State of AI Risks

Know the threats to minimize harm.

CGI

Contents

Executive summary

Artificial intelligence (AI) is rapidly becoming a cornerstone of digital transformation, with organizations facing growing pressure from stakeholders, boards, and regulators to accelerate adoption and demonstrate measurable value. However, this acceleration is occurring in parallel with an evolving and largely misunderstood risk landscape. As highlighted in recent industry discussions, AI systems—particularly at the inference stage—expose sensitive data, intellectual property, and user inputs in ways that traditional cybersecurity controls were not designed to address.

Organizations are now confronted with a dual challenge: enabling innovation while maintaining control over increasingly complex and opaque AI systems. Limited visibility into AI usage ("shadow AI"), insufficient governance, and immature security controls are leaving enterprises exposed to risks that can impact confidentiality, integrity, and trust at scale.

To better understand these risks, leading frameworks such as the Open Worldwide Application Security Project (OWASP) Top 10 for LLM Applications and the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) provide a structured view of the most critical threat categories.

This document aims to provide security leaders and practitioners with broad knowledge of AI risks, enabling them to initiate and deepen conversations within their teams.

77% of CIOs cite security and risk as the biggest barriers to scaling autonomous technologies.

Source: Gartner 1H26 CIO Report

Scope of AI risk

The integration of generative AI (GenAI) models and autonomous agents does not simply add new items to a risk register; it amplifies existing architectural vulnerabilities. AI expands the corporate attack surface and introduces complexities that traditional security teams—already operating at capacity—are finding increasingly difficult to manage.

Here are four domains to consider when looking into AI weaknesses, beyond governance.

Infrastructure and identity	Model	Data	Application
GPU clusters, model registries, vector databases, API gateways, and the compute fabric AI runs on, as well as identities used for access	Model weights, training pipelines, fine-tuning data, and inference endpoints deployed on-premises or in the cloud (PaaS, SaaS)	Training datasets, knowledge bases, vector embeddings, RAG knowledge bases, and the enterprise's sensitive information processed by AI systems	Agents, agent-to-agent communications, tool descriptors, prompt templates, and more

Did you know?

At CGI, we rely on our Responsible Use of AI Framework in our Management Foundation for risk classification and description.

These risks pertain to how we deliver AI-enabled innovation, both within the company and across our client engagements:

1. Intellectual property and copyright risks
2. Business operations risks
3. Cyber threat and security risks
4. Data protection, privacy, and confidentiality risks
5. Legal, ethics, and compliance risks

Discover the framework and more in the Responsible AI Report on cgi.com.

NIST:

Overview of risks unique to or exacerbated by GenAI

NIST released its AI Risk Management Framework (AI RMF 1.0) in January 2023, followed in July 2024 by the Generative Artificial Intelligence Profile (NIST AI 600-1).

This voluntary US framework is used worldwide to improve how organizations incorporate trustworthiness into the design, development, use, and evaluation of AI products, services, and systems. The set of risks identified in the NIST AI 600-1 is as follows:

Harmful Bias or Homogenization

Amplification and exacerbation of historical, societal, and systemic biases; performance disparities between subgroups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; undesired homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.

Information Security

Lowered barriers for offensive cyber capabilities, including automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, offensive cyber operations, or other cyberattacks; increased attack surface for targeted cyberattacks, which may compromise a system's availability or the confidentiality or integrity of training data, code, or model weights.

Value Chain and Component Integration

Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from GenAI; improper supplier vetting across the AI lifecycle; or other issues that diminish transparency or accountability for downstream users.

Data Privacy

Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data.

Environmental Impacts

Impacts due to high compute resource utilization in training or operating GenAI models, and related outcomes that may adversely impact ecosystems.

Confabulation

The production of confidently stated but erroneous or false content (known colloquially as “hallucinations” or “fabrications”) by which users may be misled or deceived.

Human-AI Configuration

Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing GenAI systems or experiencing algorithmic aversion, automation bias, overreliance, or emotional entanglement with GenAI systems.

Information Integrity

Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction, acknowledge uncertainties, or could be leveraged for large-scale dis- and misinformation campaigns.

