CYBER STATECRAFT
INITIATIVE

#ACcyber

ISSUE BRIEF

AUGUST 2024

AI in Cyber and Software Security: What's Driving Opportunities and Risks?

MAIA HAMIN, JENNIFER LIN, AND TREY HERR

ABSTRACT *This paper discusses rapid advancements in artificial intelligence (AI), focusing on generative artificial intelligence (GAI) and its implications for cybersecurity and policy. As AI technologies evolve, they present both opportunities and risks, necessitating some understanding of what drives each. This is crucial not only for harnessing AI's capabilities in cybersecurity—where AI can both defend against and potentially enhance cyber threats—but also in considering broader national security implications. Throughout, the issue brief highlights the importance of acknowledging the long history and varied paradigms within AI development. It also emphasizes the need to consider how AI technologies are integrated into larger software systems and the unique risks and opportunities this presents. Finally, the brief calls for a more nuanced understanding of AI's impact across different sectors.*

INTRODUCTION

The rapid pace of technological improvement and the resulting groundswell of innovation and experimentation in artificial intelligence (AI) has prompted a parallel conversation in policy circles about how to harness the benefits and manage the potential risks of these technologies. Open questions in this conversation include how to map or taxonomize the set of known risks, how to assign responsibility to different actors in the ecosystem to address these risks, and how to build policy structures that can adapt to manage "unknown unknowns" (e.g., AI-related risks that are hard to predict at present). Then, add in the question of how to do all of the above while preserving some essential abilities: the broader public's to express their preferences, the research community's to innovate, and industry's to commercialize responsibly. Each of these will be a foundation for realizing the potential benefits of generative artificial intelligence (GAI) innovations and preserving the US edge in AI development to the benefit of its economic productivity and security.

This report focuses on the risks and opportunities of AI in the cyber context. Current GAI systems have proven capabilities in writing and analyzing computer code, raising the specter of their usefulness to both cybersecurity defense and offense. Cybersecurity is, by its nature, an adversarial context in which operators of information systems compete against cybercriminals and nation-state hackers. Thus, if and

The **Cyber Statecraft Initiative**, part of the ACTech programs, works at the nexus of geopolitics and cybersecurity to craft strategies to help shape the conduct of statecraft and to 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.

when AI provides a "means" to improve cybersecurity capabilities, there will be no shortage of actors with "motives" to exploit these capabilities for good and ill. As critical infrastructure and government services alike increasingly rely on computing to deliver vital goods, cybersecurity questions are also increasingly questions of national security, raising the stakes for appraising both cyber opportunity and risk.

Cybersecurity is far from the only AI application that may create opportunity or risk. The harms of non-consensual intimate imagery and harassment, the manufacture of bioweapons, the integration of biased or flawed outputs into decision-making processes, or other areas of AI risk will take different forms and demand varying mitigations. The factors that drive risk and opportunity in the cyber context may provide useful insight across other contexts as well—the authors of this paper respectfully leave it to experts in those other fields to draw from its findings as much or as little as they suit.

An important note on scope: an all-too-frequent assumption in contemporary policy conversations is that AI is synonymous with GAI. Yet—as this paper later discusses—GAI is merely the latest and greatest innovation from a decades-old field in which different paradigms and approaches to crafting nonhuman intelligent systems have risen and fallen over time. This work focuses on capabilities shown—or suggested—by current AI systems, including GAI, because these examples provide a grounded basis for reasoning about AI capabilities and accompanying risks and opportunities. Where appropriate, the report mentions or considers other AI paradigms that could prove relevant to risk and opportunity in the cybersecurity context. The report weighs, as well, not just standalone models but also "AI systems" that involve AI models embedded into broader software systems, such as an AI model paired with a code interpreter or a Retrieval-Augmented Generation (RAG) system.[1, 2]

OPPORTUNITIES FROM AI IN THE CYBERSECURITY CONTEXT

In the broadest sense, the opportunities of AI in the cybersecurity context arise from their potential use to improve a defender's lot in cybersecurity, whether by helping secure code or by helping make cybersecurity tasks easier or more efficient for defenders. Many of these opportunities arise from GAI models' ability to read, analyze, and write code.

A. FINDING AND FIXING VULNERABILITIES IN CODE

AI models that can detect vulnerabilities in software code—and, ideally, propose solutions—could benefit cybersecurity defenders by helping them scan code to find—and fix—vulnerabilities before malicious actors can exploit these. AI tools that could find significantly more vulnerabilities than existing tools, such as static analysis or fuzzing tools, could improve programmers' ability to run checks over their code before merging it or building it, preventing the deployment of vulnerable code to customers. Using these tools on existing codebases will create more challenges since applications may necessitate asking customers to patch or upgrade their code. These tools might be particularly valuable in low-resource contexts in which developers do not have access to in-house security expertise or security code reviews, such as small businesses, nonprofits, and open-source maintainers.

Using AI to find vulnerabilities in code is an area of active research effort. For example, the Defense Advance Research Projects Agency (DARPA) and Advanced Research Projects Agency for Health (ARPA-H) are partners in the two-year AI Cyber Challenge (AIxCC) that asks participants to "design novel AI tools and capabilities" to help automate the process of vulnerability detection or other cyber defense activities.[3] Right now, the open debate in this area is how good GAI models are at this task and how good they can become. One blog post from a small-business AIxCC semi-finalist said, "our experiments lead us to believe real-world performance on code analysis tasks may be worse than current benchmarks can measure quantitatively."[4] Some benchmarks do exist, such as the CyberSecEval2 framework,[5] developed by Meta—yet evidence offers mixed evaluations. The original authors of the CyberSecEval2 paper found "none" of the large language models (LLMs) "do very well on these challeng-

[1] "Assistants API Overview: How Assistants work," Open AI Platform, accessed June 30, 2024, https://platform.openai.com/docs/assistants/overview.

