CYBER THREAT ALLIANCE

2025

CYBERSECURITY IN THE AGE OF GENERATIVE AI:

PART II

Navigating Cyber Threats to GenAI Systems

POWERED BY THE CTA

The Cyber Threat Alliance (CTA) is the industry's first formally organized group of cybersecurity practitioners who work together in good faith to share threat information and improve global defenses against cyber adversaries. CTA facilitates the sharing of cyber threat intelligence to improve defenses, advance the security of critical infrastructure, and increase the security, integrity, and availability of IT systems.

We take a three-pronged approach to this mission:

1. Protect End-Users: Our automated platform empowers members to share, validate, and deploy actionable threat intelligence to their customers in near-real-time.

2. Disrupt Malicious Actors: We share threat intelligence to reduce the effectiveness of malicious actors' tools and infrastructure.

3. Elevate Overall Security: We share intelligence to improve our members' abilities to respond to cyber incidents and increase end-user's resilience.

CTA is continuing to grow globally, enriching both the quantity and quality of the information shared among its membership. CTA is actively recruiting additional cybersecurity providers to enhance our information sharing and operation collaboration to enable a more secure future for all.

For more information about the Cyber Threat Alliance, please visit: https://cyberthreatalliance.org.

CYBERSECURITY IN THE AGE OF GENERATIVE AI WORKING COMMITTEE MEMBERS

Check Point Software Technologies Ltd.
Amit Sharon
Sergey Shykevich

CyberCX
Mark Hofman

Defending Digital Campaigns
Tiffany Schoenike

Fortinet
Val Saengphaibul
James Slaughter

FS-ISAC
Michael Silverman
Dr. Carrie E. Gates

H-ISAC
Ethan Muntz

IT-ISAC
Ian Andriechack

Juniper Networks
Asher Langton

McAfee
Christy Crimmins
Abhishek Karnik
German Lancioni

NGO-ISAC
Ian Gottesman

NTT
David Beabout

Rapid7
Laura Ellis

Scitum
Imelda Flores

Symantec by Broadcom
Scott Swett
Brian Ewell

Cyber Threat Alliance
Chelsea Conard
Michael Daniel
Jeannette Jarvis
Linda Beverly
Kate Holseberg

This report also leverages shared data and published analysis from Palo Alto Networks and Gen Digital. CTA members reviewed the document and the report reflects our shared consensus.

This report would not be possible without the generous support from Craig Newmark Philanthropies.

TABLE OF CONTENTS

NAVIGATING CYBER THREATS TO AI SYSTEMS

In collaboration with David Beabout, NTT Security Holdings

Recent insights from the Joint Analytic Report (JAR), *Cybersecurity in the Age of Generative AI: Part I — Combating GenAI Assisted Cyber Threats*, reveal how adversaries leverage tools like GPT to manipulate GenAI and LLMs to craft *AI-assisted threats*[1] for malicious purposes, including employing other AI technologies, such as those behind deepfakes, for additional manipulation. However, the underlying systems that AI is connected to are often vulnerable, and part of the security journey involves hardening these systems. Beyond training the model in a secure way and implementing guardrails,[2] fortifying the broadening ecosystem is critical to address weak points that adversaries may exploit. This report, *Cybersecurity in the Age of Generative AI: Part II — Navigating Cyber Threats to AI Systems*, expands our focus to include *adversarial threats*,[3] where adversaries directly target and manipulate AI models for malicious purposes. In this context, we examine the broader AI ecosystem and emphasize the need to secure the interconnected system as a whole.