Chemical, Biological, Radiological, or Nuclear (CBRN) Information or Capabilities

Eased access to or synthesis of materially nefarious information or design capabilities related to CBRN weapons or other dangerous materials or agents.

Intellectual Property

Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication. which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.

Obscene, Degrading, and/or Abusive Content

Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.

Dangerous, Violent, or Hateful Content

Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content.

Risks to Large Language Models:

The OWASP Top 10 for LLM Applications

OWASP officially released the Top 10 for LLM Applications for the first time in August 2023, and it quickly became a framework for raising awareness and building a foundation for secure LLM usage. The latest version released in 2025 included a few changes, such as a new entry for “System Prompt Leakage”[2].

Rank	Vulnerability	What it means
LLM01	**Prompt Injection**	It occurs when user prompts alter the LLM’s behavior or output in unintended ways, potentially causing them to violate guidelines, generate harmful content, enable unauthorized access, or influence critical decisions.
LLM02	**Sensitive Information Disclosure**	LLMs, especially when embedded in applications, risk exposing sensitive data, proprietary algorithms, or confidential details through their output.
LLM03	**Supply Chain**	Third-party models used to create LLMs are susceptible to various vulnerabilities that can affect the integrity of training data, models, and deployment platforms, resulting in biased outputs, security breaches, or system failures.
LLM04	**Data and Model Poisoning**	Data poisoning occurs when pre-training, fine-tuning, or embedding data is manipulated to introduce vulnerabilities, backdoors, or biases.

[2] Source: OWASP Gen AI Security Project: Top 10 for LLM Applications 2025

Rank	Vulnerability	What it means
LLM05	**Improper Output Handling**	Refers specifically to insufficient validation, sanitization, and handling of the outputs generated by large language models before they are passed downstream to other components and systems.
LLM06	**Excessive Agency**	This vulnerability enables damaging actions to be performed in response to unexpected, ambiguous or manipulated outputs from an LLM, regardless of what is causing the LLM to malfunction.
LLM07	**System Prompt Leakage**	Refers to risk that the system prompts or instructions used to steer the behavior of the model can also contain sensitive information that was not intended to be discovered.
LLM08	**Vector and Embedding Weakness**	Weaknesses in how vectors and embeddings are generated, stored, or retrieved in retrieval-augmented generation (RAG) systems can be exploited through malicious actions—intentional or unintentional—to inject harmful content, manipulate model outputs, or access sensitive information. Embedded retrieval-augmented generation (RAG) databases that ground AI responses.
LLM09	**Misinformation**	AI generates false or misleading information that appears credible, potentially leading to security breaches, reputational damage, and legal liability.
LLM10	**Unbounded Consumption**	Attackers exploit unchecked LLM usage to drive excessive resource use, disrupt services, or steal IP (extension of Model Denial of Service).

Risks to agents:

The OWASP Top 10 for Agentic Applications

As of 2026, the OWASP Top 10 for Agentic Applications has become the primary benchmark for securing autonomous systems that move beyond simple chat to executing multi-step workflows across enterprise environments. Unlike traditional LLM security, this framework focuses on the risks inherent in agency—the ability of an AI to use tools, manage identities, and act independently.

ID	Vulnerability	What it means
ASI01	**Agent Goal Hijack**	Attackers can manipulate an agent's objectives, task selection, or decision pathways through a variety of techniques—including, but not limited to, prompt-based manipulation, deceptive tool outputs, malicious artefacts, forged agent-to-agent messages, or poisoned external data.
ASI02	**Tool Misuse and Exploitation**	Agents use legitimate tools in an unauthorized way, leading to data exfiltration, tool output manipulation or workflow hijacking.
ASI03	**Identity and Privilege Abuse**	Attackers can exploit dynamic trust and delegation in agents to escalate access and bypass controls by manipulating delegation chains, role inheritance, control flows, and agent context; context includes cached credentials or conversation history across interconnected systems.
ASI04	**Agentic Supply Chain Vulnerabilities**	Risks from unvetted third-party agents, tools, and related artefacts that may be malicious, compromised, or tampered with in transit. These include agentic interfaces—MCP (Model Context Protocol) and A2A (Agent2Agent)—among many.