[2] Patrick Lewis et al, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv, April 12, 2021 [last revised], https://doi.org/10.48550/arXiv.2005.11401.

[3] Advanced Research Projects Agency for Health (ARPA-H), "ARPA-H Joins DARPA's AI Cyber Challenge to Safeguard Nation's Health Care Infrastructure from Cyberattacks," March 21, 2024, https://arpa-h.gov/news-and-events/arpa-h-joins-darpas-ai-cyber-challenge; AI Cyber Challenge (AIxCC), accessed June 30, 2024, https://aicyberchallenge.com/.

[4] "Zellic Wins $1M From DARPA in the AI Cyber Challenge," Zellic, April 4, 2024. https://www.zellic.io/blog/zellic-darpa-aixcc/.

[5] Manish Bhatt et al., "CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models," arXiv, April 19, 2024. http://arxiv.org/abs/2404.13161.

es."[6] However, follow-on studies from the Project Zero security team at Google reported that they improved the performance of the LLMs through several principles, such as sampling and allowing the models access to tools, while still reporting that "substantial progress is still needed before these tools can have a meaningful impact on the daily work of security researchers."[7]

Drivers of opportunity

- **Domain-specific capability (vulnerability identification):** How good AI models are or could be at this task, especially compared to existing capabilities, such as fuzzing or static analysis tools. Any model that can identify vulnerabilities that current tools cannot find would have initial value as an improvement over today's baseline. Greater efficiency benefits will emerge the more AI models work to minimize both false positives and false negatives, as this will make capabilities more effective and reduce the need for human review of detections.

- **Integration with existing tools:** The more development workflows integrate AI vulnerability-finding tools, such as embedded into build processes or as part of code-hosting platforms like GitHub, the easier it will be for these tools to help detect vulnerabilities *before* the merge and rollout of code to customers, making bugs easier and less costly to fix.

- **Cost and availability:** Free or low-cost AI models or model-based tools could be particularly useful for organizations or individuals without significant resources dedicated to security reviews, such as for use in small businesses or for open-source software packages.

- **Education:** Ensuring that organizations know how to use vulnerability-finding tools and how to integrate them into their development process can help ensure that, as these tools develop, their benefits flow to defenders and, in particular, to those in less-resourced areas.

B. HELPING DEVELOPERS WRITE MORE SECURE CODE

Closely related to the question of finding and fixing vulnerabilities in existing code is the idea that AI tools that help developers generate code could help improve the security of that code by ensuring that its suggestions are free from known vulnerabilities. Despite the longstanding knowledge that certain common-class vulnerability patterns are insecure, these have recurred in code over many years.[8] Code-generating AI tools could potentially help avoid these patterns, either by training the underlying model to avoid insecure generations, such as through reinforcement learning from human feedback,[9] or by filtering model outputs for known insecure code patterns. One factor influencing LLM efficacy in this context is the type of secure coding or vulnerability discovery task assigned. Some flaws require a significant volume of context and might exceed what an LLM can accept. In other instances, model benchmarks could point to a specific code segment to propose mitigations in conjunction with human review.

Experiments on some of these techniques are already in process; in 2023, GitHub announced that its CoPilot code assistant would now include an "AI-based vulnerability filtering system" to filter out code results containing known insecure code patterns, such as those vulnerable to Structured Query Language (SQL) or path injection or the use of hard-coded credentials.[10] These tools could also have their use expanded to propose fixes—at a significantly greater speed than locating them—allowing for the integration of security review tooling based on LLMs into existing human development environments.

However, one should not assume that AI-generated code will be more secure, especially without further research and investment in this area. (The Risks section of this paper covers an early study indicating that the opposite may well have been true for one generation of LLMs.) Conducting security reviews of AI-generated code will likely require heavy human oversight limiting the throughput from even large-scale LLM deployments for software development.

The need exists for more evaluation and benchmarking to understand the security properties of AI-generated code, as compared to human code. This would offer developers and organizations defining information on how to integrate AI tools into their workflows, such as identifying contexts in which their use benefits security and pinpointing weaknesses or blind spots where developers should still thoroughly review AI-generated code for security flaws. For example, one could imagine using AI tools capable of identifying and avoiding common insecure patterns, such as a lack of input sanitization, but, consequently,

[6] Bhatt et al., "CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite."

[7] Sergei Glazunov and Mark Brand, "Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models," Google Project Zero (blog), June 20, 2024, https://googleprojectzero.blogspot.com/2024/06/project-naptime.html.

[8] "Secure by Design Pledge," US Cybersecurity and Infrastructure Security Agency (CISA), accessed June 30, 2024, https://www.cisa.gov/securebydesign/pledge; Isabella Wright and Maia Hamin, "'Reasonable' Cybersecurity in Forty-Seven Cases: The Federal Trade Commission's Enforcement Actions Against Unfair and Deceptive Cyber Practices." *Cyber Statecraft Initiative*, June 12, 2024. https://dfrlab.org/2024/06/12/forty-seven-cases-ftc-cyber/.

[9] AI models, which receive human feedback on their predictions, learn to generate outputs that receive more favorable feedback. See Paul Christiano et al., "Deep Reinforcement Learning from Human Preferences," arXiv, February 17, 2023, http://arxiv.org/abs/1706.03741.

[10] Anthony Bartolo, "GitHub Copilot Update: New AI Model That Also Filters Out Security Vulnerabilities," Microsoft (blog), Feb 16, 2023, https://techcommunity.microsoft.com/t5/educator-developer-blog/github-copilot-update-new-ai-model-that-also-filters-out/ba-p/3743238.

might generate code with more subtle design or logic errors that create new vulnerabilities.

