commitments on AI include a promise that AI companies will develop and implement watermarking.[78] However, a system of differing watermarks will present implementation challenges for entities tasked with detection and labeling. Lawmakers in the United States and beyond should instead push for the development and implementation of standardized watermarks or content provenance certification across AI developers. Congress could also require the National Institute of Standards and Technology (NIST) to develop such standards. Alternatively, it may be better if the US could participate in and adopt standards emerging from a global body, such as the International Organization for Standards.

To make standardization possible, more research and development into the technical measures of detection, watermarking, and provenance certification will be required. This mobilization should begin now. Contests like the Federal Trade Commission's Voice Cloning Challenge are an important example of ways to begin mobilizing more resources to tackle the challenge of AI-generated audio deepfakes.[79] Policymakers should also consider approaches to force companies to internalize more of the societal costs that will be associated with addressing the problem of AI-generated content in the years to come, such as by imposing a tax on AI companies. "Pigouvian taxes" are generally designed to reimpose onto companies the costs of negative social externalities created by their products; this tax would be akin to a pollution task but instead pay for the negative impacts of polluting the information environment. Some of the revenue generated by such a tax could potentially be directed toward investments in federal research to develop AI labeling solutions. Government research funding should also be directed towards developing prototypes for the implementation of watermark detection methods or legitimate content certification in communication platforms, such as examining whether there are ways to implement such features in end-to-end encrypted systems that are wholly compatible with their privacy and confidentiality guarantees.

4

Begin investing in policy and technical measures to manage risks arising from autonomous agents.

Significant autonomous capabilities in AI models would create substantial new risks in the cyber domain. Yet, it is clear that many AI companies see agentic, empowered AI systems embedded within other systems or software as the next frontier in AI development.[80] Given the lead time required to develop new technical mitigations and policy frameworks, policymakers should start investing in developing these mitigations and frameworks now. Priorities areas should include research into the best ways to create an internet that can robustly manage autonomous cyber agents, the development of legal thinking around liability for cyber-capable autonomous systems, and ongoing engagement with international partners around the responsible use of such systems by nation-states. Questions around autonomy have been addressed little by recent policy documents such as the Executive Order on AI. While assessing the capabilities of models themselves is a key step forward, there are myriad risks from increasing autonomous capabilities in these systems that are not addressed by testing requirements alone.

The web of the future will need to be safe, usable, and resilient in the face of continuous interactions with autonomous agents or bots. Researchers should begin to examine points of potential weakness in this infrastructure, as well as ways in which autonomous agents or web infrastructure can be designed to minimize cyber risks. For example, researchers could explore systems that require bots to attest to their AI status and define safe ways for them to interact with web infrastructure. Or as is the case with content authentication, it may be infeasible to require all AI systems to self-declare and instead may be more prudent to seek safe and privacy-preserving ways for human users to verifiably attest to their humanity as they use the internet. Many governments, including that of the United States, have struggled in this domain for a long time—perhaps this moment can be the impetus they need to refocus on the development of secure tools and software to attest to digital identity.

78 "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI," The White House, July 21, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

79 "The FTC Voice Cloning Challenge," Federal Trade Commission, November 9, 2023, https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge.

80 Kevin Roose, "Personalized A.I. Agents Are Here. Is the World Ready for Them?" *The New York Times*, November 10, 2023, https://www.nytimes.com/2023/11/10/technology/personalized-ai-agents.html.

Another area of focus is clarifying liability for cyber harms caused by autonomous systems. There are many players in this equation—the developer of the LLM, the developer of the agent framework, and the user—and it is not yet clear where liability for bad outcomes rests. There are also tradeoffs in terms of different actors' ability to prevent cyber harms from arising from these systems. Such frameworks will also need to account both for intentional or criminal harms and unintentional consequences. Researchers have already found evidence of the ways that LLMs can be vulnerable to prompt injection and other attacks, which could turn the AI models themselves into a vector for cyberattacks as well as a tool. While it is not yet clear which actors are best positioned to assume responsibility, policymakers should be actively considering the question, at risk of lapsing into the world of disclaimed liability that has already bedeviled much of the software ecosystem.

