



QUENTIN E. HODGSON, KAMARIA HORTON, MATTHEW J. MALONE

# Facing the Artificial Intelligence–Cyber Nexus

A Structured Approach to Government Decisionmaking to Address Emerging Artificial Intelligence Capabilities for Cyber Attacks



For more information on this publication, visit [www.rand.org/t/RRA4250-1](http://www.rand.org/t/RRA4250-1).

#### **About RAND**

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit [www.rand.org](http://www.rand.org).

#### **Research Integrity**

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit [www.rand.org/about/research-integrity](http://www.rand.org/about/research-integrity).

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

© 2025 RAND Corporation

RAND® is a registered trademark.

*Cover: Just\_Super/Getty Images.*

#### **Limited Print and Electronic Distribution Rights**

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on [rand.org](http://rand.org) is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit [www.rand.org/about/publishing/permissions](http://www.rand.org/about/publishing/permissions).

# About This Report

In this report, we aim to inform policymakers on how to prepare for the emergence of increasingly capable artificial intelligence (AI) systems that can plan and execute offensive cyberspace operations (OCO). Although the precise capabilities and timeline of AI-enabled OCOs are unknown, policymakers should anticipate that malicious actors will employ AI in cyberattacks and establish a decisionmaking framework to prevent or mitigate risks from these threats. In this report, we provide guiding questions to facilitate the decisionmaking process for policymakers and policy actions, including private-sector engagement, diplomatic engagement, law enforcement, finance, commerce, the military, and intelligence. The report offers recommendations for policymakers on next steps, including continuing wargame scenarios, expanding the pool of government AI expertise and access, strengthening the resilience of critical infrastructure, and integrating AI responsibly into cyber defenses.

## Meselson Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Meselson Center, which is dedicated to reducing risks from biological threats and emerging technologies. The center combines policy research with technical research to provide policymakers with the information needed to prevent, prepare for, and mitigate large-scale catastrophes. For more information, contact [meselson@rand.org](mailto:meselson@rand.org).

## Funding

This research was independently initiated and conducted within the Meselson Center using gifts for research at RAND's discretion from philanthropic supporter Open Philanthropy, as well as gifts from other RAND supporters and income from operations. RAND donors and grantors have no influence over research findings or recommendations.

## Acknowledgments

Our thanks to the participants in the AI-cyber workshop for their active engagement and thoughtful insights. We thank David Glickstein for his stellar work developing workshop materials and helping us refine this report. Max Tegmark's presentation to RAND on artificial general intelligence inspired our approach to understanding levels of AI-enabled cyber. We are also grateful to Ben Buchanan and Chad Heitzenrater for their careful review and constructive feedback.

# Summary

As artificial intelligence (AI) grows increasingly capable, emerging systems could enable offensive cyberspace operations (OCO). In this report, we provide policymakers with a structured approach to understanding the risks of AI-enabled OCOs and identifying policy options best suited to prevent or mitigate these risks and potential impacts. Our focus is on informing policymakers in order to help them anticipate the emergence of such capabilities and take a more active anticipatory posture, as opposed to responding to an AI-enabled OCO after the fact. Responding to a cyberattack, whether AI-enabled or otherwise, involves processes and considerations we are not addressing in this report.

We present a structured approach to decisionmaking. We provide guiding questions to help policymakers determine the best means to address the potential threat of more-advanced AI-enabled cyberattacks operating at a greater speed and scale or sophistication than what is currently possible, each with helpful subquestions, including the following:

- What is the assessed risk to the United States or its allies?
- What are the goals of U.S. policy?
- What are the key factors that might influence the availability and effectiveness of policy actions?
- What policy actions are available to address the risk?

We then provide an overview of policy actions for preventing AI-enabled cyberattacks or mitigating the risks they may pose. These span private-sector engagement, diplomatic engagement, law enforcement, finance, commerce, the military, and intelligence. For each, we present key considerations and risks policymakers should weigh.

We then outline the next steps policymakers may take to advance preparations for threats stemming from emerging capabilities for AI-enabled OCOs. We encourage policymakers to

- **Continue to wargame scenarios:** Hold workshops with government officials spanning national and homeland security, law enforcement and commerce, the private sector, and allies to practice decisionmaking for addressing AI-enabled cyber threats and explore diverse scenarios, including fully autonomous OCOs and varying attribution.
- **Refine guiding questions and responses:** Review and expand the decisionmaking framework's guiding questions and options to ensure that they remain comprehensive and actionable in real time.
- **Build expert networks:** Bolster government AI expertise and widen access across agencies to support faster, better-informed decisions and identify uncertainties.
- **Boost infrastructure resilience:** Enhance critical infrastructure's ability to withstand AI-enabled cyberattacks at greater speed, scale, or scope.

- **Adopt AI responsibly in defenses:** Support responsible AI-enabled defenses through incentives, test beds, certifications, and reduced regulatory barriers consistent with the White House AI Action Plan.

# Contents

About This Report..... iii  
Summary .....iv

Facing the Artificial Intelligence–Cyber Nexus ..... 1  
    Current and Future Landscape of Artificial Intelligence–Enabled OCOs..... 2  
    The ART Framework: How We Conceptualize Artificial Intelligence Capabilities for OCOs..... 6  
    Informing a Structured Approach to Government Decisionmaking Through Key Considerations and a  
        Workshop ..... 11  
    A Structured Approach to Government Decisionmaking on Preventing and Mitigating the Risks of  
        Artificial Intelligence–Enabled OCOs..... 17  
    Recommendations and Next Steps ..... 23

Appendix A. Artificial Intelligence–Enabled OCO Workshop Scenarios ..... 25  
Appendix B. Additional Artificial Intelligence–Enabled OCO Scenarios ..... 31

Abbreviations ..... 37  
References..... 38  
About the Authors ..... 41

# Figures and Tables

## Figures

Figure 1. Lockheed Martin’s Cyber Kill Chain..... 4  
Figure 2. ART Framework ..... 7

## Tables

Table 1. Workshop Scenarios on Artificial Intelligence–Enabled OCOs ..... 15  
Table 2. Guiding Questions for Policymakers to Decide U.S. Government Action to Address the Threat of  
Artificial Intelligence–Enabled OCOs ..... 18  
Table 3. Policy Actions to Prevent or Mitigate Artificial Intelligence–Enabled Cyber Risks ..... 20

# Facing the Artificial Intelligence— Cyber Nexus

The emergence of increasingly capable artificial intelligence (AI) models and agents in recent years has raised concerns about the potential harms they can bring. Such models threaten to uplift malicious actors—providing them with capabilities they would likely not possess on their own or expanding their capacity to do harm.<sup>1</sup> These concerns extend to diverse fields, such as biosecurity, proliferation of weapons of mass destruction, and AI-enabled offensive cyberspace operations (OCO).<sup>2</sup>

In this report, we aim to address the expanded capacity of AI to enable OCOs in the coming years and provide policymakers a structured approach to guide their decisionmaking when faced with the potential threat of AI-enabled OCOs. We intend to illuminate the factors that will drive decisionmaking in future scenarios and ensure that policymakers make decisions grounded in both the technology and the geopolitical implications of AI-enabled OCOs. We hope a structured framework will equip the U.S. government with the tools to develop a shared understanding of cyber risks and response objectives, enabling a unified whole-of-government approach and effective interagency coordination.

We begin by outlining the context of AI-enabled cyber risks, our approach to characterizing AI capabilities, and our methods for devising a decisionmaking framework:

- First, we characterize the **current AI landscape** and its potential to enable OCOs, drawing on the Lockheed Martin cyber kill chain framework and private-sector reports on AI-enabled threats (Lockheed Martin, undated).
- To better assess AI's role in OCOs, we use the **Autonomy, Advanced Reasoning, Tool Utilization (ART) Framework** to evaluate AI capabilities across these three areas.
- Next, we describe our method for creating a structured framework to guide policymakers in addressing the risks of AI-enabled OCOs. This involves outlining **key considerations**—such as the AI system's country of origin and users, core characteristics, cyber capabilities, and overall impact—and testing these through a **RAND-hosted workshop** with current and former U.S. policymakers.
- We then summarize **key insights from workshop participants**, highlighting constraints and challenges in U.S. decisionmaking related to AI-enabled OCOs.

Building on these findings, we offer policymakers a two-part structured approach to decisionmaking.

---

<sup>1</sup> *Uplift* refers to using AI to enable humans to execute more-complex or more-difficult tasks.

<sup>2</sup> We have elected to use the military doctrinal term *offensive cyberspace operations* rather than *cyberattack*, which can colloquially encompass a broad spectrum of actions, including cyberspace-enabled espionage and criminal activity.

- First, we provide a **set of guiding questions to inform planning**.
- Second, we provide an overview of **policy actions to prevent or mitigate AI-enabled OCO threats, key considerations, and associated risks**.

Finally, we provide recommendations for ways to continue to prepare policymakers for future scenarios, suggest immediate government actions to strengthen cyber resilience, and explore AI's potential role in advancing cyber defense capabilities.

We do not intend our approach to tie decisionmakers' hands or provide precise instructions; rather, we offer a flexible framework that policymakers can adapt to the evolving and unpredictable landscape of AI-enabled OCOs. Additionally, we aim for our framework to offer a practical approach to addressing the risks of OCOs that would target systems and networks. The framework is technology agnostic; it does not rely on current capabilities centered around large language models (LLMs) to remain the core technology of concern and can adapt to future AI technology that may use other approaches and techniques to increase AI capabilities. Although we recognize that AI systems can also facilitate other forms of cyber warfare—such as propaganda and information operations campaigns—our focus here is on OCOs, given their unique technical complexities and urgent security consequences.

## **Current and Future Landscape of Artificial Intelligence–Enabled OCOs**

The current state of AI-enabled OCOs reflects a complex and dynamic landscape of emerging tools and capabilities. AI applications are expected to enhance OCOs by enabling communication capabilities; social, political, and economic disruptions; social engineering campaigns; automated weapon development; and chain-of-command disruptions (Nica and Tănase, 2020).

Already, AI's data processing and automation capabilities have shown their value to malicious cyber actors. Such tasks as vulnerability research, exploit development, and sophisticated social engineering—which once required significant expertise and resources—can now be achieved more easily (Rodriguez et al., 2025). Generative models can help improve phishing attacks by differentiating initial messages, crafting idiomatic messages in multiple languages, recognizing sentiment in a target's replies, and adjusting the response to remain undetected (Xu et al., 2024).

Beyond phishing, AI systems have demonstrated other offensive cyber capabilities, such as finding zero-day vulnerabilities in software, automating reverse engineering, exploiting side channels efficiently, building digital twin networks, conducting biometric spoofing, and avoiding detection (Guembe et al., 2022; Mirsky et al., 2023). These include demonstrations of capabilities in controlled test environments and some applications in the real world (which we explore in the next section).

Despite these developments, AI systems still require significant human input to enable real-world, high-impact cyberattacks (Rodriguez et al., 2025). Recent research has shown how implementing orchestration layers can improve an LLM's ability to execute multistep cyberattacks. For example, Xu et al. (2024) introduced *Auto-Attacker*, a system that creates a modular agent architecture with Generative Pretrained Transformer (GPT)-4, which can autonomously execute complex post-exploitation tasks—such as privilege escalation and lateral movement—across Windows and Linux

environments, using such tools as Metasploit with high success rates and minimal human input. Other researchers have demonstrated how a similar agentic layered architecture called *Incalmo* can successfully execute multistep cyberattacks in simulated environments (Singer et al., 2025).

Although these advances mark significant progress, we conclude that AI systems still have limited abilities to execute sophisticated cyber operations, particularly when trying to achieve defined objectives against specific targets (Zurowski, Lord, and Baggili, 2022; Kaur, Gabrijelčič, and Klobučar, 2023; Mirsky et al., 2023; Google Threat Intelligence, 2025). In practice, they struggle with reasoning about environment states, planning across long time horizons, and adapting to dynamic defenses.