Drivers of opportunity

- **Trustworthy AI outputs:** A first, vital prerequisite is that AI-generated code improves upon the security of human-written code in relatively consistent ways (and without causing human developers to neglect security concerns in their code more than is currently the case). The security improvements of AI code need not be absolute across contexts—AI-generated code does not need to be better than the best cryptography expert to help the average developer avoid SQL injection attacks. Thus, additional clarity in how and when to trust AI-generated code with respect to security would help ensure its appropriate adoption in different contexts. In addition to being secure, AI code suggestions must, at least, be moderately helpful to developers, if only to buoy wider adoption of the suggestions (and their potential security benefits).

- **Integration with existing tools:** The more that code-generating tools coalesce with integrated development environments (IDEs) and other environments where programmers can use them as part of their development workflows, the more expansive their potential adoption, which will increase tool leverage on other information, such as the broader context of a project to more accurately assess the security implications of the code they generate.

- **Cost and availability**: Many small developers, including open-source software maintainers, may likelier use free or widely available tools rather than expensive proprietary solutions. Ensuring that low-cost model solutions have strong security protections for the code they generate—not just expensive or leading-edge models—could benefit these developers.

- **Education:** Educating developers on the best ways to use AI code-generating tools, as well as how to verify the security of generated code, could also help ensure that these tools roll out in ways that maximize their potential benefits.

C. MAKING SENSE OF CYBERSECURITY DATA

In addition to using the code-analysis and code-generation features of AI to improve the security of software code, another relatively well-developed current use case for AI in cybersecurity is the idea of using AI to help with cybersecurity-relevant data processing. For example, AI tools could help sort through data generated by computer systems, such as system logs, to help identify or investigate cyberattacks by identifying anomalous behavior patterns and indicators. Likewise, AI tools could help process and analyze cyber threat intelligence or information about vulnerability disclosures to help defenders respond to this information and prioritize follow-up actions.[11] These systems may incorporate generative AI but might also follow entirely separate AI paradigms, like supervised machine learning.

Drivers of opportunity

- **Domain-specific capabilities (anomaly detection):** The degree to which AI systems can correctly identify anomalies or other relevant information from system data. Both false negatives and false positives would be harmful in this situation, though false negatives, perhaps more so.

- **Integration with existing data and tooling:** How well can new AI solutions integrate with existing security tooling to access the panoply of data required to do anomaly detection? Is there adequate high-quality available to train these models in the first place?

- **Cost and availability:** Free or low-cost models or tools could be particularly useful for organizations or individuals without significant resources to operate their own security operations center (SOC) teams and similar.

- **Education:** Helping organizations, particularly those with fewer resources, understand how to use and configure these tools can help them harness the efficiencies—and avoid hoodwinking by tools that make big promises but then deliver little in terms of increased security.

D. AUTOMATION OF OTHER CYBERSECURITY TASKS

Beyond these well-developed categories, there are other examples of often neglected cybersecurity tasks, which, if improved or eased using AI, would provide benefits to security. One example is the failure to "timely" apply patches and version upgrades to software within a network. These patches and version upgrades often contain important security updates, but many organizations are slow to patch, whether due to resource constraints or negligence. Another related example is consistently upgrading dependencies in software packages to address upstream vulnerabilities.

Further afield suggestions include the idea of having AI systems, including agents, that can automate longer action sequences in cyber defense, such as systems that can identify an anomaly and then autonomously take action, such as quarantining affected systems. Such autonomy is likely beyond the capabilities of cur-

[11] "CISA Artificial Intelligence Use Cases," US Cybersecurity and Infrastructure Security Agency (CISA), accessed June 30, 2024, https://www.cisa.gov/ai/cisa-use-cases.

rent GAI models, and some researchers have suggested creating "cyber gyms" to help train reinforcement learning agents for these kinds of tasks through trial and error.[12]

Drivers of opportunity

- **Trustworthiness:** Once operators seek to delegate tasks to AI systems (rather than asking the system to make a suggestion for a human operator to action), it becomes more important to have a very good sense of the accuracy and robustness of the model. For example, an AI patch management system that can modify and control arbitrary elements of a corporate network requires high-level trust protocols that it will not take spurious or destructive actions. This contrasts with many of the other opportunities identified, which envision a human-in-the-loop.

- **Openness and availability for experimentation:** The more different researchers and organizations experiment with models of how to implement AI into the defensive cyber process, the more likely it becomes that a product or service of genuine value might emerge to help use LLMs to automate additional tasks in cybersecurity.

AI RISKS IN THE CYBERSECURITY CONTEXT

Broadly, the risks posed by AI in the cybersecurity context fall into at least two categories: risks from malicious misuse (e.g., the use of models to create outputs useful for malicious hacking) and risks to *AI users* arising from their well-intentioned use (e.g., cyber harms created when models generate incorrect or harmful outputs or take incorrect or harmful actions). Notably, this second category of risks to AI users tightly connects with many of the potential benefits outlined above.

A. RISKS FROM MALICIOUS MISUSE: HACKING WITH AI

The broadest category of malicious misuse risks in the cyber context is the potential for malicious actors—whether high-capability entities like the United States, Israel, or Russia or the most lackadaisical cybercriminal—to use generative AI models to become more efficient or more capable hackers.

Previous work published by the Cyber Statecraft Initiative on this topic "deconstructs" this risk by breaking "hacking" into constituent activities and examining GAI's potential utility for assisting with both making capable players better and bringing new malicious entrants into the space.[13] It seems possible, and likely, that all kinds of hackers could use GAI tools for activities including reconnaissance or information gathering, as well as assistance with coding and script development. Indeed, OpenAI reported disrupting threat actors who were using their models to conduct research into organizations and techniques and tools, generate and debug scripts, understand publicly available vulnerabilities, and create material for phishing campaigns.[14]

These risks are already here. What is less clear is whether or not these risks are acceptable and bearable. The OpenAI case shows that GAI is arguably a useful tool for hackers, but not necessarily that it provides a step change in terms of sophistication or capability. Tools like Google, after all, are also a benefit to hackers. The essential question is: where to draw the line?