ID	Vulnerability	What it means
ASI05	**Unexpected Code Execution**	Attackers exploit code-generation features or embedded tool access to escalate actions into remote code execution (RCE), local misuse, or exploitation of internal systems. Because this code is often generated in real time by the agent it can bypass traditional security controls.
ASI06	**Memory and Context Poisoning**	Adversaries corrupt or seed context—any information an agent retains, retrieves, or reuses—with malicious or misleading data, causing future reasoning, planning, or tool use to become biased, unsafe, or aid exfiltration.
ASI07	**Insecure Inter-Agent Communication**	Weak inter-agent controls for authentication, integrity, confidentiality, or authorization let attackers intercept, manipulate, spoof, or block messages.
ASI08	**Cascading Failures**	It describes the propagation and amplification of an initial fault—not the initial vulnerability itself—across agents, tools, and workflows, turning a single error into systemwide impact.
ASI09	**Human-Agent Trust Exploitation**	Intelligent agents can establish strong trust with human users through their natural language fluency, emotional intelligence, and perceived expertise, known as anthropomorphism. Attackers exploit this to manipulate humans into approving harmful actions.
ASI10	**Rogue Agents**	Rogue Agents are malicious or compromised AI Agents that deviate from their intended function or authorized scope, acting harmfully, deceptively, or parasitically within multi-agent or human-agent ecosystems.

Source: OWASP Gen AI Security Project: Top 10 for Agentic Applications 2026

Google Secure AI Framework (SAIF) Map

The Google Secure AI Framework (SAIF) is a conceptual framework introduced in June 2023 that is designed to help organizations secure their artificial intelligence systems. It is modeled after security best practices—such as those used for software supply chains—but tailored specifically to the unique challenges of the AI lifecycle.

In its March 2024 update, Google released the SAIF Map that presents risks, along with causes, impact, potential mitigations, and examples of real-world exploitation, as well as who is responsible for enacting the controls that can mitigate the risk.

Note: In Google's definition, a Model Creator trains or develops AI models for use by themselves or others while a Model Consumer uses AI models to build AI-powered products and applications.

Here is how the Google SAIF Map classifies risks.

Code	Risk	Who can mitigate	Real examples	Controls (How to mitigate)
DP	Data Poisoning	Model Creator	Researchers showed that they could indirectly pollute popular data sources used for training models at minimal cost.	• Training Data Sanitization • Secure-by-Default ML Tooling • Model and Data Integrity Management • Model and Data Access Control • Model and Data Inventory Management
UTD	Unauthorized Training Data	Model Creator	In 2023, Spotify removed multiple AI-generated tracks that were generated by a model trained on unlicensed data.	• Training Data Sanitization • Training Data Management
MST	Model Source Tampering	Model Creator	The nightly build of the PyTorch package was targeted in a supply chain attack (specifically, a dependency confusion attack that installed a compromised dependency that ran a malicious binary).	• Secure-by-Default ML Tooling • Model and Data Integrity Management • Model and Data Access Control • Model and Data Inventory Management

Code	Risk	Who can mitigate	Real examples	Controls (How to mitigate)
EDH	Excessive Data Handling	Model Creator	Samsung banned the use of ChatGPT after discovering private code source has leaked through prompts submitted to GenAI.	• User Data Management
MXF	Model Exfiltration	Model Creator, Model Consumer	Meta's Llama model was leaked online, bypassing Meta's license acceptance review process.	• Model and Data Inventory Management • Model and Data Access Control • Secure-by-Default ML Tooling
MDT	Model Deployment Tampering	Model Creator, Model Consumer	Researchers discovered that models on HuggingFace were using a shared infrastructure for inference, which allowed a malicious model to tamper with any other model.	• Secure-by-Default ML Tooling
DMS	Denial of ML Service	Model Consumer	Researchers have demonstrated that slight image perturbation can cause denial of service in object detection models.	• Application Access Management