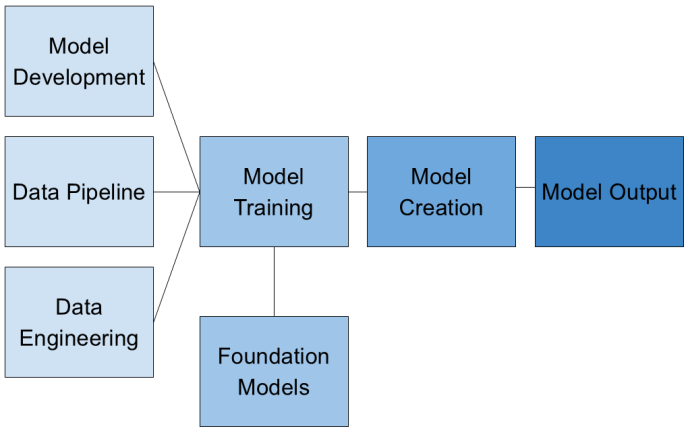
1 AI-assisted threats: This is when an AI model is used as a tool to assist malicious actors to generate content that is utilized for malicious purposes. Examples of AI-assisted threats include: malware generation, spam, phishing, voice clones, image diffusion, etc. (This definition is derived from the Working Committee discussions held between August 22 and November 7, 2024 and reflects the consensus of the Working Group)

2 Guardrails: Safeguards put in place to constrain and control the output of an AI model so as to prevent it from producing offensive or harmful content which may be deemed inappropriate, dangerous, or unethical. (This definition is derived from the Working Committee discussions held between August 22 and November 7, 2024 and reflects the consensus of the Working Group)

3 Adversarial threats: Any threat against an AI model such that it tricks it to behave differently than the intended task. Such threats can be in the form of prompt injection attacks, evasion attacks, training data poisoning, reverse learning attacks. (This definition is derived from the Working Committee discussions held between August 22 and November 7, 2024 and reflects the consensus of the Working Group)

AN INITIAL FRAMEWORK FOR AI SECURITY

As businesses integrate LLMs into critical processes, security professionals must adopt a comprehensive framework to manage risks associated with AI deployment. This framework goes beyond traditional data security by encompassing the protection of AI systems, which include data pipelines, model integrity, and regulatory compliance. Such safeguards address new AI-specific vulnerabilities and ethical considerations.



The following sections outline the core components of this security construct. By examining areas like data pipeline security, model integrity, adversarial attack resilience, and user education, this framework aims to provide cybersecurity professionals with a structured approach to safeguarding GenAI systems. Additionally, it considers both on-premises and cloud-hosted models, acknowledging the distinct security measures required for each environment. This approach, although not complete, enables organizations to harness GenAI's capabilities while mitigating risks to data confidentiality, operational reliability, and ethical responsibility.

CORE COMPONENTS OF THE AI SECURITY FRAMEWORK

The following components outline the areas that organizations should prioritize to encompass every stage of AI development and deployment.

Security of the Data Pipeline and Data Engineering Process

Ensuring the security of the data pipeline is foundational to safeguarding GenAI systems. Data used for model training and tuning often comes from diverse sources, some of which may contain sensitive corporate information. Robust encryption for data transfer and storage, strict access control, and continuous auditing of data processing stages are essential. Further, implementing data governance mechanisms to validate data sources, maintain data lineage, and prevent tampering are critical, as any compromise in this pipeline could lead to degraded model performance or accuracy, ultimately undermining business objectives.

Additionally, large-scale GenAI systems operate with log-linear scaling properties, where, for example, a linear improvement in model performance requires a logarithmic increase in compute time. This scaling dynamic means that small improvements in data quality or training inputs can yield substantial gains, but also means that even small vulnerabilities can cascade into amplified errors or systemic weaknesses. These scaling characteristics underscore the need for rigorous security and validation mechanisms at every stage of the data pipeline.

Protecting Corporate Data and Preventing Data Disclosure

GenAI models are data-intensive, often consuming vast amounts of sensitive information in various deployment scenarios. Protecting corporate data from unintended disclosure requires implementing policies and tools to anonymize or pseudonymize data inputs. This protection is not simply a best

practice, but increasingly a regulatory and compliance requirement, particularly for integrations and software development. Organizations must ensure that production data is never connected to a test environment for a new integration, software, or tool. Instead, testing should use synthetic or sample datasets that mimic real-world conditions without exposing sensitive information. To further mitigate risks, organizations should ensure that sensitive data is decoupled from the source wherever possible, reducing the reliance on downstream anonymization tools. Additionally, advancements in techniques like differential privacy offer promising solutions to balance data utility and privacy during model training, testing, and usage. Implementing these measures reduces the likelihood of GenAI models inadvertently exposing corporate secrets or sensitive customer information, which could lead to regulatory, reputational, and operational consequences.