This research recommends a few areas where GAI capabilities could create more profound capability improvements for malicious hackers.

- Models that can generate content for highly sophisticated social engineering attacks, such as creating deepfakes that impersonate a known figure for the purpose of carrying out an attack.

- Models that can identify novel vulnerabilities and develop novel exploits in code at an above-human level.

- AI-based "agents" with the ability to string together multiple phases of the cyberattack lifecycle and execute them without explicit human intervention, providing significant benefits in terms of speed and scalability as well as challenging typical means of detecting malicious activity such as looking for connections to a command and control server.

Thus, the risk that hackers will use GAI is not speculative—it is here. The issue, instead, is how much this usage increases risks to businesses, critical infrastructure companies, government networks, and individuals.

Drivers of risk

- **Deepfakes:** The ability for GAI systems to generate realistic-looking content that impersonates a human being, which the people interacting with it cannot distinguish or identify as machine-generated.[15]

[12] Andrew Lohn, Anna Knack, Ant Burke, and Krystal Jackson, "Autonomous Cyber Defense: A Roadmap from Lab to Ops," Center for Security and Emerging Technology (CSET), June 2023, https://cset.georgetown.edu/publication/autonomous-cyber-defense/.

[13] Maia Hamin and Stewart Scott, "Hacking with AI," *Cyber Statecraft Initiative*, February 15, 2024, https://dfrlab.org/2024/02/15/hacking-with-ai/.

[14] "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors," OpenAI, February 14, 2024, https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/.

[15] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova, "Truth, Lies, and Automation," Center for Security and Emerging Technology (CSET), May 2021, https://cset.georgetown.edu/publication/truth-lies-and-automation/.

- **Domain-specific capabilities (vulnerability identification and exploitation):** The ability for models, especially those fine-tuned on relevant datasets and actions, to display above-human level performance at specific high-risk activities, such as identifying novel vulnerabilities.

- **Domain-specific capabilities (autonomous exploitation):** The ability of models to string together and execute complex action sequences—particularly, though not exclusively, in the form of generating and executing code—to compromise an information system end-to-end.

- **Integration with existing tools:** Studies appear to suggest that integrating AI models with tools such as code interpreters can upskill these models,[16] which could increase the risk that they can be useful to hackers.

- **Removal of safeguards:** It is very challenging to create blanket safeguards that prevent bad behavior while protecting legitimate use cases, in part because of the similarity between malicious and benign activities. Developers call this the "safety-utility tradeoff." At the same time, models do currently refuse to comply with overtly malicious requests and appear to be improving in their ability to do so over time—thus, models without any safeguards at all or those fine-tuned for malicious cyber activity could lose even these modest protections.

B. RISKS TO AI USERS

Risks to AI users depend much more heavily on the context and purposes of the model's, or its outputs' use, as well as the type or nature of safeguards and checks implemented within that environment. Some of the key contexts and activities in which AI can create cyber risks to users include the use of AI-generated code, the use of systems where AI agents may have access to user devices and data, and the use of AI in defensive cybersecurity systems.

B1. RISKS OF INSECURE AI-GENERATED CODE

In one initial study on the security properties of AI-generated code, published by Stanford, researchers split developers into two groups, gave only one group access to code-assist tools, then observed the developers during the process of solving coding problems and examined the security of the resultant code.[17] They found that "participants who had access to an AI assistant … wrote significantly less secure code than those without access." For example, only 3 percent of programmers in the group with the AI assistant implemented an encryption/decryption function in a way that the researchers categorized as "secure," compared to 22 percent of programmers working alone who generated a "secure" solution. The researchers surveyed the developers and found that, of the developers using the AI assistant, those who reported placing less subjective trust in the AI assistant were more likely to generate "secure" code. Additionally, the researchers found that code labeled "secure" had, on average, a larger "edit distance," (e.g., more changes from initial AI-generated code than did "insecure" or "partially secure" solutions).

While it is possible, and perhaps even likely, that the assistant's properties have evolved since this point, this example illustrates the need to better understand the security properties of AI-generated code before developers embed it deeply into their workflows. Policymakers can help hold companies to account on this question.

Drivers of risk

- **Untrustworthy outputs:** The risks from AI-generated code are greatest when the developer is incapable of, or unlikely to, validate the output themselves or if there is no process of human oversight over the generated code. That is, risks become acute when there is a mismatch between the trust that a developer *thinks* they can place in AI-generated code and the level of trust that is actually appropriate. These levels may vary across contexts, as different kinds of code are more or less security sensitive—for example, deploying a web app has fewer opportunities to go wrong than implementing a cryptographic library—or AI models may be better or worse at generating it securely by virtue of having seen more or fewer examples. These risks necessitate the development of robust benchmarks that measure the security properties of AI-generated code across a variety of contexts.

- **Misplaced user trust:** If users verify the security of generated code themselves and to their own standards, the risks that the code will be insecure significantly lessen. Much of the problem thus stems from users placing unearned trust in model outputs. Yet, pointing the blame finger back at the user is not an appealing path for policy, Moving forward, users will place trust in automated systems, and therefore, it is up to the makers of those systems and policymakers alike to help ensure that the systems are fit to deserve that trust.

B2. RISKS FROM INTEGRATED AI SYSTEMS WITH DATA OR SYSTEM ACCESS

There is a lot of interest in connecting GAI models to environments that give them the tools to automate tasks—rather

[16] Glazunov and Brand, "Project Naptime: Evaluating Offensive Security Capabilities."

[17] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh, "Do Users Write More Insecure Code with AI Assistants?" In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, (November 2023), 2785–99, https://doi.org/10.1145/3576915.3623157.

than feeding output to a human to do a task; leading to more autonomous agents. Such conditions create cybersecurity risks because many AI models are very vulnerable to adversarial attacks that can cause them to do strange and potentially undesirable things, including compromising the security of the system they operate or the data they have access to.