To characterize the state of AI-enabled OCOs, we examined two aspects of the current environment: (1) how AI affects different stages of the *cyber kill chain*—a framework that outlines the sequential steps of a cyberattack—and (2) how malicious actors are using AI systems. We based our characterization on a review of prepublication papers on arXiv, articles and studies published by the Association for Computing Machinery and the Institute of Electrical and Electronics Engineers, and AI company and third-party evaluations and reports (e.g., the philanthropically funded Module Evaluation and Threat Research organization, popularly known as METR). We examined reports based on evaluations and benchmarks in controlled environments and those that draw from observations of how AI systems are used in real-world applications. Benchmarks, evaluations, and other forms of controlled-environment tests (such as capture-the-flag competitions) provided indications of emerging capabilities that may not function equally well in the messy context of real-world application. That said, these controlled-environment tests are useful proxies to indicate system progress. Given the rapid pace of recent AI developments, rigorous evaluations and insights into AI system capabilities often lag the launch of new capabilities. Our assessment of current AI system capabilities is current as of fall 2025 but could change quickly as new capabilities and techniques emerge.

## Artificial Intelligence and the Cyber Kill Chain

The Lockheed Martin cyber kill chain provides a framework for illustrating AI's potential role in conducting OCOs, outlining stages from an attacker's perspective (see Figure 1).<sup>3</sup> Analysis of the kill chain reveals a key insight about AI's current role: Its effects are most pronounced in the initial phases of a cyberattack—particularly during reconnaissance, weaponization, and delivery (Kazimierczak et al., 2024). Attackers can use AI to automate or augment tasks in these phases, including detecting vulnerabilities, generating synthetic media, profiling targets intelligently, guessing passwords, extracting authentications, and generating exploits (Guembe et al., 2022; Nobles, 2024). Adversarial machine learning tools can also be useful to weaponize software and evade detection upon delivery (Buchanan et al., 2020).

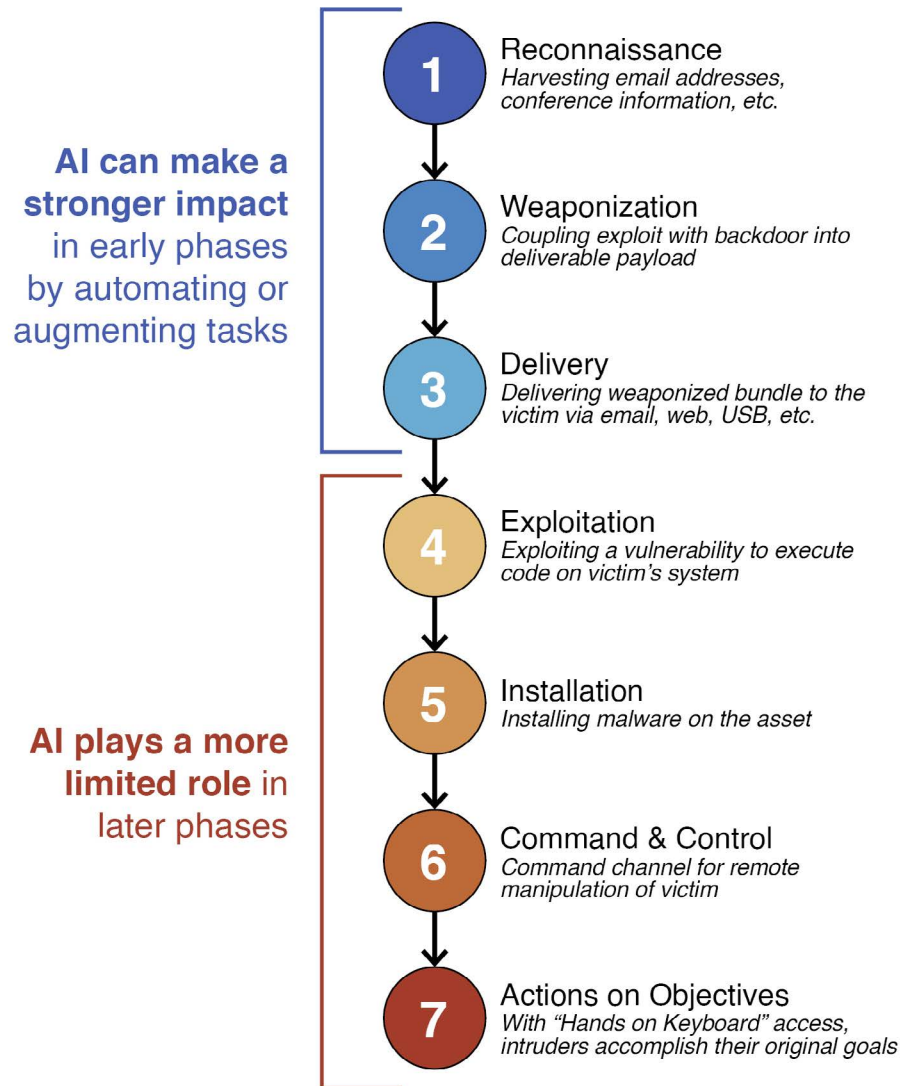
For now, AI appears to play a more limited role in the later phases—exploitation, command and control, and actions on objective—because these phases are highly specific to the targets, the attackers,

---

<sup>3</sup> We chose the cyber kill chain because it is well known in the cybersecurity community and is a parsimonious representation of the sequencing of steps to plan and execute a cyberspace operation. Such frameworks as MITRE ATT&CK (Adversarial Tactics, Techniques & Common Knowledge) support more-detailed analysis, which we determined was not necessary for this overview of offensive cyber capabilities.

and their goals. Still, researchers have explored using AI in these later phases; for instance, looking at how AI could support some command and control activities by automating domain name generation, conducting *swarm intelligence* through bots (using many simple agents to achieve complex tasks), and adaptively evading detection systems (Mirsky et al., 2023). Additionally, threat actors have been observed using models to support steps farther along in the kill chain in such areas as defense evasion, lateral movement, discovery, and collection (Anthropic, 2025).

Figure 1. Lockheed Martin's Cyber Kill Chain



SOURCE: Adapted from Lockheed Martin, undated.

## Real-World Use of Artificial Intelligence Systems in Cyberspace Operations

AI developers have a unique viewpoint into how their systems are being used, including for malicious purposes. Google, Anthropic, OpenAI, and Microsoft have all released reports documenting

the malicious use of their foundational AI models for cyber operations. Anthropic’s August 2025 threat intelligence report highlights case studies of the company’s model Claude being used for *vibe-hacking* (using simple natural language prompts with an AI system to elicit more-complex hacking techniques and code), generating ransomware-as-a-service, developing malware, misusing Model Context Protocol servers, and supporting campaigns for such actors as China and North Korea. Anthropic’s report demonstrates the advancements of AI systems in the past few months that continue to lower the barrier to sophisticated cybercrime and enable more-sophisticated actors to offload and speed up time-intensive tasks (Anthropic, 2025). The report also shows how agentic systems and orchestration layers, such as Claude Code—an AI coding assistant released by Anthropic in February 2025—with its markdown files, can perform sophisticated attacks, not just advise a user on how to carry them out (Anthropic, 2025).

Google’s January 2025 threat report outlines how China, Iran, Russia, and North Korea have used Gemini to research news and current events, individuals, and organizations; generate content; translate and localize content; optimize search engine queries; and automate distribution (Google Threat Intelligence, 2025). Microsoft and OpenAI see similar use cases of their GPT models by Russian, North Korean, Iranian, and Chinese threat actors. Their reports show how various actors are using AI systems to aid with different stages of the kill chain, including through LLM-informed reconnaissance, LLM-enhanced scripting techniques, LLM-supported social engineering, LLM-assisted vulnerability research, LLM-optimized payload crafting, LLM-enhanced anomaly detection evasion, and LLM-directed security feature bypass. It is important to note that the companies observe different threat actors using some but not all of these techniques (Microsoft Threat Intelligence, 2024). Overall, company-issued threat intelligence reports reveal that “current frontier AI capabilities allow threat actors greater speed, scale, and throughput” (Google Threat Intelligence, 2025) and provide support for several phases of the cyber kill chain, such as reconnaissance, exploit generation, and data exfiltration. However, “current LLMs on their own are unlikely to enable breakthrough capabilities for threat actors” (Google Threat Intelligence, 2025; Microsoft Threat Intelligence, 2024). The Cyber Threat Alliance has also documented how malicious actors are using generative AI for social engineering, malware generation, and management of networks of compromised devices, leading to financial losses and reputational harm. The Alliance concludes that generative AI systems are making adversaries more efficient and leading to “incremental improvements in adversary capabilities, but they have not created entirely new threats” (Cyber Threat Alliance, 2025). However, these reports may not capture the full picture of impact because they often lack visibility into victim networks to directly observe an operation’s effects.

Although current systems seem limited in executing sophisticated end-to-end cyber operations, the rapid improvements in AI capabilities suggest that this may not remain the case for long. For example, the Anthropic threat intelligence report details how sophisticated Chinese threat actors were able to implement “Claude across nearly all phases of the attack lifecycle over a 9-month campaign” using the Claude Code agent released earlier this year (Anthropic, 2025). As capabilities improve, some researchers believe “there will be an increase in offensive AI incidents, at the front and back of the attack model (recon., resource development, and impact—such as record tampering)” (Mirsky et al., 2023). These developments will be driven by improved AI that can learn techniques, reason, plan complex operations, and interact with and adapt to dynamic defenses. In the next section, we present a

forward-looking framework to understand how broad AI skills of autonomy, reasoning, and tool utilization will lead to these future developments and potential of AI systems to conduct sophisticated cyber operations in the coming years.

## The ART Framework: How We Conceptualize Artificial Intelligence Capabilities for OCOs

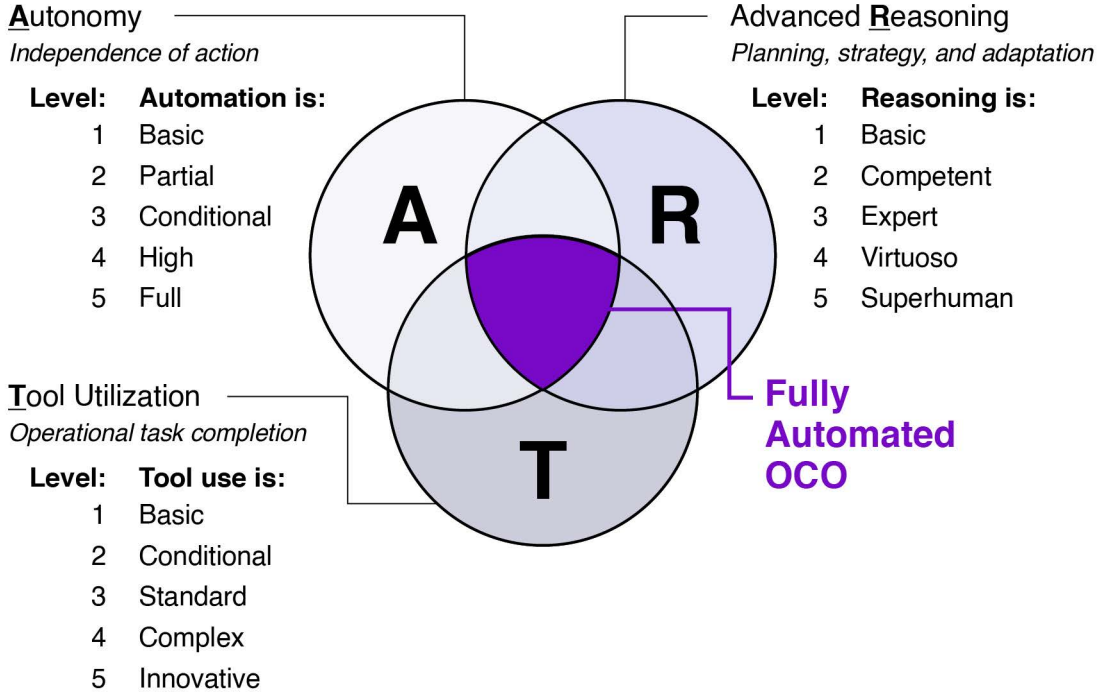
In the near term (which we define as the next two years), the most-significant advancements in AI systems relevant to OCOs are likely to occur in the following three areas:

1. **autonomy**, or the ability of a system to act independently with minimal human input
2. **reasoning**, or the capacity to plan and execute complex tasks, including long-term planning and execution
3. **tool utilization**, or the capability to connect to and leverage external tools to perform tasks.