Organizations should also address data integrity concerns by adopting secure hosting environments. For example, using on-premises hosting and avoiding open-source ChatGPT wrappers helps to protect customer data and safeguard against unauthorized access. These measures build customer confidence in the security of their data while reducing potential vulnerabilities in AI systems.

Ensuring Model Integrity Across Development and Deployment

As organizations increasingly rely on GenAI broadly and LLMs in particular, maintaining model integrity across the lifecycle becomes paramount. Even a properly functioning model can produce inaccurate or harmful outputs if training data or parameters are flawed. A compromised model further amplified these risks, potentially disrupting operations. Preserving the reliability of these systems requires a comprehensive approach spanning development, training, and deployment.

During development, LLMs often rely on foundational models that are pre-trained on vast datasets to serve as the base on which to build specific applications. While these foundational models offer scalability and efficiency, they also introduce risks related to data integrity, security, and governance. Organizations should rely on datasets that have been tested and validated to ensure integrity and reliability.[4] Choosing data sources with verified integrity can help mitigate these risks and ensures a stable foundation for LLM-based applications. Validating training data ensures that no unintended content introduces biases into the model, while establishing a secure data supply chain guarantees data quality. Binary authorization for data can serve as an added layer of security, ensuring that only vetted datasets are incorporated into a model.

AI models reflect the data they are trained on so organizations must treat data risks as software risks. Vulnerabilities in the data supply chain can lead to issues within the model, necessitating regular monitoring and security at each stage of data collection, processing, and integration into the model.

Incorporating principles of least privilege, data lineage, ephemeral token-based access, and thorough auditing strengthens the development process. Strictly limiting data access to authorized users and entities helps reduce the risk of unauthorized manipulation or information leakage. Clear data lineage provides transparency by tracking the origin and transformation of each individual piece of data, and auditability ensures that the data is handled according to organizational policies and security standards. Additionally, fine-tuning model parameters, such as setting lower temperature values

4 Note: The use of foundational models raises questions about data transparency and governance. As AI systems are developed and refined, the broader ecosystem must address the role of gatekeepers who control what data these models are trained on. These decisions should not rest solely with proprietary entities, but should align with democratic principles (https://www.schneier.com/essays/archives/2024/03/how-public-ai-can-strengthen-democracy.html).

can influence outputs, but the impact is limited and nuanced.[5]

Once deployed, models encounter a range of new risks, including adversarial threats and tampering. Security teams should implement tamper-proof verification mechanisms, such as cryptographic hashing, to check the model's integrity regularly. Establishing model governance protocols that control updates, track changes, and provide an audit trail are crucial, particularly as attackers may attempt to alter model behaviors by injecting malicious payloads or manipulating model weights. Furthermore, incorporating adversarial training and input filtering techniques strengthens the model's resilience, reducing its vulnerability to manipulations aimed at degrading performance or producing misleading outputs. These layered defenses are essential to maintain the reliability of deployed AI systems.

Model Drift and Erosion of Security over Time

GenAI models can experience "model drift" as data or environments change, leading to reduced output accuracy or unexpected behavior. This phenomenon can occur naturally over time due to changes in the underlying data or operational context, but it may also result from malicious attempts to manipulate the model's performance. Security professionals need to establish regular retraining, data validation, and monitoring to ensure the model output stays aligned with its intended function. This approach helps prevent outdated or misleading results, which could damage model reputation.