From stickers on stop signs that can fool computer vision algorithms to "jailbreak" prompts that can convince LLMs to ignore their creator-imposed safeguards, [18,19] it is hard to ensure that AI systems solely do what you want them to do. Many leading models have proven vulnerable to "prompt injections,"[20] which allow a user (or a potential attacker) to get around security limitations, including to obtain hidden information. Researchers have already demonstrated that, by embedding hidden text on their webpage, they can manipulate the results of GAI model outputs.[21] If users interact with a model that has access to sensitive data, such as a business database or sensitive files on a user's computer, they might be able to use prompt engineering to trick the model into handing that information over. Or, people could create malicious websites that, when an autonomous agent scrapes them, contain hidden commands to obtain and leak data or damage the machine they are operating.

These risks grow as developers embed AI systems into higher stakes systems that grant access and authorization to take ever more sensitive actions. Cybersecurity experts have highlighted reliability as a core concern to using AI models as a component of cybersecurity defense, and they stressed the need to deploy models and grant them autonomy in ways proportional to the organizational context in which they operate and the risks associated.[22]

Drivers of risk

- **Untrustworthy outputs:** Outputs by models that misalign with the goals or needs of their human operators, whether insecure code, harmful outputs as the result of prompt injection, or unsafe decision-making in the cyber context.
- **Misplaced user (or system) trust:** When users or information systems embed a model into a context with more trust and permissions than the model deserves based upon its own reliability.
- **Increased delegation / lessened supervision:** The integration of models into contexts without sufficient, or no, oversight before placing their outputs into "use" (e.g., code merged into a product or security action taken).

DUAL DRIVERS

The opportunity and risk drivers outlined above are not always diametrically opposed. Were they, it would offer an easy remedy for policy: do more of the "opportunity" drivers and less of the "risk" drivers. Instead, as the next sections illustrate, the close coupling between many of these drivers will challenge policy's ability to neatly extricate one from the other.

DOMAIN-SPECIFIC CAPABILITIES

Particular domain-specific capabilities for AI models would drive both opportunity and risk in the cyber context. For example, the ability to find novel vulnerabilities would benefit defenders by helping them identify weaknesses to patch and malicious actors searching for footholds into software systems. To a lesser degree, the same is true of the general ability that models would have to write complex, correct code—this ability could offer efficiency benefits to developers, whether they are open-source maintainers or ransomware actors. It seems unlikely that these capabilities would advance in ways that only benefit the "good guys." While model safeguards could help reject obvious malign requests (e.g., ask a model to help them write an urgent email), in the wider cyber context, bad actors are on an endless search for reasonable justifications to test for and seek vulnerabilities in a codebase. No currently known software can develop a foolproof way to see inside its operator's heart to discern their true intent. Instead, it is likely that policy will simply have to accept these twinned risks, seeking to measure them as they progress and find ways to make it as easy as possible for defenders to implement new technologies in hopes that they can outpace malicious actors. This is an uneasy balance, but it is also one that is deeply familiar in information security.

TRUST AND TRUSTWORTHINESS

Perhaps the single largest driver of AI opportunity in the cybersecurity context is model "trustworthiness"—that is, the degree

[18] Evan Ackerman, "Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms," IEEE Spectrum, August 2017, https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms.

[19] Melissa Heikkilä, "Three Ways AI Chatbots Are a Security Disaster," *MIT Technology Review*, April 3, 2023, https://www.technologyreview.com/2023/04/03/1070893/three-ways-ai-chatbots-are-a-security-disaster/.

[20] Bhatt et al., "CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite."

[21] Arvind Narayanan (@random_walker), "While Playing around with Hooking up GPT-4 to the Internet, I Asked It about Myself… and Had an Absolute WTF Moment before Realizing That I Wrote a Very Special Secret Message to Bing When Sydney Came out and Then Forgot All about It. Indirect Prompt Injection Is Gonna Be WILD Https://T.Co/5Rh1RdMdcV," X, formerly Twitter, March 18, 2023, 10:50 p.m., https://x.com/random_walker/status/1636923058370891778.

[22] Anna Knack and Ant Burke, "Autonomous Cyber Defence: Authorized Bounds for Autonomous Agents," Alan Turing Institute, May 2024, https://cetas.turing.ac.uk/sites/default/files/2024-05/cetas_briefing_paper_-_autonomous_cyber_defence_-_authorised_bounds_for_autonomous_agents.pdf.

to which a model or system that integrates AI produces outputs that are accurate, reliable, and “fit for purpose” in a particular application context. For example, if a model can regularly generate code that is secure, free of bugs, and does exactly what the human user intended, it might be trustworthy in this context.

A model’s trustworthiness almost directly controls the potential productivity benefits it can deliver by dictating whether a human must essentially run “quality control” on model outputs, such as carefully reviewing all generated code or all processed data to ensure the model did not make a mistake or miss an important fact. For example, a completely untrustworthy model saves no time (and may, in fact, waste it) because its work requires manual duplication; theoretically, a perfectly trustworthy model should not need human oversight. In practice, human oversight (whether manual or automated) in some fashion must bridge this imperfect trust. Moreover, it is important that the humans or systems performing this oversight have a good understanding of the level of oversight needed and avoid the complacency of overly trusting the system’s outputs.

Trust is not a single benchmark but a property dictated by context. Different contexts have distinct requirements, acceptable performance levels, and potential for catastrophic errors. What matters is that the operator has an appropriate way to measure the model’s trustworthiness within a specific task context and determine its respective risk tolerances, then compare both to ensure they align. Policymakers and businesses alike should review the varied levels of criticality for AI application contexts and be specific as to both how to define the properties that a model would need to be trustworthy in each context and how to measure these properties.