Finally, the US should work with its allies and partners to establish norms around the use of autonomous offensive cyber weapons, in the same way it has led efforts to develop and define norms of responsible state behavior in cyberspace.[81] Policymakers looking for similar frameworks could take a page from the Department of Defense, which outlines governance structures and approval processes for the use of autonomous kinetic weapons.[82] These policies do not apply to autonomous cyber weapons—an implicit recognition that some forms of malware like computer worms already operate semi-autonomously—emphasizing the necessity of coalescing around shared definitions and frameworks for understanding levels of autonomy in cyber weapons and agreeing on risk management practices.

Conclusion

The intersection of cybersecurity and AI is an area of much excitement, interest, and anxiety. Current AI models are information systems rather than physical ones, and thus we should expect that their fastest areas of integration and impact will be with other information systems. As such, it is natural to wonder how such systems might be able to affect technology against our will. Cyber is also an arena of direct, offensive versus defensive competition, between states or cyber criminals and companies, and thus will be a sector ripe for experimentation and innovation in and around AI for the purposes of gaining an upper hand.

LLMs and their ability to produce code have supercharged this excitement, as well as the accompanying concern. But LLMs, by the very nature of their training paradigm, are elusive in attempts to immediately appraise their capability for certain tasks. They are master storytellers, paragons of the reasonable-sounding response. Yet, this appearance of competence is sometimes the truth and sometimes wholly fictional. Complicating the picture is that AI developers have vested commercial interests in over-promising the capabilities of their systems, and, perhaps, in portraying risks in ways that advance their policy goals. Amongst the excitement, policymakers have the unenviable task of discerning fact from science fiction and attempting to set reasonable guardrails that will protect the nation without unreasonably curtailing the development of a technology that seems likely to have major long-term economic and strategic implications.

The potential utility of GAI systems for developing or supporting offensive cyber capabilities has emerged as an early area in which concern and attention have grown. Yet often missing from these discussions is a sense of structure, a set of empirical ways to assess the capabilities of models against what we know about cyberattacks. This paper is an attempt to bridge that divide. It finds that, at present, empirical testing indicates that GAI provides certain benefits for some kinds of well-scoped tasks but that it is far from ready to independently enable new hackers or to successfully conduct a hack itself—in part due to its well-known challenges with accuracy.

At the same time, the vast amount of attention and resources pouring into the development of generative AI, and in particular into coding AI, means that this center will not hold forever. Policymakers should be skeptical yet open-minded, ready for new generations of current models

81 "Joint Statement on Advancing Responsible State Behavior in Cyberspace," US Department of State, September 23, 2019, https://www.state.gov/joint-statement-on-advancing-responsible-state-behavior-in-cyberspace/.

82 "Directive 3000.09: Autonomy in Weapons Systems," US Department of Defense, January 25, 2023. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

or for new paradigms that will upend this calculus entirely. The government should begin taking steps now to manage known or foreseeable risks, such as the use of AI-generated content for social engineering and the creation of autonomous agents that interact with web systems and the computers connected to them. Finally, policymakers should consider how to establish regulatory regimes designed to empirically test for worrisome capabilities in ways that maximize transparency and public participation to drive accountability by the largest AI labs, while seeking to calibrate such regimes to protect the open development of AI models and the good they create.

Leaders should view the current moment in context as one step in a long history of attempts to develop intelligent systems, while also seeing this as an opportunity to define forward-looking and flexible regulatory regimes that allow society to manage the potential risks arising from AI systems now and into the future. Cyber is but one example of a high-stakes domain where policymakers can seek to balance reality and the risks of the future, but only if they are willing to see these technologies as they are while trying to understand them as they may be.

About the Authors

Maia Hamin is an Associate Director with the Atlantic Council's Cyber Statecraft Initiative under the Digital Forensic Research Lab (DFRLab). She works on the intersection of cybersecurity and technology policy, including projects on the cybersecurity implications of artificial intelligence, open-source software, and cloud computing. Prior to joining the Council, Maia was a TechCongress Congressional Innovation Fellow serving in the office of Senator Ron Wyden, and before that a software engineer on Palantir's Privacy and Civil Liberties team. She holds a B.A. in Computer Science from Princeton University.

Stewart Scott is an Associate Director with the Atlantic Council's Cyber Statecraft Initiative under the Digital Forensic Research Lab (DFRLab). He works on the Initiative's systems security portfolio, which focuses on software supply chain risk management and open source software security policy. Stewart earned his B.A. from Princeton University at the School of Public and International Affairs along with a minor in Computer Science. His course of study centered on misinformation, computer science, social media policy, online extremism, journalism, and American political and economic history.