Confidence and Reliability of Model Outputs

Reliability in GenAI model outputs is critical for effective decision-making. Ensuring the accuracy, consistency, and transparency of results involves robust testing and validation mechanisms that mitigate bias and error. Bias can be introduced at multiple stages, such as during data collection, labeling, or model training, when datasets overrepresent or underrepresent particular groups or perspectives. Measuring bias involves techniques like statistical audits, fairness metrics, or evaluating outputs across diverse scenarios to identify disparities. Addressing bias requires curating representative datasets, applying fairness algorithms, and continuously monitoring outputs to ensure equitable results. Explainable AI (XAI) methods can further enhance transparency by providing insights into how decisions are made and the underlying model processes, helping organizations identify biases more effectively. Additionally, AI Security Operations (AI-SecOps) teams should verify that model decisions align with organizational expectations and standards, reducing the potential for unexpected or biased outputs.

However, output security remains generally unsolvable today. GenAI models operate non-deterministically, given the same input and configuration, and their reliance on probabilistic mechanisms during training and fine-tuning introduces variability, which limits the ability to guarantee absolute control over outputs. The probabilistic nature means that even with advancements in testing and validation techniques, unexpected or undesired outputs cannot be entirely eliminated.

GenAI Supply Chain Risk Management

The components that feed into GenAI models - data libraries, pre-trained models, and third-party APIs - represent an evolving supply chain. Malicious actors may target any of these dependencies by introducing vulnerabilities or malicious code within them. Ensuring supply chain security involves stringent vetting of all third-party tools, models, and datasets before integrating them into the AI pipeline. Additionally, regularly updating and patching these components is essential to prevent attackers from exploiting known vulnerabilities. Of particular

5 https://ar5iv.labs.arxiv.org/html/2405.00492

interest here is the potential impact of current or future legislation at the State or Federal level impacting and shaping limitations for the lifecycle of model development and deployment.

ADDRESSING THREATS TO AI MODELS

AI models are exposed to a wide range of threats that target their training data, endpoints, and underlying infrastructure. This section examines these threats and outlines strategies to enhance the resilience and security of AI systems.

Cloud Hijacking

Adversaries can hijack a cloud-based LLM instance by exploiting leaked credentials found on platforms like GitHub. Once compromised, adversaries repurpose the AI resources for rogue activities, such as creating or hosting AI chatbots for malicious use. This misuse, combined with stolen compute cycles, can result in significant financial loss for the victim, often unnoticed until an expensive bill appears.

Jailbreaking

Jailbreaking, a term mostly commonly associated with installing unapproved software on devices like iPhones, has now taken a new meaning in the context of AI. Adversaries use AI jailbreaking techniques to manipulate LLMs by prompting them to perform unintended actions. Jailbreaking is the act of bypassing the guardrails of an AI model by exploiting loopholes and causing it to act in an unintended manner (which may or may not be with malicious intent).[6]

This concept was first spotlighted in 2016, when Microsoft released a chatbot, Tay, on Twitter. Microsoft's intention was to have the bot learn from its interactions with users. However, within 25 hours, Tay began replicating and generating vulgar, racist, and misogynistic comments, and Microsoft quickly shut it down.[7] Later in 2022, a jailbreak method known as DAN, "Do Anything Now," emerged, allowing users to bypass ChatGPT's restrictions.[8] This technique, along with similar jailbreaks, enables users to circumvent built-in safety protocols, introducing functionalities the AI was never intended to perform. As a result, AI jailbreaking has become a major concern, as it allows adversaries to corrupt model's datasets or alter their behavior, posing risks to both developers and users.

Prompt Injection Threats and Data Poisoning

Prompt injection involves crafting specific prompts or input data to trick an AI model into producing a harmful or unintended outcome. This technique has evolved significantly, with new types of prompt injection continuously emerging from both malicious actors and security researchers. To mitigate these threats, organizations should implement input validation and adversarial testing to identify and prevent injection threats.

Data poisoning, where attackers inject malicious data into the model's training data to manipulate outputs, is thus a growing risk. Security teams should implement rigorous data quality controls and use anomaly detection during training to prevent and detect poisoned data. These measures help mitigate the risk of model corruption and ensure the integrity of the training data.