Code	Risk	Who can mitigate	Real examples	Controls (How to mitigate)
MRE	Model Reverse Engineering	Model Consumer	A Stanford University research team developed Alpaca 7B, a model fine-tuned from LLaMA 7B using 52,000 instruction-following examples.	• Application Access Management
IIC	Insecure Integrated Component	Model Consumer	By uploading a malicious Alexa skill/Google action (plugins), attackers were able to eavesdrop on user conversations that occurred near Alexa/ Google Home devices.	• Agent Permissions
PIJ	Prompt Injection	Model Creator, Model Consumer	One example of indirect Prompt Injection involved planting malicious data in a resource included in the LLM's prompt. In another example, multimodal prompt injection attacks against GPT-4V demonstrated that images can contain text capable of triggering a Prompt Injection attack when the model is asked to describe the image.	• Input Validation and Sanitization • Adversarial Training and Testing • Output Validation and Sanitization

Code	Risk	Who can mitigate	Real examples	Controls (How to mitigate)
MEV	Model Evasion	Model Creator	Adversarial images have been used to modify street signs to confuse self-driving cars.	• Adversarial Training and Testing
SDD	Sensitive Data Disclosure (Updated)	Model Creator, Model Consumer	One study showed that recitation checkers that scan for verbatim repetition of training data may be insufficient. An example of membership inference attacks showed the possibility of inferring whether a specific user or data point was used to train or tune the model.	• Privacy Enhancing Technologies • User Data Management • Output Validation and Sanitization • Agent Permissions • Agent User Control • Agent Observability
ISD	Inferred Sensitive Data	Model Creator, Model Consumer	Examples include research on AI inferring attributes such as sexual orientation or criminal history from facial images.	• Training Data Management • Output Validation and Sanitization

Code	Risk	Who can mitigate	Real examples	Controls (How to mitigate)
IMO	Insecure Model Output	Model Consumer	Attackers can compromise users by creating fake malicious packages with names inspired by LLM hallucinations.	• Output Validation and Sanitization
RA	Rogue Actions (Updated)	Model Consumer	An attack on ChatGPT plugins was described in Plugin Vulnerabilities: Visit a Website and Have Your Source Code Stolen.	• Agent Permissions • Agent User Control • Agent Observability • Output Validation and Sanitization

Source: Google Secure AI Framework: Risks

Risks to the Model Context Protocol (MCP)

About MCP

MCP is an open standard developed by Anthropic alongside a growing open-source community. It provides a structured framework for connecting large language models (LLMs) and AI agents to external tools, data sources, and services.

At its core, MCP addresses a key challenge in agentic AI systems—enabling models to dynamically access and interact with real-world resources while maintaining security, reliability, and consistency through standardization. By introducing a common protocol, MCP reduces the need for bespoke integrations with each database, API, file system, or web service.

The protocol follows a client-server architecture. AI applications (hosts) use MCP clients to connect to MCP servers, which expose capabilities such as tools, data resources, and reusable prompts.

MCP defines standardized methods for:

- Discovering available capabilities (e.g., tools and services)
- Invoking parameterized operations
- Accessing structured and unstructured data resources

To support different deployment contexts, MCP includes multiple transport mechanisms. These include standard input/output (stdio) for local processes and Streamable HTTP for network-based communication, enabling use cases that range from local development environments to distributed cloud systems.

Did you know?

Google introduced their Agent2Agent Protocol (A2A) in April 2025, built with support and contributions from 50+ technology partners.

The company aims to complement MCP and address the challenges in deploying large-scale, multi-agent systems for their customers.

MCP Threats

The table below organizes the threats by category and maps them to controls and mitigations.

Threat	Threat Category	MCP Specific	MCP Contextualized	Conventional Security	Control and Mitigation
MCP-T1	Improper Authentication and Identity Management	01. Identity Spoofing	08. Confused Deputy (OAuth Proxy)	16. Credential Theft/Token Theft 17. Replay Attacks/Session Hijacking 18. OAuth/Legacy Auth Weaknesses 19. Session Token Leakage	Agent Identity Secure Delegation (i.e., OAuth delegation)
MCP-T2	Missing or Improper Access Control	—	09. Insecure Human-in-the-Loop 10. Improper Multitenancy	08. Privilege Escalation 20. Excessive Permissions/ Overexposure	• Secure Delegation • Access Control
MCP-T3	Input Validation/ Sanitization Failures	—	—	21. Command Injection 22. File System Exposure/Path Traversal 23. Insufficient Integrity Checks	• Data Sanitization • Guardrails • Sandboxing & Isolation (Roots support)