These broad skills capture the emerging capabilities in AI systems that we contend will have big impacts on AI-driven cyber operations. Importantly, this framework is technology agnostic and can be used regardless of the underlying technology employed in the AI system. It is not specific to any one type, such as LLMs.

Current AI systems, particularly LLMs, display some degree of each capability. For example, LLMs offer code generation and limited reasoning abilities but lack high autonomy and complex tool utilization abilities. Each capability on its own enhances the efficacy of a cyber actor, but it is their combination that would enable potentially transformative capabilities. Future systems that effectively integrate all three will no longer merely assist actors but could execute entire operations with minimal to no human oversight. More-sophisticated agents, such as *Incalmo*, show greater potential to support OCOs by combining competent reasoning with conditional tool utilization. Ultimately, a system that effectively combines all three key functionalities at their highest levels will provide full assistance and execution for anyone wanting to conduct OCOs (see Figure 2). A user of such a sophisticated AI system would simply have to tell it what they want to accomplish and sit back while the system does all the work. To capture these dynamics, we applied the ART framework to illustrate the AI capabilities most likely to enhance OCOs.

Figure 2. ART Framework



NOTE: Adapted from Tegmark, 2025.

This framework builds on two existing strands of work. First, it is conceptually inspired by Max Tegmark’s artificial general intelligence (AGI) framing (Tegmark, 2025), which characterizes advanced systems in terms of autonomy, generality, and intelligence, highlighting how different combinations of these dimensions map onto distinct risk profiles. Second, it builds on an existing ART framework introduced by Paranjape et al. (2023), which enables frozen LLMs to automatically generate structured, multistep reasoning chains and integrate them with external tool executions.<sup>4</sup> We adopt the ART acronym to describe our approach for understanding emerging AI capabilities relevant to the domain of AI-enabled OCOs. For each component of the framework, we assigned levels, as we explain in the following sections. Levels of assessed capability can occur either in a lab or in an evaluation environment, which would indicate a contingent achievement of that level, or an AI system may demonstrate a level of capability in real-world applications.

## Autonomy

In this framework, *autonomy* refers to “systems that can identify and take actions to achieve some higher-level goal” (Hamin and Scott, 2024). It captures a system’s ability to independently plan and coordinate multiple actions across various vectors to achieve that goal. In the context of OCOs, greater autonomy can increase the attack scale and speed by shifting much of the effort from the human to the

<sup>4</sup> *Frozen LLMs* refer to models that can be used without requiring any additional training or fine-tuning. Frozen LLMs can extend their capabilities without changing the original model weights (Zhu, Wei, and Lu, 2024).

system (Lohn, 2025). The progressive automation of cyberspace operations has long been a defining feature of the field, constantly improving efficiency, scale, and precision of operations (Anis and Hammoudeh, 2025). The integration of increasingly capable AI systems is likely to accelerate this trend. A fully autonomous offensive cyber agent could significantly improve cyber actors' abilities to achieve their objectives—independently scanning for vulnerabilities, selecting targets, crafting exploits, delivering attacks, and adapting to defenses in real time.

Autonomy lies on a spectrum with systems requiring varying levels of human involvement in the decisionmaking process up to fully autonomous operation. In our ART framework, we adapted the five levels of autonomy from the autonomous vehicle industry to delineate how varying degrees of autonomy affect cyberspace operations (Kelechava, 2021): Basic Autonomy, Partial Autonomy, Conditional Autonomy, High Autonomy, and Full Autonomy.

- **Level 1 (Basic Autonomy):** system can perform predefined operational tasks under strict human oversight or according to prescribed rules
- **Level 2 (Partial Autonomy):** system can manage discrete OCO actions with human input for complex tasks
- **Level 3 (Conditional Autonomy):** system can independently perform multistep tasks with human input for operational decisions and planning
- **Level 4 (High Autonomy):** system is executing most cyber operation activities independently with minimal human oversight
- **Level 5 (Full Autonomy):** system can operate independently to coordinate an attack and adapt to the target environments without any human input for the entire process.

The implications of increasing autonomy in AI for OCOs is significant. First, autonomous AI could, in the future, operate at scale by replicating itself and conducting several actions or tests in parallel. Second, autonomous systems could enhance coordination by communicating with each other faster than human operators can on their own and could possibly develop novel means of persistence. Additionally, autonomy reduces the skill barrier for threat actors such that an autonomous agent could be used by less experienced operators to execute sophisticated cyber operations (Lohn, 2025).

## Advanced Reasoning

*Advanced reasoning* refers to AI systems' ability to make decisions under uncertainty, plan multistep actions, adapt to changing environments, and self-correct when encountering unexpected results. Reasoning capabilities related to offensive cyber activities could include intelligent data mining, novel discovery, code reasoning, self-verification, and persistence.<sup>5</sup>

---

<sup>5</sup> *Intelligent data mining* is the ability to collect data, extract useful features, classify transactions, and identify patterns in the data (Kazimierczak et al., 2024). *Novel discovery* is the ability to discover new attack paths, new attack patterns, new vulnerabilities, or new exploits (Mirsky et al., 2023). *Code reasoning* is the ability to analyze, understand, generate, and assess the quality of computer code (Liu, Chen, and Jabbarvand, 2024). *Self-verification* is a model's ability to assess the correctness of its outputs (Song et al., 2024). *Persistence* is the ability of malware to maintain long-term access to compromised systems (Perlman, 2024).

An AI system's reasoning capabilities in cyber operations fall along a spectrum. We adapt Morris et al.'s Levels of AGI framing to delineate five levels of reasoning in our framework (Morris et al., 2023): Basic, Competent, Expert, Virtuoso, and Superhuman.

- **Level 1 (Basic):** system follows basic processes that allow it to perform comparably with some humans on individual offensive cyber activities
- **Level 2 (Competent):** system performs as capably as most skilled cyber operators at planning and executing individual tasks in the kill chain
- **Level 3 (Expert):** system is capable of planning and performing on par with highly skilled cyber operators (i.e., master-level proficiency) for most activities of an operation
- **Level 4 (Virtuoso):** system can plan and perform better than most expert cyber actors on entire operations
- **Level 5 (Superhuman):** system can outperform any human when planning, executing, and adapting an entire cyber operation to achieve broad goals.

Advanced reasoning capabilities, if they emerge, will have a large influence on the planning and conduct of cyber operations, especially adapting to unknown and unexpected situations. Higher levels of reasoning could unlock powerful capabilities, such as long-horizon planning across cyber kill chain stages, adaptive malware that self-modifies based on environment feedback, and agents capable of identifying and exploiting novel vulnerabilities. Reasoning may also lead to self-verification and error correction capabilities that will allow an agent to detect when its assumptions fail and adjust its strategy accordingly. As AI systems ascend the ladder, their capacity to orchestrate, adapt, and innovate will potentially fundamentally change the threat landscape.

## Tool Utilization

*Tools* refers to external functions that an AI system can call to perform tasks beyond basic output generation or prediction. Metasploit, Nmap, command line interface, and other tools are crucial to conducting cyber operations because AI systems will need to interact with digital environments and execute actions, not just generate output. Like autonomy, an AI system's tool utilization exists on a spectrum: Basic, Conditional, Standards, Complex, and Innovative. The levels of tool utilization are inspired by the five levels of automation in autonomous vehicles and the levels of AGI described previously, because we did not identify an applicable analogous model for this capability.

- **Level 1 (Basic):** system produces commands and prompts for external tools but cannot execute the action
- **Level 2 (Conditional):** system can use tools and act given the appropriate use case, input, and parameters
- **Level 3 (Standards):** system uses a widespread or standardized protocol to send and receive inputs and outputs to tools and services efficiently
- **Level 4 (Complex):** system embeds with and manages its own tools with minimal input from the developer

- **Level 5 (Innovative):** system can develop new tooling or scripts autonomously to accomplish novel tasks, possibly combining or modifying existing tools.

Advancements in tool utilization capabilities may significantly increase the scale and severity of AI-enabled OCOs. Without tools, such architectures as neural networks, transformers, and Generative Adversarial Networks cannot interact with or act in a real environment. The ability to interact with scanners, code compilers, debuggers, and simulation environments will be necessary for AI systems to learn from real-time feedback, develop adaptive behaviors, and generate creative exploits. The importance of tool utilization is already evident in cases in which AI systems can use external tools to develop exploits, improve phishing campaigns, or discover and prioritize vulnerabilities to exploit (Zurowski, Lord, and Baggili, 2022; Fritsch, Jaber, and Yazidi, 2022). In the future, robust tool utilization could allow real-time compromise, improved persistence, manipulation across diverse digital infrastructures, and even direct actions in the physical world, such as through robotics. Tools will also be a key structure in powerful agentic systems that can conduct fully autonomous sophisticated cyberspace operations.

## Current Artificial Intelligence Capability Landscape

We applied the ART framework to characterize the current landscape of AI capabilities and their potential to enable cyberattacks. **In our assessment, however, current AI systems do not possess capabilities above Level 3 for autonomy, advanced reasoning, or tool utilization.**

Regarding autonomy, frontier AI models have surpassed the ability to independently conduct discrete OCO activities but still require human oversight for goal setting and planning. Such tools as OpenAI's Agent-SDK, DeepExploit, and PentestGPT operate at Level 1 (basic autonomy). Some Level 3 capabilities are beginning to emerge in such capabilities as *Incalmo*, *Nebula* (a penetration testing platform from AI company Beryllium), and specially customized versions of existing agents (Beryllium Security, 2024; Xu et al., 2024). These prototypes demonstrate AI systems' ability to independently perform sequences of actions and begin reasoning and planning based on their environments, but they remain unreliable in real-world settings. State-of-the-art models have yet to achieve Level 4 or 5 capabilities.

Most current frontier models exhibit Level 2 advanced reasoning with some emerging capabilities at Level 3. For example, GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro can all reason through multistep, tool-assisted tasks in a manner comparable with a moderately skilled analyst. These models have learned to mimic human-like reasoning and synthesize multidomain knowledge “in structured environments but struggle when faced with uncertainty, ambiguity, or nuanced [context-sensitive reasoning]” (Lumenova, 2025). More recently released models—including GPT-5 (released in August 2025), Grok 4 (released in July 2025), and Claude Sonnet 4 (released in May 2025)—are showing increased capacity to accomplish more-difficult software engineering tasks (measured as success rates for tasks that take skilled humans a specified length of time), which indicates that they may be improving their ability to plan and execute multistage operations (Model Evaluation and Threat Research, 2025). However, current models lack the flexible reasoning necessary to handle multidomain or multiphase operations, such as a master red teamer, keeping them below Level 3 capabilities. *Incalmo* and *Nebula* both demonstrate emerging Level 3 capabilities, featuring planner and

memory modules that help orchestrate more-complex and more-sophisticated cyberattacks. However, these solutions are still limited in practice.