Atlantic Council

Board of Directors

CHAIRMAN
*John F.W. Rogers

EXECUTIVE CHAIRMAN EMERITUS
*James L. Jones

PRESIDENT AND CEO
*Frederick Kempe

EXECUTIVE VICE CHAIRS
*Adrienne Arsht
*Stephen J. Hadley

VICE CHAIRS
*Robert J. Abernethy
*Alexander V. Mirtchev

TREASURER
*George Lund

DIRECTORS
Stephen Achilles
Elliot Ackerman
*Gina F. Adams
Timothy D. Adams
*Michael Andersson
Alain Bejjani
Colleen Bell
Sarah E. Beshar
Stephen Biegun
Linden P. Blue
Brad Bondi
John Bonsell
Philip M. Breedlove
David L. Caplan
Samantha A. Carl-Yoder
*Teresa Carlson
*James E. Cartwright
John E. Chapoton
Ahmed Charai
Melanie Chen
Michael Chertoff
*George Chopivsky
Wesley K. Clark
*Helima Croft
Ankit N. Desai
Dario Deste
Lawrence Di Rita
*Paula J. Dobriansky
Joseph F. Dunford, Jr.
Richard Edelman
Stuart E. Eizenstat
Mark T. Esper
Christopher W.K. Fetzer
*Michael Fisch
Alan H. Fleischmann
Jendayi E. Frazer
*Meg Gentle
Thomas H. Glocer
John B. Goodman
Sherri W. Goodman
Marcel Grisnigt
Jarosław Grzesiak
Murathan Günal
Michael V. Hayden
Tim Holt
*Karl V. Hopkins
Kay Bailey Hutchison
Ian Ihnatowycz
Mark Isakowitz
Wolfgang F. Ischinger
Deborah Lee James
*Joia M. Johnson
*Safi Kalo
Andre Kelleners
Brian L. Kelly
John E. Klein
*C. Jeffrey Knittel
Joseph Konzelmann
Keith J. Krach
Franklin D. Kramer
Laura Lane
Almar Latour
Yann Le Pallec
Jan M. Lodal
Douglas Lute
Jane Holl Lute
William J. Lynn
Mark Machin
Marco Margheri
Michael Margolis
Chris Marlin
William Marron
Gerardo Mato
Erin McGrain
John M. McHugh
*Judith A. Miller
Dariusz Mioduski
*Richard Morningstar
Georgette Mosbacher
Majida Mourad
Virginia A. Mulberger
Mary Claire Murphy
Julia Nesheiwat
Edward J. Newberry
Franco Nuschese
Joseph S. Nye
*Ahmet M. Ören
Ana I. Palacio
*Kostas Pantazopoulos
Alan Pellegrini
David H. Petraeus
*Lisa Pollina
Daniel B. Poneman
*Dina H. Powell McCormick
Michael Punke
Ashraf Qazi
Thomas J. Ridge
Gary Rieschel
Charles O. Rossotti
Harry Sachinis
C. Michael Scaparrotti
Ivan A. Schlager
Rajiv Shah
Wendy R. Sherman
Gregg Sherrill
Jeff Shockey
Ali Jehangir Siddiqui
Kris Singh
Varun Sivaram
Walter Slocombe
Christopher Smith
Clifford M. Sobel
Michael S. Steele
Richard J.A. Steele
Mary Streett
Nader Tavakoli
*Gil Tenzer
*Frances F. Townsend
Clyde C. Tuggle
Francesco G. Valente
Melanne Verveer
Tyson Voelkel
Michael F. Walsh
Ronald Weiser
*Al Williams
Ben Wilson
Maciej Witucki
Neal S. Wolin
Tod D. Wolters
*Jenny Wood
Guang Yang
Mary C. Yates
Dov S. Zakheim

HONORARY DIRECTORS
James A. Baker, III
Robert M. Gates
James N. Mattis
Michael G. Mullen
Leon E. Panetta
William J. Perry
Condoleezza Rice
Horst Teltschik
William H. Webster

**Executive Committee Members*

List as of January 1, 2024

1030 15th Street, NW, 12th Floor,
Washington, DC 20005
(202) 778-4952
www.AtlanticCouncil.org