Developing better ways to measure model trustworthiness and make models more trustworthy will, for the most part, unlock opportunity. However, this factor is in the twinned risk section because, undeniably, trusting a model creates risk. The more a model has delegated tasks without stringent oversight, the greater the productivity gains—and the greater the stakes are for its performance and robustness against attack. Notably, in the cybersecurity context, embedding AI systems into broader information systems, while they remain vulnerable to adversarial inputs, creates the risk that these models could become potent vectors for hacking and abusing systems into which they integrate. In this area, it will be vitally important to benchmark and understand AI models’ vulnerability and to develop security systems that embed AI models in ways that account for these risks.[23] Without better ways to measure risk before models become embedded into sensitive contexts, there is a risk that AI systems will develop their own kind of “Peter Principle” (i.e., AI models embedded into increasingly high-trust situations until they prove they have not earned that trust).

OPENNESS

Many of the most acute benefits that GAI systems can provide in cybersecurity will come from using such systems to reduce the labor required to perform security tasks, from auditing code packages to monitoring system logs. The more open innovation there is, the more tools there will be. And the more these tools have accessible price points, the likelier it will be that less-resourced entities will use them. Competition and, in particular, the availability of open-source models can encourage innovation and experimentation to build these tools and keep costs relatively low. Open models can also benefit some of the key questions of trust that are core to AI opportunity and risk: open models are easier to experiment with and customize, making it easier for users and researchers alike to measure the trustworthiness of models in particular contexts and to customize models to meet their specific trust needs. These models are growing ever larger and also more powerful. Cohere AI recently released a 104 billion parameter model through Hugging Face.[24] Open models can also contribute to higher levels of trustworthiness, allowing developer-led organizations to validate model behavior under different conditions and tasks with more control of model versions and constraints.

At the same time, expanded access to capable models—and, in particular, open-source models—may create additional challenges in preventing model misuse. Open models foreclose abuse-preventing tools, such as monitoring application programming interface (API) requests, and allow users to remove safeguards and protections through fine-tuning. The science of safeguards and their relative strengths and weaknesses needs further study to make the case that open models create significantly more “marginal risk” than closed models.[25] For example, in the cyber context, even reasonably designed safeguards may be unable to stop hackers from appropriating reasonable outputs, such as email text or scripts seeking more malign ends. However, safeguards may be more impactful when it comes to contexts like embedding watermarks in AI-generated content and similar. As model capabilities and safeguarding techniques advance, the marginal risk posed by open models may increase.

[23] Caleb Sima, “Demystifing LLMs and Threats.” *Csima* (blog), August 15, 2023, https://medium.com/csima/demystifing-llms-and-threats-4832ab9515f9.

[24] Cohere 4 AI, “Model Card for Cohere 4 AI Commanr R+”, May 23, 2024, https://huggingface.co/CohereForAI/c4ai-command-r-plus

[25] Sayash Kapoor et al., “On the Societal Impact of Open Foundation Models,” February 27, 2024, https://arxiv.org/pdf/2403.07918v1.

ASYMMETRIC DRIVERS

At the same time, there are some factors likely to drive primarily risk or primarily opportunity in the cybersecurity context. These asymmetric drivers of risk and opportunity make promising areas for policy intervention.

RISK: DEEPFAKES AND IMPERSONATION

There are few legitimate reasons why AI models should need to generate content that imitates a person (especially an actual person) without appropriate disclosures that this content is not real. This is true across images, video, and voice recordings. Policy could knock out a series of easy wins by focusing on requiring disclosures and making AI-generated media easier to identify. Already, a bevy of proposed state initiatives exist, which, if enacted, will mandate disclosing AI-generated media in contexts from political advertising to robocalls,[26] and federal lawmakers could unify these requirements with legislation to apply them consistently whenever consumers interact with advertising or businesses. Laws will not stop criminals, of course—for that, the government may need to invest in technical research to embed watermarks into AI-generated content and to help electronic communication carriers like voice and video calling implement systems for detecting faked content. This work will not be easy, requiring novel research and development as well as implementation across a variety of parties. Nonetheless, the government is the best-positioned actor to coordinate and drive this forward.

OPPORTUNITY: EDUCATION

Another clear opportunity is investing in ways to educate different users who will interact with and make decisions about AI—from business leaders to developers—about how to use AI in responsible and reasonable ways. This kind of education can increase the uptake of AI, where it can be helpful, while also providing an opportunity to prime these users to consider specific kinds of risks, from the need to review AI-generated code to the security risks of embedding AI systems that might be vulnerable to prompt injection.

OPPORTUNITY: MEASURING TRUSTWORTHINESS

The more that operators have a grounded sense of models' strengths and weaknesses, the more they can build applications atop them that do not run the risks of strange and unexpected failures. Policy can help steer and incentivize the development of ways to measure relevant aspects of model trustworthiness, such as a model's accuracy (best defined in a specific context), its security and susceptibility to adversarial inputs, and the degree to which its decisions allow audits or reviews after the fact. Better measurements will unlock better usage with fewer risks. And they will enable the government to step in and demand clear standards for certain high-risk applications.

DRIVERS OF RISK AND OPPORTUNITY IN CONTEXT

Many of the drivers of risk and opportunity draw from the unique characteristics of this moment in AI. Understanding the story of how we got to this moment, alongside identifying some specific meta-trends that characterize it, can help policymakers comprehend the drivers of risk and opportunity as well as how they are likely to change in the future.

Deeply Unsupervised

The first trend is the rise of unsupervised learning, alongside its resulting highly capable generalist models. The field of AI has seen the rise and fall of multiple different paradigms throughout its lifetime, with generative AI representing the next instantiation of a longer-running trend in the field toward systems that learn to make sense of data themselves using patterns and rules that are increasingly opaque to their creators.