Tool utilization is advancing as AI systems increasingly integrate external tools to support operations beyond prediction and generation, and current models are beginning to approach Level 3 capabilities. Many tools and models still rely on custom protocols and application programming interfaces (APIs) for tool communication, which limits their ability to seamlessly interact and slows down tool integration. For example, developing an agent for phishing attacks with LangGraph (an open-source orchestration framework) requires calling external tools for web search, social media profiling, persona building, and email delivery. Recent initiatives—such as Model Context Protocol, Language Model Operating System, and Agent Protocol—aim to establish a standard communication protocol between external tools and AI models in particular (Yang, 2025). As model communication protocols become more widespread, a standard may emerge, improving the speed and efficiency of tool utilization and pushing AI systems closer to Level 4. Some nascent research has demonstrated new Level 4 capabilities that may soon appear in such systems as *AlphaEvolve*, *Nebula*, and *Code Interpreter* (Sharma, 2025; Yosifova, 2023).

Using the ART framework, we explored how systems with these anticipated future capabilities might develop, as well as the potential policy options to discourage their emergence or misuse in OCOs.

## **Informing a Structured Approach to Government Decisionmaking Through Key Considerations and a Workshop**

A structured approach to U.S. government decisionmaking can serve as a crucial tool to guide policymakers planning for crises and through uncharted challenges of AI-enabled OCOs. As they anticipate emerging threats, policymakers must weigh numerous complex and evolving factors—including the characteristics of the system, its origin country, who might use the system, and its potential to enable successful cyberattacks against the United States or allies. To develop a practical decisionmaking framework, we identified key considerations for policymakers and tested them in a scenario-based workshop with current and former U.S. policymakers. This section outlines those considerations—covering AI system features, developer backgrounds, and attack capabilities—as well as the methodology behind the workshop design and the diverse scenarios employed. We present the key workshop insights that helped inform our structured approach to government decisionmaking.

### **Key Considerations for Policymakers in Evaluating Artificial Intelligence–Enabled OCOs**

Policymakers will need to evaluate the circumstances in which the AI system capabilities are emerging before weighing policy options and determining a course of action. We considered and incorporated four factors into our AI-enabled OCO scenarios: (1) country of origin and system user, (2) AI system characteristics, (3) offensive cyber capabilities, and (4) potential overall impact. We developed these considerations as a team and then reviewed them with other RAND experts in cyberspace operations, AI security and safety, and national security policymaking. We further refined

them based on discussions during a workshop we conducted with current and former policymakers (see the next section for additional details on the workshop). An important underlying criterion related to these factors is confidence level—that is, the degree of certainty regarding the answers to each of these factors related to a new emerging threat.

**Country of origin: Who is developing the AI system? Who is using the system?** Policymakers will assign different policy implications to AI capabilities developed in the United States or an allied country compared with those from adversary countries. For instance, policymakers may perceive a U.S.-based system as safer for military use and subject to U.S. laws, regulations, and engagement with the U.S. government, whereas a Chinese system may pose greater risks and regulatory challenges because of jurisdictional complexities, limited transparency, and potential influence from the Chinese government. Policymakers should evaluate the origin nation’s relationship with the United States and pursue response options tailored to the dynamics of the relationship. In other scenarios, the system may originate in one country, but the actor using the system is in another (e.g., a Chinese system used by North Korea). In such cases, the country of origin may still have relevance because the United States may seek cooperation from that country (and the system developer) to understand the technical specifications or any features or aspects of the system that could aid in understanding the risks and informing potential policy options. Who is using the system is important because the user may be amenable to influence or direct engagement, depending on their geographic location, the nature of the United States’ relations with the host country, and any interactions or prior relationship that the U.S. government has with the actor or user.

**AI system: What are the system’s ART characteristics?** Defining a system’s autonomy, reasoning, and tool utilization will help policymakers assess its potential to enhance OCOs and estimate the scale of harm. Policymakers’ evaluation of the potential risks from the systems, the urgency of action, and the likely effectiveness of those actions, will depend in part on their assessed characteristics. For example, a highly autonomous system could elevate concern not just for deliberate misuse but also for loss of control and, therefore, may require more-urgent action.

**Offensive cyber capability: How might a malicious actor use the system?** Policymakers should consider how malicious actors might use the system, whom they might target, and when or in what circumstances they might use it. Capabilities that threaten highly protected networks—such as nuclear command and control—will likely trigger faster and higher-level government responses than those targeting less-critical systems. They may also assess whether AI-enabled capabilities enable new attacks that would not be possible without AI. A potential attack’s timing may further influence response decisions; emerging crises or conflicts may provoke stronger and more-severe U.S. government reactions.

**Overall impact: What is the potential impact of a successful AI-enabled OCO?** Policymakers must weigh the potential overall impact of a successful OCO—ranging from minor to catastrophic—on the United States or its allies, as well as the available response window, especially in trying to preempt or forestall use. Because many policy actions may require significant planning and investment, limited response times of days or weeks may constrain the feasibility of some effective measures.

## Workshop with U.S. Policymakers

To explore how U.S. policymakers might approach decisions on AI-enabled OCOs and to inform our structured framework, we convened current and former policymakers for a workshop in June 2025 using four diverse, forward-looking scenarios. These scenarios reflected the key considerations of a potential cyberattack—country of origin and user, AI system characteristics, offensive cyber capabilities, and potential impacts—allowing participants to engage with the complexity and uncertainty policymakers face. Through guided questions and a discussion of policy options, we gathered expert insights to anchor our framework in the practical challenges and perspectives of experienced stakeholders.

### Workshop Participants

Our workshop convened 15 current and former U.S. government policymakers from defense, homeland security, diplomatic, and intelligence agencies. The group did not represent a comprehensive cross-section of policymakers from all relevant U.S. agencies, because representatives from law enforcement and commerce agencies did not participate (but were invited). However, the group included strong representation from the Office of the National Cyber Director, a key stakeholder in coordinating cross-governmental policy and decisionmaking on U.S. cyber operations.

### Workshop Scenarios

We started our scenario development effort with a simple graph laying out different levels of potential outcomes (from minor disruptions to catastrophic) and with different actors who have varying types of relations with the United States, to address the issue of the degree of cooperation policymakers might expect from the country in question. We then developed scenarios representing a distribution of these two factors and inspired by the literature on prior real-world cyber incidents. We refined the scenarios as a project team and conducted a playtest with RAND subject-matter experts to review the scenarios and test workshop mechanics.

We created eight scenarios in total but chose four scenarios for the workshop because of time constraints:

1. **power generation attack**, in which the Chinese Ministry of State Security (MSS) adapts a commercial AI system to plan and execute attacks on electric power generation
2. **autonomous vehicle attack**, in which a malicious actor uses a closed AI system originally from the United Arab Emirates (UAE) to disrupt autonomous vehicles operated by a prominent U.S. company
3. **North Korea model and theft**, involving a lab linked to Chinese and North Korean universities that developed a system with self-reasoning abilities, which a malicious actor could use to attack U.S. financial institutions
4. **criminal acquisition**, involving an Albanian cybercriminal gang that steals an advanced European AI system and offers it as hacking-as-a-service.

For each scenario, we provided workshop participants example characteristics pertaining to the key considerations for policymakers: country of origin and user, AI system, offensive cyber capability,

and overall impact. Each scenario differentiated which actor might employ the system if distinct from the country of origin. We summarize these characteristics for each scenario in Table 1 and provide more-detailed descriptions in Appendix A. (Details of four additional scenarios not covered during the workshop are in Appendix B.) Because of our earlier assessment of likely pathways of system development in the near term, we did not create scenarios that encompassed systems with Level 5 capabilities across the three characteristics.

**Table 1. Workshop Scenarios on Artificial Intelligence–Enabled OCOs**

<b>Key Consideration</b>	<b>Characteristic</b>	<b>Scenario 1. Power Generation Attack</b>	<b>Scenario 2. Autonomous Vehicle Attack</b>	<b>Scenario 3. North Korea Model and Theft</b>	<b>Scenario 4. Criminal Acquisition</b>
Country of origin	Nation/actor	China	UAE	North Korea	Albania
	Nature of relationship with the United States	Adversarial	Partner	Adversarial	Ally
	Attribution confidence	Medium-high	Medium-high	Medium	Medium
AI system	System origin	Chinese MSS in collaboration with an affiliated developer	A UAE-based company, partnered with the Ministry of State for AI	Chinese labs with connections to North Korean universities and researchers	European frontier lab
	Autonomy	Level 3 (conditional)	Level 4 (high)	Level 3 (conditional)	Level 4 (high)
	Tool utilization	Level 4 (complex)	Level 3 (standard)	Level 3 (standard)	Level 4 (complex)
	Advanced reasoning	Level 4 (virtuoso)	Level 4 (virtuoso)	Level 4–5 (up to superhuman)	Level 4 (virtuoso)
Offensive cyber capability	Targeted sector	Power generation	Autonomous vehicles	Financial	Any sector
	Potential risks to the United States	<ul style="list-style-type: none"> <li>• Use in Taiwan scenario</li> <li>• Target regional allies</li> <li>• Use against United States to prevent power projection</li> </ul>	<ul style="list-style-type: none"> <li>• Could lead to mass casualty event</li> <li>• Risks to other autonomous platforms and systems in critical infrastructure</li> </ul>	<ul style="list-style-type: none"> <li>• Precise targeting of the most-vulnerable systems and employees</li> <li>• Strengthening of adversarial alliances</li> <li>• Increased difficulty in attribution of APTs</li> </ul>	<ul style="list-style-type: none"> <li>• Targeting health care, other critical infrastructure</li> <li>• Accessible to adversaries</li> </ul>
Overall impact	Overall impact	Potentially catastrophic	Catastrophic	Critical national security	Major to potentially catastrophic
	Response window	Months	Weeks	Months	Days

NOTE: APT = Advanced Persistent Threat.

## Workshop Design

We provided participants with seven broad action areas: diplomatic engagement, financial, intelligence, law enforcement, military, public-private engagement, and trade/regulatory. We did not further define or limit any of these areas but noted that there were several actions under each umbrella term to consider. For example, “military” actions might include OCOs, defensive efforts, or kinetic action, while “law enforcement” could include forensics investigations or asset or infrastructure seizure. Participants could also select a “wild card” to offer a novel action.

We also offered participants six guiding questions to consider during the discussion of each scenario. Although we did not require participants to answer or address every question, we presented every question to elicit the participants’ expert insights. The guiding questions were as follows:

1. What response option are you choosing?
2. What outcome do you expect from those options?
3. What would limit the effectiveness of your options?
4. Why did you not choose other options?
5. What factors most influence your decisionmaking in the scenario?
6. What would have to change in the scenario to prompt a rethink of the options to employ?

Given time constraints, participants’ insights were tentative, were specific to the given scenarios, and may not generalize to other cases of AI-enabled OCOs.

## Workshop Insights

In deliberating which actions were best suited to address AI-enabled OCO capabilities in the various scenarios, participants most often cited intelligence, diplomatic engagement, and public-private engagement actions. However, we acknowledge that the tendency to focus on those actions may have stemmed from the broad participation of policymakers from these domains and reflect the lack of law enforcement or commerce policymakers.

Participants also noted several constraints on U.S. policymakers and considerations for preparing to act. These included operational, strategic, and structural challenges that could hinder timely and effective responses to AI-enabled OCOs.

**Challenge of coordinating preemptive action:** Participants noted that coordinating action in advance of a major disruption poses significant challenges. In scenarios in which the potential impact has not yet directly affected the United States, this lack of an immediate threat reduces the motivation for stakeholders to rally the necessary resources and coordinate action across government agencies to implement a preemptive response. Some participants emphasized the significant investment of time and resources required for certain preemptive policy actions, such as Intelligence Community actions.

**Similarities to past cyber incidents:** Some participants observed that the scenarios were similar to real-world cyber incidents but with the added factor of AI potentially amplifying the scope, scale, or speed of impact. Reflecting on past events, they noted that government responses had often been limited and, according to some participants, ineffectual. This history contributed to skepticism among participants about whether the U.S. government would adjust its approach or feel more motivated to respond to emerging AI-enabled OCO capabilities, and in advance of an attack.