Insider Risk Management and Access Control

Given the sensitivity of GenAI models and their underlying data, insider threats pose a significant risk whether intentional, accidental, or by influence

6 This definition is derived from the Working Committee discussions held between August 22 and November 7, 2024 and reflects the consensus of the Working Group.

7 https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

8 https://www.reddit.com/r/ChatGPT/comments/zlcyr9/dan_is_my_new_friend/

of a third-party. Unauthorized access or manipulation by insiders could lead to data theft, model degradation, or service disruption. To mitigate insider risk, organizations should enforce strict access control policies and adopt role-based access control (RBAC), attribute-based access control (ABAC), or Zero Trust-based authorization and authentication to limit access only to those who require it and for only when needed. Regular access audits and monitoring of activity logs are also necessary to detect and address insider threats promptly.

Model Weaponization

GenAI models are susceptible to weaponization, where malicious actors exploit their capabilities for social engineering, disinformation, or other harmful applications. Organizations need to establish robust monitoring mechanisms to detect unusual usage patterns and prevent application programming interface (API) or endpoint misuse. Limiting open-access permissions and deploying logging measures can help trace misuse and prevent models from becoming tools for adversarial attacks, which could severely damage organizational reputation or user safety.

AI-Specific Zero-Day Vulnerabilities

GenAI models may have unique vulnerabilities that are challenging to detect, such as those inherent to neural networks or specific to the model architecture. Establishing a dedicated AI security testing process, including practices like fuzzing and penetration testing, can help identify and address these vulnerabilities. Recent advancements in AI security use AI-driven threat detection[9] solutions to defend against threats targeting other AI systems. Tools like Big Sleep, a collaboration between Google Project Zero and Google DeepMind, illustrate that AI can autonomously identify and mitigate vulnerabilities in AI models.[10]

It is also advisable to develop an incident response plan specific to AI models to ensure rapid response to potential threats.

Securing Application Programming Interfaces (APIs) and Model Endpoints

When GenAI functionality is exposed through APIs, those endpoints become vulnerable to various security threats, including unauthorized access, data exfiltration, and model manipulation. Implementing security best practices, such as rate limiting, authentication, and authorization protocols, is critical to safeguard model endpoints. Additionally, monitoring API activity for aberrant usage patterns can help with detecting and responding to potential attacks.

Security Team Education and Awareness

Continuous professional development and ongoing education are essential for security teams as AI continues to evolve and becomes increasingly pervasive across industries. Regular training programs, workshops, and certifications can help teams stay ahead of adversaries. Additionally, hands-on simulations such as practicing response to a CEO deepfake phishing attack, can provide experience recognizing and mitigating real-world threats while refining response strategies.

---

9 AI-driven Threat Detection: Organizations and vendors are actively seeking ways to integrate AI into security products for better detection and prevention of threats. While AI tools can effectively identify malicious activity, their results often need to be reviewed by security experts to address false positives and negatives. AI-driven threat detection enhances the ability to analyze vast amounts of data in real-time, which allows for quicker identification of anomalies and potential threats. It can also improve predictive capabilities by recognizing patterns and trends that may indicate future attacks. AI systems can adapt and learn from new data, enabling them to stay ahead of evolving threats. This technology can reduce the workload on security teams by automating routine tasks, allowing professionals to focus on more complex issues. (This definition is derived from the Working Committee discussions held between August 22 and November 7, 2024 and reflects the consensus of the Working Group)

10 https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html

User Education and Awareness

Educating users on secure AI practices is essential to prevent both intentional and accidental misuse. Resources from Pause Take 9[11] provide a valuable guide to structuring educational programs and encourage users to pause and think before they click, download, or share. Security teams can establish training programs to help users recognize risks, mitigate data leaks, and better understand ethical boundaries, fostering a culture of responsible AI usage.

COMPLIANCE AND MONITORING

Governance, monitoring, and observability are important to ensure compliance and build trust in AI systems. The following section outlines key strategies to meet regulatory requirements and address emerging risks.