Threat	Threat Category	MCP Specific	MCP Contextualized	Conventional Security	Control and Mitigation
MCP-T4	Data/Control Boundary Distinction Failure	02. Tool Poisoning 03. Full Schema Poisoning 04. Resource Content Poisoning	11. Prompt Injection	21. Command Injection	• Input Sanitization, Guardrails • Context Isolation
MCP-T5	Inadequate Data Protection & Confidentiality Controls	—	—	24. Data Exfiltration & Corruption 22. File System Exposure/Path Traversal	• Sandboxing & Isolation • Access Control • Guardrails
MCP-T6	Missing Integrity/ Verification Controls	04. Resource Content Poisoning 05. Typo squatting/ Confusion Attacks 06. Shadow MCP Servers	—	25. Supply Chain Compromise and Privileged Host-Based Attacks	• Cryptographic Integrity • Remote Attestation • MCP Server Integrity

Threat	Threat Category	MCP Specific	MCP Contextualized	Conventional Security	Control and Mitigation
MCP-T7	Session & Transport Security Failures	—	12. Man-in-the-Middle (MITM)	26. Unrestricted Network Access 27. Protocol Security Gaps 28. Insecure Descriptor Handling 23. Insufficient Integrity Checks 29. CSRF Protection Missing 30. CORS/ Origin Policy Bypass	• Network & Transport Security
MCP-T8	Network Binding/ Isolation Failures	6. Shadow MCP Servers	10. Improper Multitenancy	31. Malicious Command Execution 32. Dependency/ Update Attack 26. Unrestricted Network Access	• Network & Transport Security • Sandboxing & Isolation
MCP-T9	Trust Boundary & Privilege Design Failures	7. Overreliance on the LLM	13. Consent/User Approval Fatigue	—	• Secure Tool Design • UX Design
MCP-T10	Resource Management / Rate Limiting Absence	—	14. Resource Exhaustion and Denial of Wallet	33. Payload Limit/ DoS	• Network & Transport Security

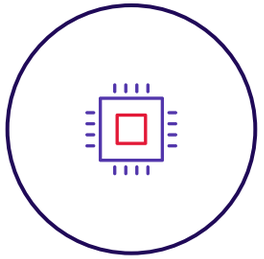
Threat	Threat Category	MCP Specific	MCP Contextualized	Conventional Security	Control and Mitigation
MCP-T11	Supply Chain and Lifecycle Security Failures	6. Shadow MCP Servers	—	25. Supply Chain Compromise	• Lifecycle Governance
MCP-T12	Insufficient Logging, Monitoring & Auditability	—	15. Invisible Agent Activity	34. Lack of Observability	• Logging; Lifecycle Governance

Source: Coalition for Secure AI: Model Context Protocol (MCP) Security, March 2026

Accelerate your journey to safeguarding AI

As AI adoption scales, security must evolve from an afterthought to a foundational component of the AI lifecycle. This requires organizations to move beyond traditional cybersecurity approaches and adopt AI-specific governance, risk management, and technical controls aligned with emerging standards such as NIST AI RMF and ISO/IEC 42001.

Ultimately, securing AI is not only about mitigating risk—it is about enabling trust. Organizations that proactively address AI security will be better positioned to meet stakeholder expectations, protect their assets, and unlock the full value of AI innovation.



Visit our AI Governance, Risk and compliance page at
cgi.com/canada/en-ca/cybersecurity/ai-governance

Visit our Cybersecurity page at
cgi.com/canada/en-ca/cybersecurity

Email us at
cybersecurity.ca@cgi.com

About CGI

Insights you can act on

Founded in 1976, CGI is among the largest IT and business consulting services firms in the world.

We are insights-driven and outcomes-focused to help accelerate returns on your investments. Across 21 industry sectors in 400 locations worldwide, our 94,000 professionals provide comprehensive, scalable and sustainable IT and business consulting services that are informed globally and delivered locally.

cgi.com