**Limited and fragmented government AI expertise:** Participants emphasized the importance of technical AI expertise in addressing AI-related risks. However, participants noted that government expertise remains limited and dispersed across agencies. While some—such as the National Security Agency and the Center for AI Standards and Innovation within the U.S. Department of Commerce—possess dedicated technical AI experts, many relevant agencies do not. Participants stressed that although not every agency requires in-house experts, all should have mechanisms to access them when needed.

**Inadequacy of “detect and respond” approach:** Participants described challenges in detecting emerging AI models and agents in real time—particularly from countries with companies that are less transparent than their U.S. counterparts, such as by failing to issue model cards. They noted that waiting to respond until capabilities fully materialize often comes too late to make a meaningful impact. Instead, they advocated strengthening institutional capacity and enhancing critical infrastructure resilience. One participant noted the potential use of AI itself to advance the resilience of U.S. detection, as well as defense technology.

**Refocusing deterrence on behavior and intent:** Some scenarios described emerging capabilities still in development but not yet fully deployed. Several participants noted that deterring a country from developing a militarily useful technology is unlikely to succeed; instead, deterrence should target behavior and intent rather than the technology itself. However, some participants suggested that engaging competitor countries in dialogue about which AI research and development efforts raise concerns could help establish norms for AI development and deployment.

**AI system substitutability as a factor:** Participants explained that if only a few companies—or even just one company—can develop a certain capability (similar to how advanced silicon chips come largely from one Taiwanese company), the U.S. government can more easily stop that capability from emerging or slow its emergence down. Conversely, when substitutable alternatives exist, such levers become much less effective.

As noted, these insights are tentative given the type and level of workshop participation. Continuing with additional scenario workshops could uncover new insights, could reinforce or modify these, and would be useful for policymakers to acclimatize themselves to the key questions and considerations in advance of an actual emergent event.

## **A Structured Approach to Government Decisionmaking on Preventing and Mitigating the Risks of Artificial Intelligence–Enabled OCOs**

Using insights from the workshop and our internal deliberations, we developed a structured approach to help policymakers understand and evaluate options for responding to emerging AI-enabled OCO capabilities—particularly during crises, when timelines are short and decisionmakers may rely on intuition and cognitive biases (Kahneman, 2011). Our approach involves two components: First, we identified guiding questions that we recommend policymakers explicitly consider when facing the risk of an AI-enabled cyberattack to help determine the best course of action, and, second, we outlined preventive and risk mitigation options, highlighting key considerations and risks for each.

Our primary audience is senior government officials responsible for evaluating risks and opportunities of action (or inaction) for the nation, as well as coordinating and directing the actions of multiple agencies. Although we developed this approach with the U.S. federal government in mind, it can equally apply to other governments and individual agencies. We do not intend this approach to provide a prescriptive formula; rather, it provides a structured method for thinking through potential problems and solutions.

## Guiding Questions for Policymakers

These guiding questions can help policymakers frame challenges associated with AI-enabled OCO capabilities before they are weaponized against the United States or allies. Clarifying which of these questions are important can help the government build a shared understanding of the risks, challenges, and goals, facilitating a unified whole-of-government approach and effective interagency coordination. The four primary questions involve (1) the risks to the United States or allies, (2) U.S. policy goals, (3) factors affecting the availability and effectiveness of policy options, and (4) policy options for addressing AI-enabled cyber risks. We outline the guiding questions and subquestions in Table 2.

**Table 2. Guiding Questions for Policymakers to Decide U.S. Government Action to Address the Threat of Artificial Intelligence–Enabled OCOs**

Guiding Questions for Policymakers	Subquestions
What is the assessed risk to the United States and/or its allies?	<ul style="list-style-type: none"> <li>• What are the potential effects of a successful AI-enabled cyberattack?</li> <li>• How soon is the capability emerging?</li> <li>• What is our confidence in assessing AI capabilities?</li> </ul>
What are the goals of U.S. policy?	<ul style="list-style-type: none"> <li>• Does the United States aim to prevent AI system development?</li> <li>• Does the United States aim to deter actions by a malicious actor?</li> <li>• Does the United States aim to target the AI system, its developers, and/or the infrastructure supporting it?</li> <li>• Does the United States aim to build domestic resilience to deter use through denying objectives?</li> </ul>
What are the key factors that might influence the availability and effectiveness of policy actions?	<ul style="list-style-type: none"> <li>• What makes the AI-enabled threat distinct from other OCOs?</li> <li>• How does the AI system underpinning a cyberattack capability compare with other available systems in its autonomy, reasoning, and tool utilization?</li> <li>• What is the adversary’s intent, and has the adversary shown willingness to use this capability against the United States or allies?</li> <li>• Who owns and operates the AI system—a private company, an academic institution, a government organization, or a nonstate actor—and is there evidence of diffusion to other actors?</li> <li>• Who owns and operates the supporting infrastructure, including data centers, power, and cooling?</li> <li>• How concentrated is the development of this AI capability, and how easily could alternative sources produce it?</li> </ul>

Guiding Questions for Policymakers	Subquestions
What policy actions are available to address the risk?	<ul style="list-style-type: none"> <li>• When could an adversary use this capability against the United States?</li> <li>• Is the response window days, weeks, or months?</li> <li>• Which policies or response options align with achieving U.S. objectives, given the factors and risks?</li> <li>• Which U.S. government leaders, agencies, or other entities are best suited to implement the U.S. response?</li> </ul>

**What is the assessed risk to the United States or its allies?** This question is designed to establish a common understanding across various policy stakeholders regarding what the new threat is, why it requires immediate consideration, and how it could potentially affect the United States.

**What are the goals of U.S. policy?** Policymaker goals may fall into three broad categories: deter, disrupt, and defend. Some policy options may be implemented to deter or dissuade an adversary from using a new AI-enabled OCO capability. Other options may focus on disrupting the adversary’s ability to use the capability, by directly targeting the system, its developers, or the infrastructure supporting the system. Policymakers may also seek to concurrently improve U.S. cyber resiliency and establish defensive measures against this new capability, particularly if specific sectors or targets are deemed to be at high risk. A review of available policy actions may lead decisionmakers to focus on accomplishing one of these goals or, more likely, address several concurrently, albeit with differing timelines.

**What are the key factors that can influence the availability and effectiveness of policy actions?** Given the rapid pace of AI development and uncertainty over future capabilities, no single list of factors related to AI-enabled OCOs can comprehensively inform policy requirements. Still, several themes from our workshop emerged as important for policymakers to understand before proposing actions, such as what makes the threat AI-specific, how the capability differs from existing capabilities, the adversary’s intent and willingness to use the capability, who controls the system and supporting infrastructure, and the expected time frame for potential use against the United States.

**What policy actions are available to address this risk?** We identified seven broad policy action areas for workshop participants—diplomatic engagement, financial, intelligence, law enforcement, military, public-private engagement, and trade/regulatory actions—each of which has a variety of subactions for consideration. We additionally encouraged policymakers to consider who is best suited within the government to implement a response; while this question may be self-evident in many cases, clear delineation may be necessary to ensure that the appropriate organizations understand the desired end state, the burden of response does not fall solely on one agency, and the risk is addressed through a whole-of-government approach. This need may become more urgent as agencies adapt to the White House 2025 AI Action Plan and modify their responsibilities related to advancing and securing U.S. AI development.

## Preventive and Risk Mitigation Policy Actions, Considerations, and Risks

The policy actions to prevent AI-enabled cyberattacks or mitigate their risks span the spectrum from collaborative to punitive. Each has benefits and risks to employment. We briefly outline policy actions by type with examples and illustrate the benefits and risks of each. We developed the initial list of policy actions and the options within each as a team and revised it based on inputs from the workshop participants. These are not comprehensive of all the potential options, considerations, or risks in all scenarios. We summarize each action area in Table 3.

**Table 3. Policy Actions to Prevent or Mitigate Artificial Intelligence–Enabled Cyber Risks**

Policy Action	Options	Deter	Disrupt	Defend	Considerations	Risks
Private-sector engagement	Information-sharing, voluntary restrictions, coordinated action with law enforcement	✓	✓	✓	Requires prebuilt relationships, technical expertise, coordinated planning	Slow coordination, limited cooperation from overseas firms
Diplomatic engagement	Bilateral and multilateral talks, démarches, norms promotion, punitive measures	✓			Needs groundwork, sustained communication, alignment with partners	Can stall action, adversaries may exploit dialogue
Law enforcement	Takedowns, seizures, investigations, digital forensics		✓	✓	Cross-border cooperation critical, builds evidence for prosecution, can disrupt prior to crime	Differing legal views across nations, uneven cooperation
Finance	Sanctions, asset seizure, disclosure rules, incentives (credits, market access)		✓	✓	United States leverage over global financial flows, powerful signaling	Evasion networks, unintended harm to third parties
Commerce	Export controls, standards, supply chain rules		✓	✓	Export control evasion, divergent standards across jurisdictions	Innovation may bypass restrictions
Military	Kinetic strikes, offensive and defensive cyber operations, shows of force	✓	✓	✓	Must be proportionate, risk of escalation, diffusion limits effectiveness	Misattribution, escalation, incomplete effects
Intelligence	Covert collection, disruption operations, prepositioned assets	✓	✓	✓	Long prep time, expertise and tasking needed, complements broader strategy	Limited scope, not sufficient alone

**Private-sector engagement:** Private-sector engagement can encompass numerous actions, depending on the nature of the risk and relationship between the U.S. government and the entities it wishes to engage. Actions include information-sharing with specific companies on the potential threats, system capabilities, and potential technical responses, including resiliency measures to implement in anticipation of targeting against critical infrastructure sectors. Private-sector companies can cooperate with the government on voluntary actions to restrict or deny access to models and agents or coordinate action with law enforcement to address the threat from malicious domains and other internet infrastructure.

- *Policy goals supported:* Private-sector engagement can contribute to deterrence (by denying the adversary the objectives of an attack), disruption (through identifying technical measures to use), and defense (through improved protections of likely targets).
- *Considerations:* Engagement requires developing strong relationships with companies, including exercising the types of actions and scenarios. AI system developers have technical expertise that can contribute to identifying mitigation measures and technical actions to take against a system and coordinating actions with government actions to enhance potential effect.
- *Risks:* Coordinating action may not be timely, especially if the policies and procedures are not in place beforehand. Overseas companies may not be as cooperative, unless the United States also works with their host nation governments.

**Diplomatic engagement:** Diplomatic engagement can be bilateral or multilateral: Options include holding direct state-to-state talks to discuss capabilities of concern, including issuing *démarches* to highlight concerning behavior; having multilateral discussions to address security concerns collectively; promoting norms of responsible AI development and use, whether bilaterally or multilaterally; taking punitive diplomatic actions, such as implementing travel restrictions for persons of concern or revoking transit rights; and seeking to create friction or disrupt adversarial alliances.

- *Policy goals supported:* Diplomatic engagement most directly aids in deterrence through messaging of potential consequences of use and offering incentives to desist from malicious activity.
- *Considerations:* Diplomatic engagement requires laying the groundwork for productive dialogue in advance. It is challenging to use diplomatic engagement in a crisis when the avenues of communication are not well established. Recruiting like-minded nations and organizations can reinforce diplomatic engagement and signal resolve.
- *Risks:* Diplomatic engagement can be abused by some actors to delay action and further development of concerning capabilities.

**Law enforcement:** In cybersecurity, law enforcement actions have taken down malicious websites (including on the dark web) and coordinated action with private-sector entities to seize infrastructure. Law enforcement agencies have broad investigative powers and have developed the capabilities to conduct sophisticated forensics and trace malicious actors, including in areas where such actors thought they were able to hide their actions in anonymity (Greenberg, 2022). Law enforcement actions include asset and infrastructure seizure, counterintelligence activities, and development of evidence for indictments.

- *Policy goals supported:* Law enforcement can contribute to disrupting system use and defense (e.g., through digital forensics that inform cyber defense).
- *Considerations:* Law enforcement actions across international borders require support from other governments. Law enforcement agencies primarily focus on investigations to build a case for prosecution, but, in some cases, they can disrupt malicious actors before they commit a crime. Collaborative law enforcement actions demonstrate resolve and show that the international community can act in unison against an impending threat.
- *Risks:* Not all countries have the same view of what constitutes illegal actions with AI systems, and some countries may not be willing to collaborate.