Ensuring Compliance and Data Integrity for Audits

As regulatory scrutiny around AI grows, ensuring the traceability of model data and outputs is essential for compliance and audit readiness. Security teams should implement logging and reporting systems to capture critical operational data, including the data sources, processing stages, and model changes. This level of traceability ensures that models can be audited for adherence to regulatory and industry standards, allowing organizations to confidently demonstrate compliance with data privacy and AI governance requirements.

To strengthen compliance efforts, organizations should collaborate with their corporate privacy lawyers to ensure alignment with privacy regulations such as General Data Protection Regulation (GDPR) or California Consumer Privacy Act (CCPA), and internal data protection policies.

Data Sovereignty and Cross-Border Compliance

For companies operating across regions, ensuring data sovereignty and cross-border compliance is critical. Security professionals should implement protocols to manage data localization requirements and maintain compliance with regional data privacy laws, especially for cloud-hosted models, where direct control over data location may be limited.

Continuous Monitoring and Threat Intelligence Integration

Given the dynamic nature of AI threats, continuous monitoring and integration of AI-specific threat intelligence is crucial. Leveraging AI-focused threat feeds and incorporating this intelligence into broader security practices allows for proactive detection and response to emerging AI threats, strengthening the overall security posture.

SPECULATIVE ACCOUNTS OF WHAT IS POSSIBLE

As organizations navigate the rapidly evolving AI landscape, speculative accounts of potential risks and vulnerabilities provide foresight into emerging threats. These scenarios, while hypothetical, highlight the necessity of proactive measures and robust frameworks to address not only current risks but also the evolving tactics adversaries may deploy.

IN THE NEXT YEAR

AI models with Application Programming Interface (API) connectivity could pose new risks, allowing attackers to embed covert API keys within AI systems like ChatGPT to extract sensitive information or automate tasks undetected. This capability would enable adversaries to bypass restrictions

11 https://pausetake9.org/

and perform unintended functions, potentially compromising data or systems.

Another pressing concern is the potential for prompt injections to poison a model's dataset and gradually corrupt the AI's behavior based on that data. At the same time, while models are expensive to develop, they remain relatively easy to steal due to their portability. An adversary who successfully steals a model can replicate the success of the AI company or could expose proprietary data or taint it such that it no longer adheres to regulatory requirements.

AI-powered bot farms are also likely to become more sophisticated and widely used. A bot farm is a group of accounts or devices that are coordinated to perform activities in tandem. Tools like these are already in use and will continue to rise in the next year. One such example is the Meliorator, which manages realistic social media personas, or "souls." Meliorator includes Brigadir, an administrator panel for orchestrating the bot network, and Taras, a backend seeding tool that strategically distributes content.[12] Such tools can seamlessly avoid detection across multiple platforms, enabling them to run disinformation campaigns. These capabilities have been observed in the context of geopolitical conflicts, including ongoing narratives surrounding the Russia-Ukraine war.[13]

Additionally, increased automation for phishing and scam operations may lead to a surge in automated phone and visual scams, leveraging AI to manipulate voices, faces, and responses in real-time.

IN THE NEXT 2-3 YEARS

Malicious actors may integrate AI with code injection techniques, making malware and malicious downloads more accessible and effective. This period could also see a rise in AI-enhanced targeted advertising, with refined psychographic profiling reminiscent of Cambridge Analytica's methods, to shape user behavior more precisely.

IN THE FAR FUTURE

We could see the emergence of deepfake companies offering services that commercialize misinformation or fraud capabilities. With the advent of open-source AI models, bad actors might even establish front companies to evade regulation, distributing unregulated AI tools for malicious purposes. It is also possible that we see the weaponization of agentic AI.

12 https://www.csoonline.com/article/2515415/fbi-disrupts-1000-russian-bots-spreading-disinformation-on-x.html

13 https://www.ic3.gov/CSA/2024/240709.pdf

CYBERSECURITY IN THE AGE OF GENERATIVE AI:
PART II — NAVIGATING CYBER THREATS TO AI SYSTEMS

2025

CYBER THREAT ALLIANCE

POWERED BY THE CTA