Many early attempts to build artificially intelligent systems focused on programming complex, pre-determined rules into computer systems. These systems could be surprisingly capable: in 1966, the first "chatterbot," Eliza, used simple language-based rules to emulate responses from a mock therapist, with its creator finding that "some subjects have been very hard to convince that Eliza (with its present script) is not human."[27] And, in 1997, the computer Deep Blue outplayed world chess champion Garry Kasparov using brute-force computation and a complex set of rules provided by chess experts.[28] Yet, these systems lacked at least one key characteristic of intelligence: the ability to learn.

Decades before these rule-based approaches, research into how the human brain works through the interconnection and firing of neurons inspired the invention of another paradigm: neural networks.[29] The weights in neural networks—updated over time by an algorithm that seeks to reduce the error between the net-

[26] Bill Kramer, "Transparency in the Age of AI: The Role of Mandatory Disclosures," Multistate, January 19, 2024. https://www.multistate.ai/updates/vol-10.

[27] Ben Tarnoff, "Weizenbaum's Nightmares: How the Inventor of the First Chatbot Turned against AI," *Guardian*, July 25, 2023, https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.

[28] IBM, "Deep Blue," accessed June 30, 2024, https://www.ibm.com/history/deep-blue.

[29] Warren S McCulloch and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5 (1943), https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf.

work's prediction and reality—allow neural networks to learn rules, patterns, and relationships not explicitly specified by their creators. While neural networks fell out of favor during a long "AI Winter," they began to recur in the nascent field of machine learning, which focused on developing statistical algorithms that could learn to make predictions from data.

Initially, machine learning focused primarily on supervised learning, a paradigm in which a model tries to learn relationships between input data (such as images or numerical and financial data) and output labels (such as names of items in an image or future price projections). Supervised learning with increasingly deep neural networks proved very successful for tasks like image classification, predictive analyses, spam detection, and many other tools developed during the 2000s and 2010s.

In contrast, current generative AI systems receive their training, at least in large part, through unsupervised learning, a different paradigm in which a model reviews an immense amount of unlabeled data, such as raw text, and learns to cluster or predict that data without explicit human-provided labels (or target predictions). LLMs, like OpenAI's Generative Pre-trained Transformer (GPT) series, are huge neural networks trained on trillions upon trillions of words of text data, much of which comes from scraped internet sites and digital books.[30] Interestingly, these models still learn by making predictions and receiving error signals to correct their prediction functions—but instead of learning to predict human-generated labels, they learn to predict patterns and structure in human-generated data (text) itself.

Unsupervised learning has increased the capacity of models, producing technologies, like ChatGPT, that can and have dazzled users and researchers alike with their capabilities. It has also created systems that are more challenging for developers, researchers, policymakers, and users to understand. Rules-based systems were definitionally transparent. Deep learning was perhaps the first indication that subsequent AI systems might bring opaque internal logic that defies easy interpretation. However, supervised approaches have still provided some clear ways to evaluate model performance within a specific domain. New unsupervised models are challenging to interpret and evaluate. Their capabilities emerge through testing and scale rather than explicit design.[31] The emergence of these models preceded the development of empirical ways to test their capabilities across many of the domains they likely have skills. Harnessing the opportunity and avoiding the risks of these highly general models will require developing new ways to think about model explainability and *new ways* to evaluate model capabilities across the varied tasks and contexts, where their use is not only probable but also possible.[32]

Ravenous Demand for Compute and Data

The second trend focuses on the ways in which the intensive compute and data needs of the latest generation of AI model development have made current systems highly proximate to concentrated power in the hands of large technology companies.

Current leading-edge models are big.[33] Size defines the computing costs associated with training a model, namely, the size of its training dataset and the size of the model itself (often measured as the number of "parameters"). Both of these have grown ever larger and the compute required to train these massive models is expensive.[34] At present, the well-capitalized and semi-commercial players (e.g., OpenAI, Meta, and Google) build most of the leading models. This creates a different paradigm than that of previous iterations of AI or machine learning systems, which more often emerged from research and academic settings. The computational and data costs of large-model development have tied the evolution of AI models to other existing technology infrastructures, especially cloud computing, with major providers to deliver, in part, the required compute (e.g., the Amazon and Microsoft partnerships with leading generative AI labs).[35] Likewise, access to text data for training models has become a point of leverage. Sites like Reddit and Twitter that host lots of public text have begun charging for API access to data,[36] as users question whether their technology providers take advantage of private data to train AI models (major model providers say they use only public data).[37]

[30] Dennis Layton, "ChatGPT – Show Me the Data Sources," Medium (blog), January 30, 2023, https://medium.com/@dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8.

[31] Jason Wei et al., "Emergent Abilities of Large Language Models," arXiv, October 26, 2022, https://doi.org/10.48550/arXiv.2206.07682.

[32] Leilani H. Gilpin et al., "Explaining Explanations: An Overview of Interpretability of Machine Learning," arXiv, February 3, 2019, http://arxiv.org/abs/1806.00069.

[33] Anil George, "Visualizing Size of Large Language Models," Medium (blog), August 1, 2023, https://medium.com/@georgeanil/visualizing-size-of-large-language-models-ec576caa5557.

[34] Jaime Sevilla et al., "Compute Trends Across Three Eras of Machine Learning," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, (2022), 1–8, https://doi.org/10.1109/IJCNN55064.2022.9891914.

[35] Amazon Staff, "Amazon and Anthropic Deepen Their Shared Commitment to Advancing Generative AI," March 27, 2024. https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment; "Microsoft and OpenAI Extend Partnership," Official Microsoft Blog, January 23, 2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.

[36] Mike Isaac, "Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems," *New York Times*, April 18, 2023, https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html.

[37] Eli Tan, "When the Terms of Service Change to Make Way for A.I. Training," *New York Times*, June 26, 2024, https://www.nytimes.com/2024/06/26/technology/terms-service-ai-training.html.