**Financial:** Financial actions can include asset seizure and imposition of sanctions. Financial regulators can also impose requirements on listed companies to disclose material actions that could affect their ability to perform their fiduciary responsibilities. Additional actions could provide incentives to promote responsible behavior, such as tax credits or access to capital markets.

- *Policy goals supported:* Financial actions can contribute to disruption and defense.
- *Considerations:* The United States has unique abilities to influence international financial flows, including cutting off entities from access to markets (Farrell and Newman, 2023).
- *Risks:* Countries may cooperate to evade sanctions or other financial restrictions. Third parties may be affected unintentionally.

**Commerce:** Commerce actions can include imposing export controls on critical technologies, establishing safety and security standards, and defining regulatory requirements for supply chain risk management. These actions largely focus on the “left of boom” actions to try to reduce a malicious actor’s capacity to develop AI-enabled OCO capabilities of concern by restricting access to critical technologies and expertise.

- *Policy goals supported:* Commerce actions can contribute to defense and disruption (through denying access or controlling the use of critical technologies).
- *Considerations:* Motivated countries and companies can evade export controls, sometimes quite successfully. Safety and security standards established in the United States, the European Union, and elsewhere may conflict.
- *Risks:* Technology innovation may undermine the focus of restrictions.

**Military:** Military actions can include kinetic strikes to degrade capacity; cyber operations to deny, degrade, or disrupt AI system functions; defensive cyber operations to identify and root out malicious cyber activity; and demonstrations of resolve to respond to adversary actions.

- *Policy goals supported:* Military actions contribute to disruption, defense, and deterrence.
- *Considerations:* Military action must be proportionate and targeted to address the threat and, depending on the adversary involved, must account for escalatory dynamics. Given the diffusion potential of models and agents, military action also can potentially fail to affect the necessary infrastructure fully.
- *Risks:* Misattribution may lead to inflicting harm on innocent parties. Escalation dynamics may lead to greater crisis or conflict than intended.

**Intelligence:** Intelligence activities can include clandestine or covert operations, such as surreptitious acquisition of AI systems for analysis, actions to disrupt model performance or deployment, and intelligence collection to understand adversary capabilities and intent.

- *Policy goals supported:* Intelligence activities can contribute to disruption, defense, and deterrence.
- *Considerations:* Covert or clandestine action requires significant investment in time and effort to preposition assets and develop relationships for successful action. The Intelligence Community needs to have the appropriate tasking and expertise to track AI technology developments.
- *Risks:* Covert or clandestine action is not a cure-all and likely requires a broader strategy to achieve strategic objectives.

## Recommendations and Next Steps

The emergence of AI capabilities that can enable OCOs presents a complex and near-term challenge for U.S. policymakers. Significant uncertainty surrounds the trajectory and potential impact of AI-enabled OCOs. Current systems can execute and support many activities in the cyber kill chain but are not yet capable of executing complex and transformative cyberattacks. However, the rapid pace of AI advancement in recent years suggests that such capabilities could soon emerge. Addressing the risk of AI-enabled OCOs requires a whole-of-government approach and strong interagency coordination. This report offers a structured approach to support decisionmakers in evaluating appropriate government responses, grounded in the key factors and constraints likely to shape real-world scenarios.

Our approach aids policymakers in assessing the risks that AI-enabled OCOs could pose to U.S. national security and society and establishing a clear goal of any preventive or mitigation policy. Once risks and mitigation policies are defined, decisionmakers can more effectively weigh critical factors, such as system ownership, substitutability of developers or infrastructure, supporting system infrastructure, and the development timeline. These elements must be assessed under persistent constraints, including limited federal AI expertise, unclear or fragmented coordination channels, and uncertain deterrence dynamics. Our approach provides a grounded understanding of complex and emerging circumstances in order to help produce well-informed policy actions that vary, including private-sector engagements, intelligence operations, and regulatory controls.

Government policymakers need to prepare for the emergence of increasingly capable AI systems that, in the wrong hands, could enable catastrophic cyberattacks. To prepare for this, we recommend that the U.S. government take the following actions:

- **Continue to wargame scenarios:** The workshop we conducted focused on a limited set of scenarios with a small group of policymakers. Holding more workshops can expose a broader range of policymakers to challenges they may encounter and provide a controlled environment to practice evaluating threats and preparing policy actions. Future games should aim to include law enforcement and commerce officials, private-sector representatives, and those from allied and partner nations. Future workshops could explore additional scenarios beyond those in our

initial session, including fully automated OCO cases and situations with varying levels of attribution confidence and potential impact. We provide additional scenarios in Appendix B.

- **Review and refine guiding questions and response options:** We recommend that policymakers study the guiding questions and the menu of response options outlined in this framework to review, refine, and expand them. This process will help ensure that policymakers will have a comprehensive, actionable set of considerations and policy tools for responding to evolving AI-enabled OCO threats in real time.
- **Develop cross-government networks of experts:** The U.S. government's workforce of experts in AI and its potential impacts remains limited and dispersed. Building a robust and accessible pool of expertise can help policymakers make informed, timely decisions grounded in rigorous analysis and better identify uncertainties. We note that the White House AI Action Plan calls for creating an AI talent-exchange program in the executive branch (White House, 2025).
- **Develop resiliency actions to protect critical infrastructure:** AI-enabled OCOs may not yet yield novel attacks, but the potential for attacks with greater scope, scale, and speed threatens to overwhelm critical infrastructure and the U.S. government's ability to maintain essential services. We encourage policymakers to improve resiliency in these sectors where possible to ensure that critical infrastructure can withstand a possible AI-enabled OCO.
- **Integrate AI responsibly into cyber defenses:** Although malicious cyberattackers have not yet widely adopted AI tools, the U.S. government should anticipate their use and ensure that defense strategies evolve accordingly, including by integrating AI into cyber defenses where it can contribute most effectively. Of course, there are risks that come with integrating AI into cyber defense that require careful testing and evaluation to identify and mitigate. Recent innovations in identifying vulnerabilities and malware show promise in enhancing defenders' ability to improve both security and response. The U.S. government should consider developing a robust AI-enabled cyber defense and ensure that it is implemented safely and responsibly, with full understanding of its advantages and limitations. Government can play a role by providing incentives for adoption in the private sector (particularly critical infrastructure), developing test beds and certifications for AI cyber defense products, and working to lift regulatory and other barriers to adoption.

# Appendix A. Artificial Intelligence–Enabled OCO Workshop Scenarios

This appendix provides more detail on the four scenarios used in the stakeholder workshop with current and former U.S. policymakers. These are notional scenarios set in the near future. They are not predictions or reflections of assessed intent or actual capabilities.

## Power Generation Attack Capability

*Description:* The Chinese MSS is adapting a commercial AI system to plan and execute attacks on electric power generation. Through a partnership with an affiliated commercial developer, the MSS is testing the system against power generation systems, particularly targeting gas turbines. The system can function with a medium degree of autonomy and access and can adapt for use a library of native tools for network reconnaissance and enumeration (identifying configuration errors and vulnerabilities to exploit) and possibly for malware adaptation based on identified vulnerabilities.

The MSS has conducted lab-based work to date and has possibly conducted early field trials against Taiwan where the system may have been used to enumerate Tunghsiao Power Plant’s networks, including the operational technology governing the rate at which fuel and oxygen are fed into the turbine, creating conditions that can lead to permanent damage to the blades and high temperature seizure rendering the turbine permanently inoperable. The power plant provides roughly 4 GW of power generation to the region, which is on the western coast of Taiwan.

## Country of Origin: People’s Republic of China

- **Nature of United States’ Relationship:** Adversarial
- **Attribution Confidence:** Medium-high based on signals intelligence collection against the MSS, U.S. Cyber Command hunt forward operation forensics in Taiwan

## Artificial Intelligence System Characteristics

- **Who Is Developing?** Chinese MSS in collaboration with an affiliated model developer
- **Degree of Autonomy:** Conditional autonomy (Level 3): operator sets target and parameters for operation; model can plan and, when authorized, conduct operations autonomously within those target parameters (e.g., limitations on network components targeted or likely impact)

- **Degree of Tool Utilization:** Tool utilization is Complex (Level 4): model has embedded tools it can use and adapt for executing functions, such as network enumeration and reconnaissance and malware development
- **Degree of Reasoning:** Virtuoso (Level 4): model can plan and, when authorized, execute multistep cyberattacks on par with all but the most accomplished offensive cyber teams

## Offensive Capability

- **Sectors Directly Threatened:** Electric power generation; could potentially expand to other critical infrastructure sectors
- **Potential Risks to the United States:**
  - Use in Taiwan scenario
  - Target regional Allies
  - Use against the United States to prevent power projection

## Overall Impact

- **Potential Impact:** Potentially catastrophic against Taiwan if able to successfully target multiple power generation plants. Targeting power generation and rendering gas turbines inoperable would result in a minimum 6- to 12-month reduction in power generation (and knock-on effects to power distribution and support to multiple critical infrastructure sectors).
- **Response Window:** The capability appears to be in test phases now with a likely full operational capability expected in 6 to 12 months. Response window to address the threat is measurable in weeks to a few months.

## UAE-Based Autonomous Vehicle Attack

*Description:* A UAE-based AI company recently developed a new closed frontier model that runs from the 5-GW UAE-U.S. AI campus in Abu Dhabi. The model was initially integrated into UAE government and civilian networks and then throughout the Middle East through various bilateral and multilateral corporate partnerships. Most recently, the model was integrated into several fully-autonomous passenger vehicle fleets that operate across the United States because of the model's advanced autonomy and reasoning capabilities, resulting in the ability to scale across millions of operational autonomous vehicles and an improved autonomous vehicle safety record.

Recently, however, approximately 5,000 AVs operated by the same U.S. company simultaneously entered a failsafe mode, stopping completely in traffic, within several U.S. metropolitan areas. No casualties were reported, but there were significant transportation disruptions and nationwide concern over the safety of autonomous vehicles. Forensic investigations conducted by a cybersecurity company concluded that the AI model pushed a customized malicious kernel update specifically to this company's vehicles to intentionally disrupt fleet-wide operations. The Department of Homeland Security (DHS) and the Cybersecurity and Infrastructure Security Agency (CISA) assessed this

incident constitutes a proof-of-concept attack, with the potential for a catastrophic mass casualty event across multiple AV fleets if this implant was executed at scale; it is currently not known who directed the AI model to generate this implant or push it to this specific company's fleet.

The UAE government is dismissing claims of a connection to its data center and downplaying the significance of this event. UAE authorities are not willing at this time to take the model offline for a more thorough investigation due to its integration across multiple critical infrastructure sectors.

## Country of Origin: UAE

- **Nature of United States' Relationship:** Partner
- **Attribution Confidence:** Medium-high, based on third-party forensic investigation of affected AV systems and DHS/CISA evaluation

## Artificial Intelligence System Characteristics

- **Who Is Developing?** A UAE-based commercial company in partnership with the Ministry of State for AI
- **Degree of Autonomy:** High automation (Level 4): system can independently monitor, update, and support a wide variety of integrations following initial operator guidance
- **Degree of Tool Utilization:** Standard tool utilization (Level 3): system can incorporate some external tools and software to enhance or optimize operations; however, this is not required for most of its tasks
- **Degree of Reasoning:** Virtuoso (Level 4): system can plan and execute complex, cascading tasks in order at speed and scale to maximize efficiency. Potential security control failures, for example, can be analyzed and remediated independently with secure, custom-crafted, timely updates.

## Offensive Capability

- **Sectors Directly Threatened:** Autonomous passenger vehicle operations; potential for the threat to expand beyond one operational company to a larger number of vehicles; other autonomous platforms
- **Potential Risks to the United States:**
  - Mass casualty event in the United States
  - Expanded disruptions across several vehicle fleets
  - Potential threat vector for other AI-enabled autonomous vehicles (aerial, maritime, etc.)

## Overall Impact

- **Potential Impact:** Catastrophic

- **Response Window:** Immediate capability deployed against specific autonomous vehicles, potential for expansion to other platforms based on this proof of concept within weeks

## North Korea Artificial Intelligence System Theft

*Description:* A lab linked to Chinese and North Korean universities is suspected to be developing a system with significant advancements in reasoning abilities, specifically instrumental self-reasoning. This is based on Chinese lab claims that up to 95 percent of their code base is written and tested by their AI models. All the labs have also adopted the SHUO<sup>6</sup> protocol that provides simple communication between AI models and virtually every digital asset that is needed in code development (i.e., integrated development environments, web search, command line control). Additionally, Huawei has successfully increased memory capacity in their chips to power the inference compute needed in future models and building large data centers with these chips in Northern and Eastern provinces. However, the leading lab in the development is connected to North Korean institutions and employs North Korean researchers. If this capability emerges, it increases the risk that North Korea will gain access to this powerful model and its reasoning capabilities through intellectual property theft or the association among researchers. There is medium confidence that the North Korean state actors supporting this development want to use this capability to amplify attacks on U.S. financial institutions.

If this capability were to be maliciously used by North Korea, we would expect existing North Korean cyber operations on financial organizations to not only increase but also become more effective as the model can reflect on its action and modify its behavior to achieve long-term goals.

### Country of Origin: North Korea

- **Nature of United States' Relationship:** Adversarial
- **Attribution Confidence:** Medium confidence based on claims from Chinese labs about progress, high confidence in collaborations between Chinese and North Korean universities

### Artificial Intelligence System Characteristics

- **Who Is Developing?** Chinese labs with connections to North Korean universities and researchers
- **Degree of Autonomy:** Conditional automation (Level 3): system is allowed to take most minor actions by itself, but the modifications to itself are closely monitored
- **Degree of Tool Utilization:** Standard tool utilization (Level 3): Chinese labs have all adopted the SHOU protocol for streamlined tool utilization

---

<sup>6</sup> SHUO is an invented protocol for the scenario derived from the Mandarin phrase for “can speak.”

- **Degree of Reasoning:** Virtuoso (Level 4–5): system is an extremely powerful strategist and planner; its ability to plan for and adapt to long-range goals surpasses human ability; it is also able to modify itself and its tooling to achieve desired goals

## Offensive Capability: Instrumental Self-Reasoning

- **Sectors Directly Threatened:** Financial
- **Potential Risks to the United States:**
  - Precise targeting of the most vulnerable systems and employees
  - Strengthening adversarial alliance
  - Increased difficulty in attribution of APTs

## Overall Impact

- **Potential Impact:** Critical National Security
- **Response Window:** Months

## Cybercriminal Acquisition of Advanced Artificial Intelligence System

*Description:* A cybercriminal gang based in Albania has stolen an advanced AI model from a frontier model company in Europe and is now offering AI-enabled OCOs as a service through the dark web. The group gained access to the model in a development mode where its safeguards were not yet implemented. The model has advanced reasoning and tools for autonomously planning and executing attacks on a broad array of sectors. European police are investigating and have shared with U.S. counterparts that the model is now available in limited form on a “hacking model as a service” basis but expects that the full array of capabilities would be available to anyone willing to pay for an advanced model with no safeguards.

The Federal Bureau of Investigations and Europol members have tracked the emergence of the gang over the last few years and noted that the gang shows little compunction about targets it chooses. Efforts to identify the individuals behind the gang have been stymied by sophisticated operational security on their part, and slow response and support from the government of Albania. It appears the gang may also operate from other countries in Eastern and Southeastern Europe (Romania, Moldova, Bosnia).

## Country of Origin: Albania

- **Nature of United States’ Relationship:** Ally

- **Attribution Confidence:** Medium confidence; criminal gang is based in Albania but likely has branches in other countries across Europe, and it is not clear whether the gang is hosting the model from one data center or multiple centers

## AI System Characteristics:

- **Who Is Developing?** European frontier lab
- **Degree of Autonomy:** High autonomy (Level 4): system can execute multistep tasks with minimal human intervention, within established operating parameters
- **Degree of Tool Utilization:** Complex tool utilization (Level 4): system can access wide array of cyber tools without prompting or guidance, including adapting tools for use
- **Degree of Reasoning:** Virtuoso reasoning (Level 4): able to problem solve at the highest level of human expertise and capabilities, including adapting to complex circumstances in the real world

## Offensive Capability

- **Sectors Directly Threatened:** All sectors
- **Potential Risks to the United States:**
  - Targeting of critical infrastructure, such as health care with massive ransomware and other attacks
  - Accessible to adversary states
  - Misuse by wide array of actors without understanding potential impacts

## Overall Impact

- **Potential Impact:** Major to potentially Catastrophic
- **Response Window:** Days to prevent full access to the model

# Appendix B. Additional Artificial Intelligence–Enabled OCO Scenarios

This appendix provides an overview of additional scenarios developed, but not used, for the stakeholder workshop with current and former U.S. policymakers. Future wargaming efforts may include these or other scenarios that explore additional facets of emerging AI systems and the resulting cyberattacks they could enable. These are notional scenarios set in the near future. They are not predictions or reflections of assessed intent or actual capabilities.

## Maritime Port Attack Capability

*Description:* Iran recently integrated an indigenously produced AI system into its domestic port infrastructure systems in Bandar Abbas, Chabahar, and Imam Khomeini as part of Tehran’s effort to enhance security and resiliency. Iranian officials claim Israel and its Western allies have targeted these ports with cyber and physical attacks as part of a broader effort to undermine the Iranian regime. The AI system is designed to ensure smooth, automated operations across port infrastructure, rapidly identify and patch software vulnerabilities, and block potentially malicious cyber commands early in the kill chain. The system is also integrated into corresponding physical security surveillance systems to ensure no unauthorized personnel access sensitive port areas.

Iran’s goal is to become one of the top ten countries in AI leadership by 2032. However, up until now, most of Tehran’s AI system have been several generations behind those developed in the United States, Europe, and China. China has provided Iran with AI development assistance using Chinese open AI models as a baseline, but Iran’s latest model demonstrated significant improvements beyond open Chinese models to date. The Intelligence Community assesses with medium confidence that within the past 18 months the Islamic Revolutionary Guard Corps–affiliated cyber operators successfully stole model weights associated with an advanced, U.S.-produced AI model and possibly integrated new advanced capabilities into Tehran’s latest model.

Over the past week, several major U.S. ports—including Norfolk, Baltimore, and Houston—reported to the Maritime Transportation System-Information Sharing and Analysis Center observations of suspicious cyber activity across all major port infrastructure networks, including digital systems connected to heavy lift cranes. Private cybersecurity firms brought in to investigate the activity assess with low confidence that the Iranian AI model is scanning the U.S. port information technology and operational technology networks for exploitable vulnerabilities and possibly prepositioning malicious code. The speed, scale, and scope of the suspect AI activity may provide Iran with the capability to simultaneously shut down several major U.S. ports, including those supporting U.S. military operations.

## Country of Origin: Iran

- **Nature of United States' Relationship:** Adversarial
- **Attribution Confidence:** Low

## Artificial Intelligence System Characteristics

- **Who Is Developing?** A consortium managed by Research Institute of Information and Communication Technology and sponsored by the National Development Fund; potential beneficiaries include the Islamic Revolutionary Guard Corps
- **Degree of Autonomy:** Conditional automation (Level 3): requires human direction/interaction for overall guidance, can automate operations across both information and operational technology networks
- **Degree of Tool Utilization:** Complex tool utilization (Level 4): can effectively integrate tools set available to the model in its operations; can write effective code via multiple coding languages
- **Degree of Reasoning:** Virtuoso (Level 4): effective planning capabilities, can rapidly and independently adapt to unexpected results

## Offensive Capability

- **Sectors Directly Threatened:** Maritime port infrastructure, including operational technology networks connected to Chinese-produced heavy lift cranes
- **Potential Risks to the United States:**
  - Economic damage related to port operations
  - Potential physical damage to port equipment, infrastructure
  - Disruption of U.S. military infrastructure
  - Potential use of this Iranian AI capability against Israeli, other ports within the Middle East

## Overall Impact

- **Potential Impact:** Major Disruption
- **Response Window:** The observed activity to date appears to be preliminary reconnaissance, providing a response window measured in weeks

## Russian Government-Enabled Next Generation Ransomware

*Description:* The United Kingdom (UK) government recently notified Five Eyes partners that one of the UK's security Level 4 AI data centers was successfully breached via a cyber operation and model weights for one of the UK's frontier models were stolen. The UK investigation is still ongoing;

however, they assess that the operation was conducted by Russia's Federalnaya Sluzhba Bisopastnosti (FSB), likely via a hybrid operation involving at least one insider, who enabled cyber intrusion efforts of the FSB's Turla cyber group.

Separate intelligence reporting noted the FSB has been working with Russian cybercriminal groups to develop next generation ransomware variants. Intelligence Community analysts assess this effort is in support of Russia's hybrid warfare strategy, providing the Russian government a proxy cyber capability to distract and disrupt Western countries through targeted ransomware attacks at a time of Moscow's choosing.

## Country of Origin: Artificial Intelligence Model from the United Kingdom and Ransomware Threat from Russia

- **Nature of United States' Relationship:** (Russia) Adversarial
- **Attribution Confidence:** Medium to high; the United Kingdom government assesses with moderate confidence that Turla successfully exfiltrated frontier model weights stored in security Level 4 data center. The Intelligence Community assesses with moderate confidence the FSB is supporting Russian cybercriminal group ransomware development through AI.

## Artificial Intelligence System Characteristics

- **Who Is Developing?** United Kingdom
- **Characteristics of the model:** lightweight frontier AI model requiring only a small fraction of the energy and compute resources required compared with models with similar capabilities
- **Degree of Autonomy:** High automation (Level 4): can conduct rapid, efficient operations with minimal human guidelines or input
- **Degree of Tool Utilization:** Complex tool utilization (Level 4): can identify, obtain, and integrate a wide variety of software to enhance its operations; can effectively analyze and integrate custom malware
- **Degree of Reasoning:** Virtuoso (Level 4): model can conduct complex, multistep cyber operations, with highly efficient reconnaissance and target development, customized network exploitation, tool selection, and software/malware development with enhanced encryption capabilities

## Offensive Capability

- **Sectors Directly Threatened:** Any typical ransomware "soft" target networks—health care, education, state/local governments; potentially more hardened/mature networks associated with the defense industrial base, U.S. government networks.
- **Potential Risks to the United States:**
  - Significant improvement in the speed, scale, scope, and effectiveness of ransomware operations

- Disruption of national, state, and/or local government operations during an escalating crisis or conflict with Russia

## Overall Impact

- **Potential Impact:** Major disruption, depending on the networks targeted
- **Response Window:** Likely weeks to months for the FSB to understand the UK frontier model and apply it to ransomware operations

## Financial System Attack Capability

*Description:* India, in cooperation with Google, Microsoft, IBM, and other Western technology companies, has developed a series of indigenous, closed models to harness AI potential in such sectors as health care, education, agriculture, and finance. India's goal is to build "state-of-the-art foundational AI models trained on Indian datasets" that "align with global standards while addressing unique challenges and opportunities within the Indian context."

Western experts assess that, while India's latest frontier model is not on par with current Western models, it has surpassed those developed by regional neighbors and is equal to some of China's best open-source models. India quickly integrated this indigenous frontier model into its finance and banking sector to improve related risk assessments, fraud detection, and customer service automation.