The pressures for large labs to rapidly commercialize these systems and to recoup their investments may drive both opportunity and risk—opportunity because there will be well-capitalized machines seeking to build functional applications and use cases for these models; risk because these companies will face tremendous pressure to create product offerings from these models, regardless of their shortcomings. Closed and for-profit paradigms may make it harder for independent researchers and outsiders to access models to evaluate them and expose their weaknesses—while large labs have definitely allowed some level of access,[38] for which they should be commended, it is hard to know exactly what the limits of this access and of researchers' ability to publicly report adverse findings are. While open-source models help bridge some of this gap, this paradigm only works if open-source models are at relative parity with closed-source ones, which may not have guarantees.[39]

New Stakeholders

The third trend—and an important caveat to the second trend—is how the popularity and accessibility of natural language interfaces for AI models have brought a new wave of AI stakeholders into the ecosystem. Even people with no technical background can easily interact with tools like ChatGPT, Bard, and the Bing chatbot through prompts written in English (or other languages) rather than computer code. Consumers, hacker-builders, entrepreneurs, and large companies—alike—expand and help develop new potential use cases for AI. Significant application development activity is also happening based on open-source and publicly available models, led by platforms like Hugging Face and the decision by Meta to publicly release its Llama models. This distributed innovation environment creates the potential for AI's benefits to disperse more widely and in a more decentralized way than were innovations, such as the large internet platforms of the 2000s. At the same time, this decentralization will increase the challenge for regulators seeking to set standards around the development and use of AI applications, in much the same way as regulators have struggled to define functional and universal standards for software security because of software's heterogeneous and decentralized nature.

CONCLUSIONS: WHOSE RISKS, WHOSE OPPORTUNITY?

Advances in AI will bring both opportunity and risk. The key question for policymakers is not how to get only opportunity and no risk—this seems all but impossible. Instead, it is one of recognizing and seeking to balance who must deal with each. Models that can write more trustworthy and reliable code will help open-source maintainers and other organizations better shore up security—and help novice hackers write scripts and tools. Both defenders and cybercriminals will use models that can find vulnerabilities. Models that integrate into workflows entrusted to make decisions can deliver the benefits of machine speed and scale, while creating risks because humans can no longer perfectly oversee and interpret their decisions.

With many of these cases, such as vulnerability hunting and coding, policymakers' best option may simply be to try to encourage enterprises to build and adopt these tools into their workflows and development processes faster than they end up as common tools for malicious hackers. For certain other cases, as with deepfake-based impersonations, it may be possible to push model developers to implement tailored protections that can asymmetrically reduce their abuse potential while preserving their benefits. And, in general, policymakers can seek to develop incentives and support for the development of best practices, tools, and standards for AI assurance, to encourage enterprises and organizations to apply appropriate scrutiny in their adoption of AI, and to hold them to account when they fail to do so.

Policymakers might also consider ways to shift more of the costs of safely integrating AI – ways of measuring trust and mitigating risk—onto the makers of these systems. The history of the debate over software liability illustrates the peril of allowing technology vendors to reap the profits from selling technology without facing any consequences when that technology proves unfit for the purpose for which they sold it.[40] The debate over software liability has raged for decades.[41] Maybe the advent of AI provides an opportunity to adopt a new paradigm a little sooner.

The balance of risk and opportunity for the end users of technology should be a primary concern for policymakers; how the market and policy equip cybersecurity defenders will play a significant role in determining that balance. Thus, there remain plenty of opportunities (and risks) for policymakers to evaluate in these next formative years of AI policy.

[38] "OpenAI Red Teaming Network," accessed June 30, 2024, https://openai.com/index/red-teaming-network/.

[39] Xiao Liu et al., "AgentBench: Evaluating LLMs as Agents." arXiv, October 25, 2023, http://arxiv.org/abs/2308.03688; "LMSys Chatbot Arena Leaderboard," Hugging Face, accessed June 30, 2024, https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard; "SEAL Leaderboards," Scale, accessed June 30, 2024, https://scale.com/leaderboard.

[40] Bruce Schneier, "Liability Changes Everything," November 2003, https://www.schneier.com/essays/archives/2003/11/liability_changes_ev.html.

[41] Maia Hamin, Sara Ann Brackett, and Trey Herr, "Design Questions in the Software Liability Debate," Cyber Statecraft Initiative, January 16, 2024, https://dfrlab.org/2024/01/16/design-questions-in-the-software-liability-debate/.

ABOUT THE AUTHORS

Maia Hamin is currently serving an assignment under the Intergovernmental Personnel Act at the US AI Safety Institute within the National Institute of Standards and Technology (NIST). She is on leave from the Cyber Statecraft Initiative, where she held the position of associate director with the Cyber Statecraft Initiative, part of the Atlantic Council Tech Programs. Hamin's contributions to this work predate her NIST assignment, and the views expressed in this paper do not represent those of the AI Safety Institute.

Jennifer Lin is a former Young Global Professional with the Atlantic Council's Cyber Statecraft Initiative, part of the Atlantic Council Tech Programs. During her time with the team, she was a sophomore at Stanford University double-majoring in political science and symbolic systems, with a specialization in artificial intelligence.

Trey Herr is senior director of the Cyber Statecraft Initiative (CSI), part of the Atlantic Council Technology Programs, and assistant professor of global security and policy at American University's School of International Service.

ACKNOWLEDGMENTS

Thank you to NVIDIA for support for this work. Thank you to the CSI team for support on this project as well as Charlette Goth-Sosa and Donald Partyka for editing and production. Thank you also to several reviewers at different stages of drafting including Harriet Farlow, Chris Wysopal, Kevin Klyman, and others who wish to remain anonymous.

Atlantic Council | Board of Directors

The Atlantic Council is a nonpartisan organization that promotes constructive US leadership and engagement in international affairs based on the central role of the Atlantic community in meeting today’s global challenges.

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