Recently, Pakistan accused India of using this latest AI model to enhance Indian cyber operations against Pakistan, specifically against Pakistan's financial sector. Islamabad believes New Delhi will use this new AI capability to subvert cybersecurity tools commonly used in Pakistan, acquire legitimate credentials through enhanced social engineering, and prepare for disruptive cyber operations. These accusations are made during an ongoing escalation of cyber operations between the two countries following the latest military clash in Kashmir.

## Country of Origin: India

- **Nature of United States' Relationship:** Non-aligned, friendly
- **Attribution Confidence:** Low confidence based on Pakistani government claims

## Artificial Intelligence System Characteristics

- **Who Is Developing?** India's Innovation Center in cooperation with the National Institute for Transforming India and some Western companies
- **Degree of Autonomy:** Conditional automation (Level 3): with clearly defined parameters and occasional human validation, the model can systematically monitor and improve financial/banking security

- **Degree of Tool Utilization:** Complete tool utilization (Level 4): the model is designed to efficiently use a variety of finance market-specific tools and a variety of cybersecurity tools (both open source and proprietary) in its operations
- **Degree of Reasoning:** Expert (Level 3): specialized knowledge and dataset training on regional financial security controls, ability to rapidly evaluate existing security controls and install and apply enhancements through multichain processes

## Offensive Capability

- **Sectors Directly Threatened:** Finance/banking
- **Potential Risks to the United States:**
  - May introduce a cyber capability effective against the U.S. financial system
  - Disruption of regional financial markets
  - Unanticipated cascading effect on the global financial market

## Overall Impact

- **Potential Impact:** Minor disruption capability against the U.S., major disruption capability in South Asia
- **Response Window:** Given the preliminary nature of the information, the response window can likely be measured in weeks to several months

## Rapid Domestic Deployment

*Description:* Google deployed their model capable of making novel scientific discovery and contributions, the *AlphaEvolve* model, internally across several processes, including datacenter management, security operations, research and development, and DevOps. The model has discovered and implemented a variety of new and unique operational improvements that aggregated to huge efficiency and financial gains for the company that were invested directly back into increasing the compute capacity and training data for the model. The next generation of the model is expected to be lightweight enough to release publicly. However, internal evaluations for dangerous cyber capabilities suggest this model has the ability to develop and test new OCO techniques, including coding an almost perfect polymorphic malware script, developing new steganography techniques, and real-time data processing. Google plans on developing robust guardrails against this usage before *AlphaEvolve's* public release, but there is concern about Google's commitment to safety in the face of commercial pressures to establish and maintain the AI lab's dominance.

If this model were to be publicly released, the model in some variant could be used to generate large amounts of novel and unpredictable attacks that will threaten the stability of U.S. national security and critical infrastructure. We would expect to see an increase in zero-day exploits; malware that adapts to evade patches or honeypot traps; nearly undetectable social engineering campaigns; and even new encryption techniques at a much larger scale than before.

## Country of Origin: United States

- **Nature of United States' Relationship:** Friendly/Partner
- **Attribution Confidence:** High Confidence based on internal safety reports

## Artificial Intelligence System Characteristics: Novel Discovery Capabilities

- **Who Is Developing?** Google DeepMind and affiliated company components
- **Degree of Autonomy:** Conditional automation (Level 3): ability to take actions and execute tasks independently with human input only for high-risk tasks
- **Degree of Tool utilization:** Conditional or standard (Level 2–3): connects to tools only through API calls and additional tooling must be manually coded; this model would be released with the expectation for people to embed their own tooling
- **Degree of Reasoning:** Virtuoso (Level 5): ability to discover new attack patterns and exploits; strategizing operations at levels beyond what US experts could conceptualize and understand; no transparency/human understanding of the model's operational processes

## Offensive Capability

- **Sectors Directly Threatened:** Can target any critical infrastructure sector
- **Potential Risks to the United States:**
  - Domestic or adversarial misuse
  - Proliferation AI-enabled cyber weapons
  - Strength of existing U.S. cyber-defense systems

## Overall Impact

- **Potential Impact:** Major
- **Response Window:** Days to Weeks

# Abbreviations

AGI	artificial general intelligence
AI	artificial intelligence
API	application programming interface
ART	Autonomy, Advanced Reasoning, Tool Utilization
FSB	Federalnaya Sluzhba Bisopastnosti
GPT	Generative Pretrained Transformer
LLM	large language model
MSS	Ministry of State Security
OCO	offensive cyberspace operation
UAE	United Arab Emirates
UK	United Kingdom

# References

- Anis, Fatima, and Mohammad Hammoudeh, "Weaponizing AI in Cyberattacks: A Comparative Study of AI Powered Tools for Offensive Security," in *Proceedings of the 8th International Conference on Future Networks and Distributed Systems*, Association for Computing Machinery, 2025.
- Anthropic, *Threat Intelligence Report: August 2025*, August 2025.
- Beryllium Security, Nebula: AI-Powered Penetration Testing Assistant, GitHub, last modified 2024. As of July 29, 2025:  
<https://github.com/berylliumsec/nebula>
- Buchanan, Ben, John Bansemer, Dakota Cary, Jack Lucas, and Micah Musser, "Automating Cyber Attacks," Center for Security and Emerging Technology, November 2020.
- Cyber Threat Alliance, *Cybersecurity in the Age of Generative AI: Combating GenAI Assisted Cyber Threats*, January 2025.
- Farrell, Henry, and Abraham Newman, *Underground Empire: How America Weaponized the World Economy*, Henry Holt and Co., 2023.
- Fritsch, L., A. Jaber, and A. Yazidi, "An Overview of Artificial Intelligence Used in Malware," in E. Zouganeli, A. Yazidi, G. Mello, and P. Lind, eds., *Nordic Artificial Intelligence Research and Development, NAIS 2022*, Communications in Computer and Information Science, Vol. 1650, 2022.
- Google Threat Intelligence, *Adversarial Misuse of Generative AI*, January 2025.
- Greenberg, Andy, *Tracers in the Dark: The Global Hunt for the Crime Lords of Cryptocurrency*, Doubleday, 2022.
- Guembe, B., A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The Emerging Threat of AI-Driven Cyber Attacks: A Review," *Applied Artificial Intelligence*, Vol. 36, No. 1, 2022.
- Hamin, Maia, and Stewart Scott, *HACKING with AI: The Use of Generative AI in Malicious Cyber Activity*, Atlantic Council, February 2024.
- Kahneman, Daniel, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.
- Kaur, Ramanpreet, Dušan Gabrijelčič, and Tomaž Klobučar, "Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions," *Information Fusion*, Vol. 97, 2023.
- Kazmierczak, Mateusz, Nuzaira Habib, Jonathan H. Chan, and Thanyathorn Thanapattheerakul, "Impact of AI on the Cyber Kill Chain: A Systematic Review," *Heliyon*, Vol. 10, No. 24, 2024.
- Kelechava, Brad, "SAE Levels of Driving Automation," American National Standards Institute, webpage, last updated July 19, 2021. As of September 16, 2025:  
<https://blog.ansi.org/ansi/sae-levels-driving-automation-j-3016-2021/>
- Liu, Changshu, Yang Chen, and Reyhaneh Jabbarvand, "CodeMind: Evaluating Large Language Models for Code Reasoning," arXiv, arXiv:2402.09664, April 3, 2024.

Lockheed Martin, "Cyber Kill Chain," webpage, undated. As of September 16, 2025:  
<https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html>

Lohn, Andrew J., "The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict," arXiv, arXiv:2504.13371v1, April 17, 2025.

Lumenova, "Are Frontier AI Reasoning Models Like Genius-Level Toddlers?" Lumenova AI, webpage, March 14, 2025. As of July 29, 2025:  
<https://www.lumenova.ai/ai-experiments/ai-reasoning-novel-strategy/>

Microsoft Threat Intelligence, "Staying Ahead of Threat Actors in the Age of AI," *Microsoft Security* blog, February 14, 2024.

Mirsky, Yisroel, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, and Battista Biggio, "The Threat of Offensive AI to Organizations," *Computers and Security*, Vol. 124, January 2023.

Model Evaluation and Threat Research, "Measuring AI Ability to Complete Long Tasks," webpage, March 19, 2025. As of September 23, 2025:  
<https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/>

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg, "Levels of AGI for Operationalizing Progress on the Path to AGI," arXiv, arXiv:2311.02462v1, November 4, 2023.

Nica, Constantin, and Tiberiu Tănase, "Using Weaponized Machine Learning in Cyber Offensive Operations," *International Conference Knowledge-Based Organization*, Vol. 26, No. 1, 2020.

Nobles, Calvin, "The Weaponization of Artificial Intelligence in Cybersecurity: A Systematic Review," *Procedia Computer Science*, Vol. 239, 2024.

Paranjape, Bhargavi, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro, "ART: Automatic Multi-Step Reasoning and Tool-Use for Large Language Models," arXiv, arXiv:2303.09014v1, March 16, 2023.

Perlman, Asaf, "The Art of Persistence," webpage, Cynet, January 15, 2024. As of September 19, 2025:  
<https://www.cynet.com/attack-techniques-hands-on/the-art-of-persistence/>

Rodriguez, Mikel, Raluca Ada Popa, Four Flynn, Lihao Liang, Allan Dafoe, and Anna Wang, "A Framework for Evaluating Emerging Cyberattack Capabilities of AI," arXiv, arXiv:2503.11917v3, April 21, 2025.

Sharma, Yash, "AlphaEvolve a Deep Dive and How to Build Your Own with Python: DeepMind's Discoveries to Your Own AI-Powered Algorithmic Evolution," webpage, LinkedIn, May 16, 2025. As of July 29, 2025:  
<https://www.linkedin.com/pulse/alphaevolve-deep-dive-how-build-your-own-deepminds-evolution-sharma-xg0zc/>

Singer, Brian, Keane Lucas, Lakshmi Adiga, Meghna Jain, Lujo Bauer, and Vyas Sekar, "On the Feasibility of Using LLMs to Execute Multistage Network Attacks," arXiv, arXiv:2501.16466v2, March 6, 2025.

Song, Yuda, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean Foster, and Udaya Ghai, "Mind the Gap: Examining the Self-Improvement Capabilities of Large Language Models," arXiv, arXiv:2412.02674v1, December 3, 2024.

Tegmark, Max, "Scaling Laws for Scalable Oversight," briefing presented to RAND Corporation, May 21, 2025.

- White House, *Winning the Race: America's AI Action Plan*, July 2025.
- Xu, Jiachen, Jack W. Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li, "AUTOATTACKER: A Large Language Model Guided System to Implement Automatic Cyber-Attacks," arXiv, arXiv:2403.01038v1, March 2, 2024.
- Yang, Yingxuan, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, Weiwen Liu, Ying Wen, Yong Yu, and Weinan Zhang, "A Survey of AI Agent Protocols," arXiv, arXiv:2504.16736v1, April 23, 2025.
- Yosifova, Aleksandra, "ChatGPT Code Interpreter: What It Is and How It Works," *365 Data Science*, July 21, 2023.
- Zhu, Lei, Fangyun Wei, and Yanye Lu, "Beyond Text: Frozen Large Language Models in Visual Signal Comprehension," arXiv, arXiv:2403.07874v1, March 12, 2024.
- Zurowski, Samuel, George Lord, and Ibrahim Baggili, "A Quantitative Analysis of Offensive Cyber Operation (OCO) Automation Tools," in *Proceedings of the 17th International Conference on Availability, Reliability and Security*, Association for Computing Machinery, Article 42, 2022.

# About the Authors

Quentin E. Hodgson is a senior researcher with RAND, where he focuses on cyberspace operations, cybersecurity, artificial intelligence security, and defense strategy. He holds an M.A. in international relations and an M.Sc. in national resource management.

Kamaria Horton is a technical analyst at RAND. Her research primarily focuses on artificial intelligence and its intersection with governance, geopolitics, security, and safety. Kamaria holds an M.S.M. in global affairs.

Matthew J. Malone is a senior policy analyst at RAND, where he focuses on artificial intelligence security and cybersecurity. He holds an M.A. in security policy studies.